Interpretation of Shotgun Proteomic Data

THE PROTEIN INFERENCE PROBLEM*

Alexey I. Nesvizhskii‡§ and Ruedi Aebersold‡¶

The shotgun proteomic strategy based on digesting proteins into peptides and sequencing them using tandem mass spectrometry and automated database searching has become the method of choice for identifying proteins in most large scale studies. However, the peptide-centric nature of shotgun proteomics complicates the analysis and biological interpretation of the data especially in the case of higher eukaryote organisms. The same peptide sequence can be present in multiple different proteins or protein isoforms. Such shared peptides therefore can lead to ambiguities in determining the identities of sample proteins. In this article we illustrate the difficulties of interpreting shotgun proteomic data and discuss the need for common nomenclature and transparent informatic approaches. We also discuss related issues such as the state of protein sequence databases and their role in shotgun proteomic analysis, interpretation of relative peptide quantification data in the presence of multiple protein isoforms, the integration of proteomic and transcriptional data, and the development of a computational infrastructure for the integration of multiple diverse datasets. ***Molecular & Cellular Proteomics 4:1419–1440, 2005.***

An explicit goal of proteomics is the identification and quantification of all the proteins expressed in a cell or tissue (1). Although not yet at the levels of data throughput and automation achieved in other genomic analyses such as DNA sequencing or microarray gene expression analysis, global protein profiling methods are rapidly evolving. This has been possible because of recent improvements in MS instrumentation, protein and peptide separation techniques, computational data analysis tools, and the availability of complete sequence databases for many species. As a result, analysis of complex protein mixtures using shotgun proteomics, a strategy based on the combination of protein digestion and MS/MS-based peptide sequencing (2–4), has become widely adopted. The method allows protein identifications and, when combined with stable isotope labeling, quantification of the changes in the protein expression levels for hundreds of proteins in a single experiment (1).

Compared with other MS-based proteomic technologies such as intact proteins sequencing (5, 6) or 2D[1] gel-based protein analysis (7), shotgun proteomic analysis has achieved a relatively high throughput. This is the result of a combination of several factors. Proteolytic digestion of proteins into shorter peptides simplifies MS/MS sequencing (peptides are easier to fragment in the mass spectrometer than intact proteins), whereas elimination of the 2D gel-based separation at the protein level simplifies sample handling and increases the overall data throughput. At the same time, computational analysis and interpretation of the data become more challenging (8–13). The first and foremost computational challenge is the need to process large volumes of acquired MS/MS data with the purpose of identifying peptides that gave rise to observed spectra. This challenge is now well understood, and a number of computational methods and software tools, including programs for assigning peptides to MS/MS spectra (14–20) and for statistical validation of those assignments (21–25), have been developed. However, identification of peptides resulting from proteolytic digestion of sample proteins represents only an intermediate step because the ultimate goal of most experiments is to identify (and quantify when appropriate) the proteins that are present in the original sample. Increasingly it has been realized that the protein inference problem, *i.e.* the task of assembling the sequences of identified peptides to infer the protein content of the sample, is far from being trivial and requires special attention (22, 26–30).

The difficulty of assembling peptide identifications back to the protein level results from the same factors that made the shotgun proteomic approach so successful in the first place, *i.e.* protein digestion at an early stage of the process and elimination of extensive separation at the protein level. Protein digestion makes peptides, and not the proteins, the currency of the method, and the connectivity between peptides and proteins is lost at the digestion stage. This loss of connectivity complicates computational analysis and biological interpretation of the data especially in the case of higher eukaryote organisms. The same peptide sequence can be present in multiple different proteins. Therefore, the identification of such *shared*[2] peptides can lead to ambiguities in the determination of the identities of the sample proteins (see Fig. 1). In general,

From the ‡Institute for Systems Biology, Seattle, Washington 98103 and ¶Institute for Molecular Systems Biology, ETH-Zurich, CH-8093 Zurich, Switzerland

Received, June 6, 2005, and in revised form, July 5, 2005

Published, MCP Papers in Press, July 11, 2005, DOI 10.1074/mcp.R500012-MCP200

[1] The abbreviations used are: 2D, two-dimensional; EST, expressed sequence tag; SILAC, stable isotope labeling by amino acids in cell culture; CNBP, cellular nucleic acid-binding protein.

[2] Also referred to as *degenerate* peptides (see *e.g.* Ref. 22).

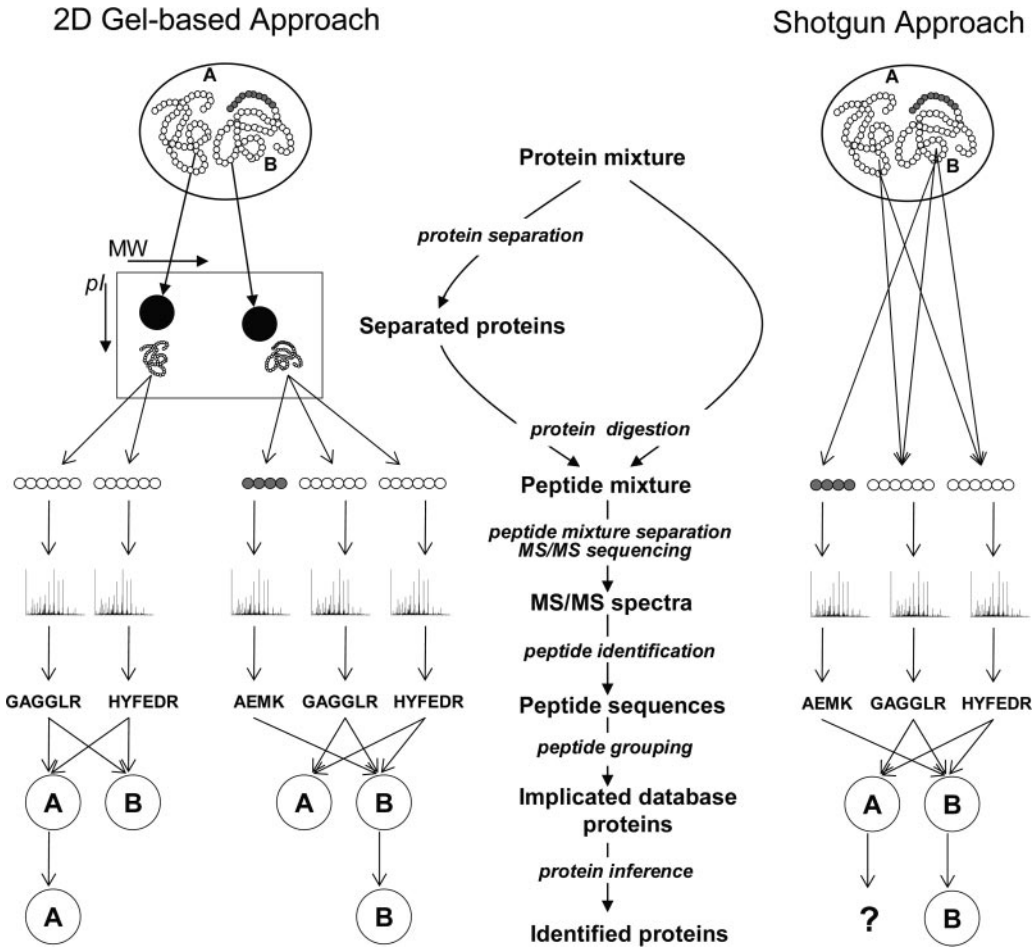


FIG. 1. **Protein identification using MS/MS.** On the *left* is a depiction of the process in a typical 2D gel-based approach: the proteins are separated, visualized, and excised from the gel. On the *right* is the process involved for shotgun proteomics. In both approaches, sample proteins are proteolytically digested into peptides, and resulting peptides mixtures are separated using liquid chromatography. Peptides are ionized, and selected peptide ions are subjected to MS/MS sequencing. Peptide sequences are determined from MS/MS spectra using a database search approach. Any given peptide may be a part of the sequences of several different proteins. The protein inference problem involves figuring out which proteins are present in the sample given the sequences of identified peptides. In this example, the sample contains two proteins, A and B, which share extensive sequence homology. All three identified peptides, AEMK, GAGGLR, and HYFEDR, are present in the sequence of protein B, and the last two peptides are also in the sequence of protein A. In the shotgun proteomic approach, the connectivity between peptides and proteins is lost; no information on the number of proteins in the sample or their properties (*e.g.* molecular weight) is available. It is not possible to conclude that protein A is present in the sample because protein B can account for all observed peptides. This is less of a problem in the case of the 2D gel-based approach where proteins are separated prior to digestion and MS/MS analysis.

protein identification is a less complex issue[3] if proteins are first separated using a multidimensional protein separation technique (*e.g.* 2D gels) where additional information, such as the protein molecular weight and isoelectric point, can assist in determination of the protein identities (see *e.g.* Refs. 7 and 31–33).

In this article we illustrate the protein inference problem of shotgun proteomics using a set of examples and discuss the need for common nomenclature and transparent informatic approaches for assembling peptides into proteins and presenting the results of shotgun proteomic experiments to the user. We also discuss related issues such as the state of protein sequence databases and their role in shotgun proteomic analysis, interpretation of quantitative proteomic data in the presence of multiple protein isoforms, correlation of proteomic and transcriptional data, and comparison and integration of shotgun proteomic data generated in different experiments.

THE PROTEIN INFERENCE PROBLEM: CASE STUDIES

Although the shotgun proteomic approach is peptide-centric, in most cases researchers are ultimately interested in knowing what proteins are present in the analyzed sample.

[3] This is true, however, only under the assumption that only a single protein is present in a spot on the gel, which is not always the case.

Inferring protein identities given a set of identified peptides becomes difficult in the case of higher eukaryote organisms. This is due to sequence redundancy, *i.e.* the presence of distinct proteins having a high degree of sequence homology, as is the case in protein families, alternative splice forms of the same gene, differentially processed proteins, and more (26, 34).

Although the identification of a single peptide is often sufficient to conclude that a product of a certain gene is present in the sample, it is often not possible to discriminate between different proteins that share extensive homology or are isoforms arising from alternatively spliced genes as illustrated in Fig. 2. All examples used in this work, except where noted, were found in the dataset from an experiment on lipid raft plasma membrane domains from human Jurkat human T cells (27).[4] In the first example, Fig. 2*A*, several peptides were identified that are present in two different splice forms of the F-actin capping protein β subunit, CAPB_HUMAN, P47756-1 and P47756-2. Because there are no peptides identified in the experiment that would correspond to one of the isoforms only, both isoforms are equally likely; any one of them, or both, could be present in the sample. In this particular case, the sequences of the two isoforms differ significantly at the C terminus. Thus, discrimination between these two isoforms using sequence information alone would be possible only with the identification of peptides spanning the areas where the sequences diverge.

Conclusive identification of alternative splice forms arising from skipping of one or more consecutive exons at the 5′- or 3′-end of the gene sequence (without introduction of any divergent sequence) is more challenging. One such example is shown in Fig. 2*B*. The sequences of the two shorter isoforms of the epithelial protein lost in neoplasm, EPLI_HUMAN (Q9UHB6-2 and Q9UHB6-3, isoforms α and 3, respectively) are included in the sequence of the longer isoform (Q9UHB6-1, isoform β). No conclusive evidence for the presence of the shorter isoforms in the sample can be obtained without any additional information, *e.g.* molecular weight of the sample proteins.[5]

Furthermore in some cases it is not possible to discriminate between proteins that are products of different genes from the same gene family (gene paralogues) (22, 26, 35, 36). This is illustrated in Fig. 3. A total of 11 peptides were identified in the dataset that are shared between more than a dozen different members of the α-tubulin family. None of the identified peptides is unique to any of the proteins. Thus, although those peptides clearly indicate the presence of one or more α-tubulin proteins, it is not possible to determine which particular member(s) of that family is present in the sample.

Interpretation of the data is further complicated due to artificial redundancies, *e.g.* truncated sequences, sequence alternatives arising from sequencing errors, and existence of essentially the same sequences under different gene names, features that exist in many protein sequence databases. This is illustrated in Fig. 4 where a group of peptides was identified that are shared between four different entries in the Entrez Protein sequence database maintained by the National Center for Biotechnology Information (NCBI).[6] Manual examination revealed that all four database entries represent the same protein, heat shock 70-kDa protein 9B (HSPA7B). Three of those entries are derived from mRNA sequences containing small sequence variations. In this particular case, the sequence variations are likely due to sequencing errors. However, these variations could also be polymorphisms, *i.e.* real sequence variants of the same protein from a different individual. In many cases, such database redundancies can only be resolved on a case by case basis by researchers analyzing the data. This presents an additional challenge for the development of automated informatic tools for dealing with large scale proteomic datasets.

Another problem encountered in shotgun proteomics is the difficulty of assigning the correct peptide sequence to an MS/MS spectrum. Two of the amino acids (Ile/Leu) have identical masses, and the difference between several other amino acid combinations (*e.g.* Asp/Asn and Glu/Gln/Lys) cannot be resolved using low mass accuracy instruments such as the commonly used ion traps. If the database contains several peptides with a similar molecular weight and having a high degree of sequence homology, determination of the correct peptide sequence among the alternatives becomes difficult or even impossible (in the case of Ile/Leu substitutions). This can result in the assignment of incorrect (although homologous) peptide sequences to MS/MS spectra, which in turn can result in incorrect protein identifications. Some but not all ambiguities can be resolved when using high mass accuracy instruments such as LTQ-FT or even Q-TOF.

Distinguishing between different proteins having a high degree of sequence similarity becomes increasingly difficult with decreasing protein sequence coverage, *i.e.* the fraction of the protein sequence covered by the identified peptides. The protein coverage observed in shotgun proteomic experiments is typically low. The number of "identifiable" peptides is obviously limited by the size of the protein but also by such factors as the enzymatic digestion constraint and the detection mass range of the mass spectrometer. In some cases, the

[4] Peptide and protein identification part of the analysis presented in Ref. 27 was repeated in this work using the Human IPI version 2.35 protein sequence database. Quantitative information (not discussed in the original publication) was extracted from the data using an automated tool, ASAPRatio (63), and then confirmed by manual inspection.

[5] It should be noted that the absence of evidence should not be interpreted as the evidence of absence of the protein in the sample.

[6] For the purpose of this discussion, the dataset of MS/MS spectra from Ref. 27 was also searched against the Entrez Protein database (downloaded in November 2004).

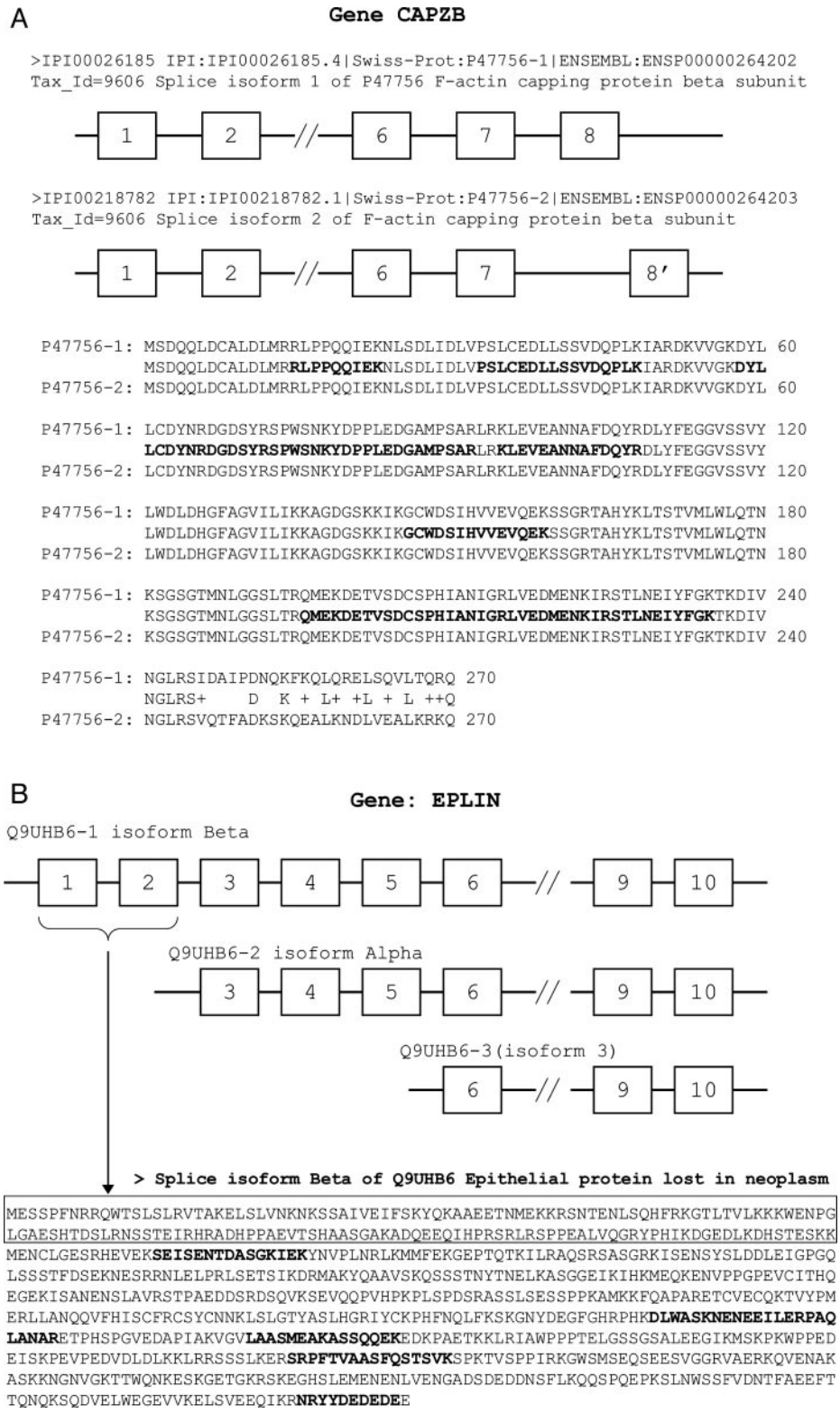


FIG. 2. **Sequences of identified peptides often do not allow discrimination between different protein isoforms.** *A*, multiple peptides are identified that are present in two different splice forms, P47756-1 and P47756-2, of the F-actin capping protein β subunit. The alignment of the sequences of the two isoforms is shown, and the sequences of the identified peptides are shown in *bold*. The isoforms are indistinguishable given the available data. The discrimination would be possible if peptides spanning the areas where the sequences diverge, *e.g.* SIDAIPDNQK (unique to P47756-1) and SVQTFADK (unique to P47756-2), were identified. *B*, protein isoforms of epithelial protein lost in neoplasm. Three isoforms result from splicing of several consecutive exons located at the 5′-end in the gene sequence. Given the sequences of the identified peptides (shown in *bold*), it is not possible to determine precisely which isoform is present. The sequences of the two shorter isoforms (Q9UHB6-2 and Q9UHB6-3) are included in the sequence of the longer isoform (Q9UHB6-1). Identification of a peptide from the region present in the isoform β only (sequence shown in a *box*) would allow conclusive identification of this isoform (no such peptides were actually observed in the experiment). Conclusive identification of the shorter isoforms would be difficult because they do not contain any unique sequence.

pool of potential peptide identifications is further reduced as a result of selective enrichment for a particular class of peptides, *e.g.* cysteine-containing peptides in quantitative proteomic experiments based on ICAT reagents (3). The identification of some peptides can be prevented by unexpected post-translational modifications. Furthermore because multiple different peptide ions are injected in the mass spectrometer (operated in top-down ion selection mode) at any given

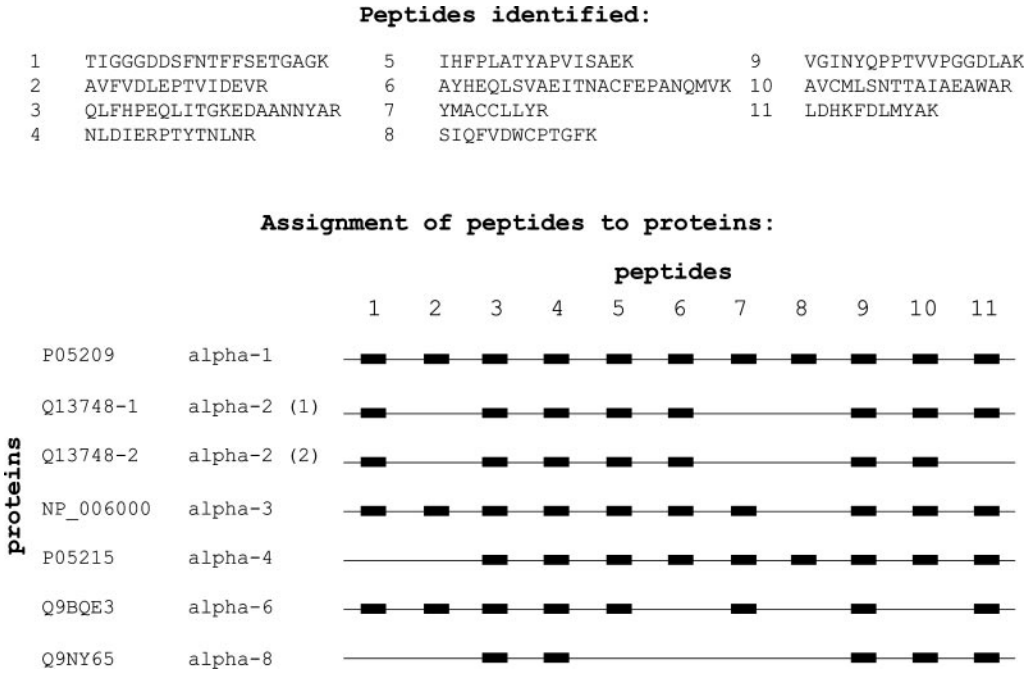


FIG. 3. **An example of a protein family.** Eleven tryptic peptides are identified that are shared between the members of the α-tubulin family. None of the proteins is identified by a peptide that is unique to it, thus making it impossible to determine which particular member(s) of the family is present in the sample.

Sequence entry 1:

gi|24234688|ref|NP_004125.3| heat shock 70kDa protein 9B precursor; heat shock 70kD protein 9

gi|21264428|sp|P38646|GR75_HUMAN Stress-70 protein, mitochondrial precursor (75 kDa glucose regulated protein) (GRP 75) (Peptide-binding protein 74) (PBP74) (Mortalin) (MOT)

gi|477763|pir||B48127 dnaK-type molecular chaperone precursor, mitochondrial - human

Sequence entry 2: *G → R at position 540*

gi|292059|gb|AAA67526.1| MTHSP75

Human mRNA L15189

Sequence entry 3: *Q → R at position 74; H → R at position 184*

gi|12653415|gb|AAH00478.1| Heat shock 70kDa protein 9B, precursor [Homo sapiens].

Human mRNA BC000478

gi|18645123|gb|AAH24034.1| Heat shock 70kDa protein 9B, precursor [Homo sapiens].

Human mRNA BC024034

Sequence entry 4: *M → V at N terminus*

gi|21040386|gb|AAH30634.1| heat shock 70kD protein 9B (mortalin-2) [Homo sapiens].

Human mRNA BC030634

FIG. 4. **Sequence database redundancies complicate the analysis of shotgun proteomic data.** Four separate entries in the Entrez Protein database represent the same protein, heat shock 70-kDa protein 9B. Three of them (entries 2–4) are derived from mRNA sequences containing small sequence variations.

time, low intensity ions, produced by low abundance or poorly ionizing peptides, are less likely to be selected for MS/MS sequencing (1). Finally some peptides, due to their physical-chemical properties, cannot be efficiently ionized or fragment in an atypical way producing MS/MS spectra unidentifiable by the current database search tools. As a result, more than 30% of all proteins that are detected in a typical shotgun proteomic experiment, including many low molecular weight or low abundance proteins, are identified by a single peptide.

ASSEMBLING PEPTIDES INTO PROTEINS

Results of large scale proteomic experiments are often presented as lists of protein identifications. At present, significant inconsistencies exist in the way different research groups assign peptides to proteins and deal with biological and database redundancies. The criteria for calling a protein "identified" are not always described, and there is no generally accepted way to do it. Shared peptides (peptides present in more than one sequence database entry) are sometimes

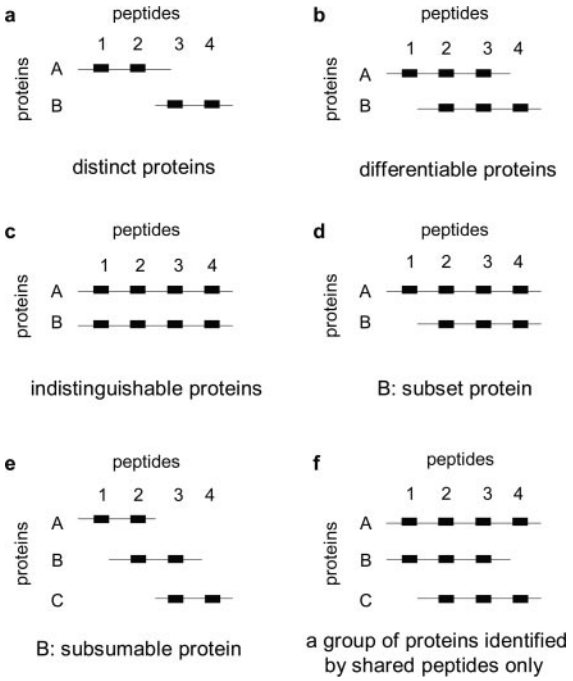


FIG. 5. **Basic peptide grouping scenarios.** *a*, distinct protein identifications. *b*, differentiable protein identifications. *c*, indistinguishable protein identifications. *d*, subset protein identification. *e*, subsumable protein identification. *f*, an example of a protein group where one protein can explain all observed peptides, but its identification is not conclusive.

assigned to a particular protein among several possibilities in a random fashion. Different sequence database entries could be counted as separate protein identifications when in fact all of them share the same set of peptides and, therefore, are indistinguishable. In most cases not only do these redundancies inflate the total number of proteins reported as identified, but they can also lead to incorrect biological interpretation of the data. The problem is further complicated when no statistical analysis is performed to determine the validity of peptide and protein identifications (10, 29). Thus, there is a need to develop a common nomenclature and a set of guidelines for assigning peptides to proteins and for interpreting resulting protein identification datasets.

The nomenclature described below provides a consistent way for presenting the results of large scale proteomic experiments. In creating a protein summary list that accurately represents the data, various peptide grouping scenarios have to be considered that are schematically illustrated in Fig. 5 (22, 30). The diagram in Fig. 5*a* describes a case of two *distinct* proteins, A and B, each identified by *distinct*[7] peptides only, *i.e.* peptides corresponding to that one protein and no other proteins (peptides 1 and 2 are unique to protein A, and peptides 3 and 4 are unique protein B). Fig. 5*b* shows a case of two *differentiable* proteins, which are identified by at least one distinct peptide (peptide 1 is unique to A, and peptide 4 is unique to protein B) but also by one or more *shared* peptides (peptides 2 and 3 are shared between the two proteins). A different scenario is shown in Fig. 5*c* where all peptides are shared between proteins A and B. These two proteins are *indistinguishable* given the sequences of the identified peptides, and either protein A, protein B, or both can be present in the sample. Fig. 5, *d* and *e*, each show a situation where all identified peptides corresponding to protein B are shared and can be accounted for by another protein (protein A in Fig. 5*d*) or a combination of several other proteins (proteins A and C in Fig. 5*e*) certain to be in the sample because they are identified by at least one distinct peptide. In general, no conclusion can be made regarding the presence of a *subset* (protein B in Fig. 5*d*) or a *subsumable* (Protein B in Fig. 5*e*) protein in the sample. A special case is shown in Fig. 5*f* where all identified peptides are shared by a group of proteins. The presence of protein A in the sample is sufficient to explain all observed peptides (B and C are subset protein identifications). Although protein A is the most likely candidate, its presence in the sample is not required to explain the data; it is identified by shared peptides only. In the absence of protein A, a combination of proteins B and C would account for all four peptides. Such situations are often observed in the case of extended protein families, such as the tubulin example shown in Fig. 3. The examples discussed above are exhaustive, *i.e.* it should be possible to explain more complicated cases observed in real datasets by reducing them to a combination of several basic grouping scenarios.

The nomenclature described here, coupled with the Occam's razor constraint (22), would provide a minimal list of proteins sufficient to explain all observed peptides. Such a minimal list would contain all distinct and differentiable proteins, *e.g.* proteins A and B in Fig. 5, *a* and *b*, and proteins A and C in Fig. 5*e* but no subsumable or subset proteins, *e.g.* only protein A would be included in the list in the cases shown in Fig. 5, *d* and *f*. In the case of indistinguishable protein identifications, Fig. 5*c*, it would be most accurate to collapse all such identifications into a single entry in the protein summary report as there is often no basis to eliminate any of them.

Presenting results of large scale shotgun experiments in terms of such minimal lists of protein identifications has several advantages. It significantly simplifies the interpretation of the data by allowing the user to focus on proteins that are conclusively determined to be present in the sample. It also allows calculation of a consistent measure for the number of proteins identified in the experiment as the smallest number of proteins that can explain all observed peptides (*i.e.* the number of entries in the minimal protein list).

At the same time, presenting only the minimal list of proteins has limitations. For example, a researcher interested in a particular gene might want to observe all related protein isoforms annotated in the protein sequence database that are implicated by at least one peptide identified in the experiment.

[7] Also referred to as *discrete* peptides (see Ref. 30).

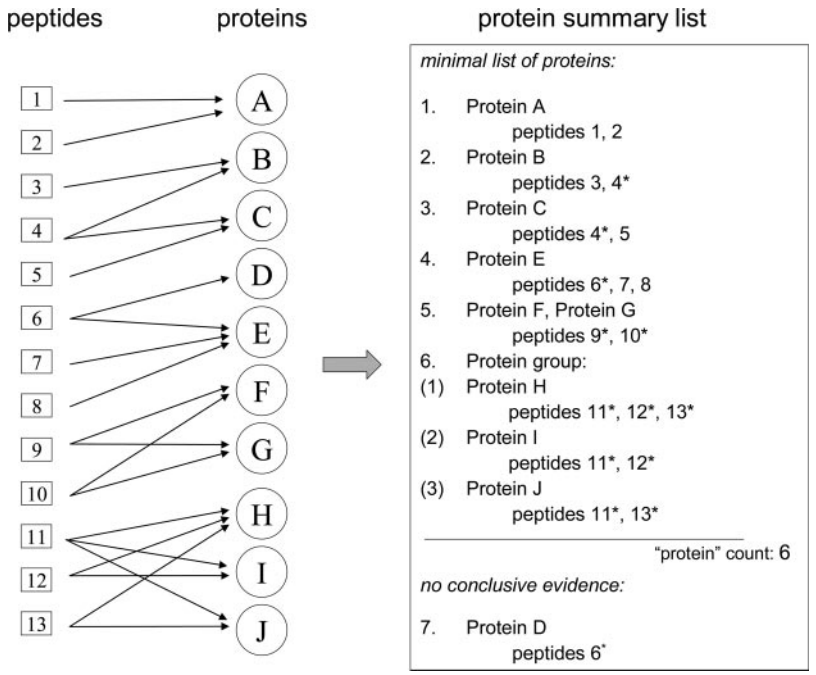


FIG. 6. **A simplified example of a protein summary list.** Peptides are apportioned among all their corresponding proteins, and the minimal list of proteins is derived that can explain all observed peptides. Proteins that are impossible to differentiate on the basis of identified peptides are collapsed into a single entry (*F* and *G*) or presented as a group (*H*, *I*, and *J*). Shared peptides are marked with an *asterisk*. Proteins that cannot be conclusively identified are shown at the end of the list but do not contribute toward the protein count.

Moreover the strict implementation of the Occam's razor approach can be misleading when applied to complex protein families. In the α-tubulin protein family example shown in Fig. 3, none of the identified peptides are unique to the tubulin α-1 protein. Thus, although this protein can explain all observed peptides, its identification is not conclusive. In fact, in the absence of the α-1 tubulin, all peptides can be accounted for by a combination of several other tubulins, *e.g.* α-3 and α-6. Because it is not possible to determine which particular member(s) of that family is present in the sample, in creating a minimal list it is more accurate and informative to present all members together as a group (22). Therefore, the most advantageous presentation would include the following: (*a*) a minimal list with indistinguishable proteins collapsed into a single entry (but showing all protein names) and with all members of protein groups listed and (*b*) means to observe the proteins implicated by at least one peptide that cannot be called conclusively identified. A simplified illustration of such a format of presentation is shown in Fig. 6.

COMPUTATIONAL TOOLS

A number of computational tools for assembling peptides into proteins in large scale shotgun proteomic experiments have been described (22, 28, 30, 37, 38). In general, the process of peptide assembly consists of the following steps. First, peptide assignments obtained by searching acquired MS/MS spectra against a protein sequence database using algorithms such as SEQUEST (14) or Mascot (16) are filtered using a user-specified set of criteria to remove false identifications. Second, accession numbers and annotations of protein sequence database entries corresponding to each peptide are retrieved from the sequence database. Third, peptides are grouped by their corresponding sequence database entries. Fourth, shared peptides are apportioned among all corresponding proteins, and a summary protein list is created. Ideally the apportionment of peptides to proteins should be done using a probability-based approach, *i.e.* taking into account the probabilities of peptide assignments (22). This has an advantage in that it allows calculation of statistical confidence measures for protein identifications and estimation of false identification error rates resulting from filtering the data (10, 22).

The format in which the results of shotgun proteomic experiments are presented to the user varies between the tools. In ProteinProphet (22), each separate entry in the protein summary file is assigned a probability that the corresponding protein is present in the sample. Indistinguishable proteins are collapsed into a single entry, and all members of protein groups, such as the α-tubulin family shown in Fig. 3, are presented together. All subset and subsumable protein entries are assigned zero probability, which is to be interpreted as the absence of conclusive evidence for the presence of those proteins in the sample. The subset and subsumable protein entries can be located and viewed using interactive web-based options. In the Experimental Peptide Identification Repository (EPIR) (38), the notion of protein groups introduced in Ref. 22 is extended, and all entries with shared peptides are organized into a single group. The protein that contains most of the peptides is selected as an anchor, and all group members that are identified by at least one distinct peptide are marked as conclusively identified. Additional visualization tools, *e.g.* a tool for aligning the sequences of all proteins within a protein group, are provided to assist in the interpretation of the data. Other software tools such as Iso-

TABLE I

Summary of the protein sequence databases that are commonly used in shotgun proteomic analysis

Database sizes and the number of sequences are given for the human subset of each database only. EBI, European Bioinformatics Institute; SIB, Swiss Institute of Bioinformatics.

Database, date (version)	Number of sequences; size of file (human)	Description; source databases	Organisms	Release; update frequency; maintained by
Uni-Prot/Swiss-Prot, 02/15/2005	11,898; 7.8 Mb	Expertly curated; high level of annotation; minimum level of redundancy; high level of integration with other databases.	Many	Release every 4 months; updates every 2 weeks; EBI, SIB, Georgetown University
Uni-Prot/TrEMBL, 02/15/2005	52,052; 23.3 Mb	Computer-annotated supplement to Uni-Prot/Swiss-Prot. Contains translated coding sequences from GenBank™ nucleotide database, protein sequences extracted from the literature or submitted to Uni-Prot/Swiss-Prot but not yet manually curated.	Many	Release every 4 months; updates every 2 weeks; EBI, SIB, Georgetown University
RefSeq, 08/26/2004 (R 9)	27,960; 17.7 Mb	Ongoing curation by NCBI staff; non-redundant; explicitly linked nucleotide and protein sequences; stable reference; high level of integration with other databases.	Many	Release every ~3 months; NCBI
Ensembl, 02/2005 (version 28-35a)	33,860; 21.1 Mb	Created using automated genome annotation pipeline; eukaryotic genomes only; explicitly linked nucleotide and protein sequences; stable reference; high level of integration with other databases. Peptides identified by MS/MS can be mapped to the genome via Ensembl Protein database and visualized using Ensembl Genome Browser.	16 organisms	Every 1–2 months; EBI and Wellcome Trust Sanger Institute
IPI, 02/2005 (version 3.03)	48,953; 28.9 Mb	Good balance between degree of redundancy and completeness; references to the primary data sources; attempts to maintain stable identifiers (with incremental versioning), but still in flux. Assembled from Uni-Prot (Swiss-Prot + TrEMBL), RefSeq, Emsembl, H-Invitational database.	5 organisms	Monthly; EBI
Entrez Protein (NCBInr), 02/17/2005	115,926; 58.5 Mb	More complete with regard to sequence polymorphisms and splice forms; annotations extracted from curated databases; high degree of sequence redundancy makes interpretation difficult. Assembled from GenBank™ and RefSeq coding sequence translations, Protein Information Resource (PIR), Protein Data Bank (PDB), Uni-Prot/Swiss-Prot, Protein Research Foundation (PRF).	Many	Frequent updates; NCBI

form Resolver (28) and DBParser (30) create protein summary lists containing all protein sequence database entries identified by at least one peptide with proteins that share a set of peptides placed adjacent to each other. In Isoform Resolver, the protein summary lists are presented in a text format, and a peptide-centric numbering scheme is used to specify what proteins are identified conclusively. DBParser outputs the results in an interactive web-based format that allows the user to view both the redundant and the minimal list of proteins.

DTASelect (39) is another widely used tool for processing of shotgun proteomic data. However, it does not provide any statistical confidence measures for protein and peptide identifications, and its approach for assembling peptides into proteins in the presence of shared peptides has not been fully described. In addition, new tools are being developed at increasing speed, including commercial programs that combine the process of peptide identification and the subsequent assembly of peptides into proteins (40).[8] This diversity of computational tools, a positive development reflecting the increased used of shotgun proteomics, nevertheless presents a significant challenge for developing any kind of standards for the analysis and journal publication of proteomic datasets

[8] SpectrumMill (www.chem.agilent.com).

(10, 29). It is thus essential that the computational tools are made transparent (published) and extensively tested and that the methods for assembling peptides into proteins and presenting the results all follow the same set of general guidelines such as those described in this article.

PROTEIN SEQUENCE DATABASES

Computational analysis and biological interpretation of shotgun proteomic data requires selection of a reference protein sequence database. For some organisms, *e.g.* human, several different databases exist that vary in terms of completeness, degree of redundancy, and quality of sequence annotation (42). Table I and Fig. 7 summarize some of the existing protein sequence databases that are commonly used with mass spectrometry data. The choice of a particular database should be based on the goals of the experiment.

When peptides are assigned to MS/MS spectra using the database search approach, the universe of all potential peptide assignments is limited to the sequences present in the searched protein sequence database. The completeness of the sequence database thus can be a decisive factor in experiments where identification of sequence polymorphisms is crucial for the biological interpretation of the data. In those cases, a large database such as Entrez Protein (also known as the non-redundant NCBI database, NCBInr) (43) would have an advantage over smaller databases such as Uni-Prot/Swiss-Prot (44) or RefSeq (45). The Entrez Protein database, for example, contains twice as many unique tryptic peptide sequences as Uni-Prot/Swiss-Prot (Fig. 7*A*). At the same time, large sequence databases contain, in addition to true biologically significant sequence variants, numerous artificial redundancies arising *e.g.* from partial mRNAs or sequencing errors (see example in Fig. 4). Fig. 7*B* plots the average number of database entries containing each unique tryptic (with no missed cleavages) peptide sequence as a function of peptide molecular weight. For example, in the range of molecular weights around 1000, the majority of tryptic peptides in the Swiss-Prot database are distinct ($N_{tot}/N_{unique} \sim 1$), whereas in the Entrez Protein database each peptide is present on average in three different entries ($N_{tot}/N_{unique} \sim 3$). In the absence of good sequence annotation in large protein sequence databases such as Entrez Protein database, it becomes necessary to perform time-consuming manual analysis and elimination of database redundancies. Furthermore searching such large databases makes it more difficult to separate the correct from random (incorrect) peptide assignments to MS/MS spectra.

When the quality of the sequence annotation and the ease of data interpretation are more important than the ability to identify sequence variants, it is more appropriate to use well curated databases such as Swiss-Prot or RefSeq. A good balance between the completeness and the level of redundancy is found in the International Protein Index (IPI) database (46), which is available for a number of organisms including human and mouse. The sequence- and identifier-based construction of this database significantly reduces the need for manual filtering while maintaining cross-references to all its source data, which include Ensembl (47), Uni-Prot (Swiss-Prot and its supplement TrEMBL) (44), and RefSeq (45). Minor sequence variants, however, are not represented in the IPI database.

Genomic databases can also be used for MS/MS database searching (48, 49), which can lead to the identification of novel

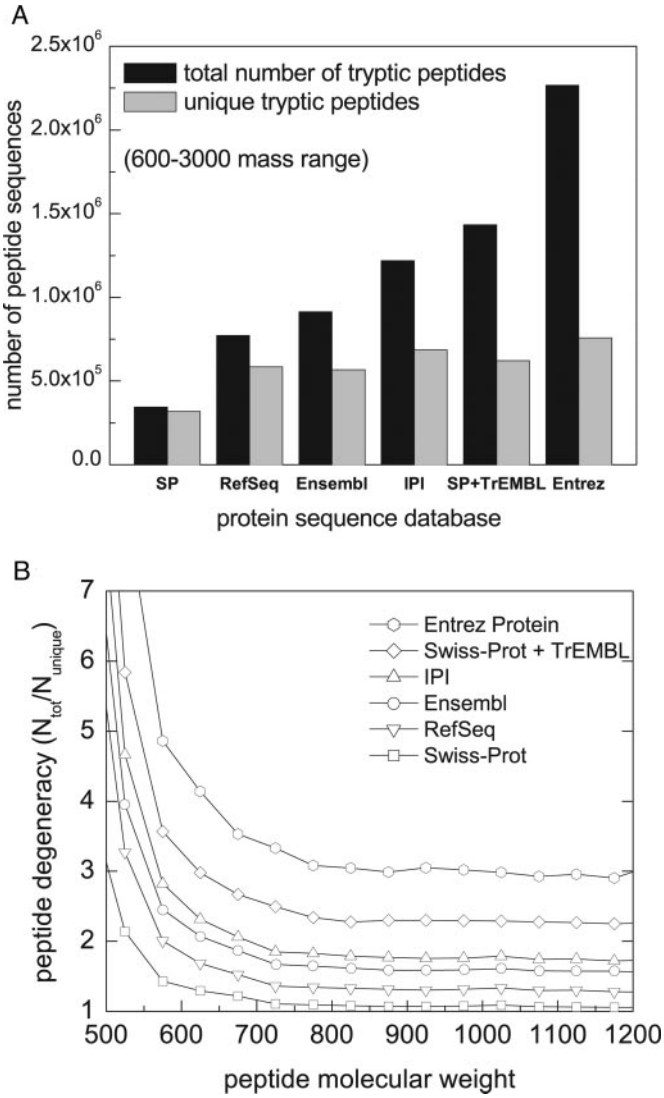


FIG. 7. **Protein sequence databases differ in terms of their completeness and the degree of sequence redundancy.** *A*, the total number of tryptic peptides with no missed cleavages (N_{tot}) and the number of unique sequences among them (N_{unique}), in the range of molecular weights between 600 and 3000, in each of the human protein sequence databases listed in Table I. *B*, a measure of the database sequence redundancy (average number of database entries containing each unique tryptic peptide sequence), estimated by taking the ratio N_{tot}/N_{unique}, plotted as a function of peptide molecular weight (bin size of 50 mass units) for the same databases. *SP*, Swiss-Prot.

alternative splice forms and sequence polymorphisms not present in the protein sequences databases. However, this type of computational analysis can be computer-intensive because of the large size of those databases. It is also complicated due to frameshifts, incorrectly predicted open reading frames, and poor quality of many EST sequences. This combined with the poor quality of many experimental MS/MS spectra can lead to high numbers of false identifications. A more efficient strategy for the identification of novel alternative splice forms or sequence variants is to perform computational analysis in an iterative fashion. In this approach, the analysis would start with searching MS/MS spectra against a well annotated database (*e.g.* RefSeq or IPI). The high quality spectra left unassigned in the initial search are then reanalyzed more extensively, first searching for post-translationally modified peptides and only then against large genomic databases.[9]

An important caveat to keep in mind when interpreting shotgun proteomic data is that the protein sequence databases are constantly in flux especially with regard to minor sequence variants, alternative splice forms, and other less well characterized gene products. With each new database update, some protein sequences disappear, the annotation and accession numbers of the remaining sequences can change, and new sequences can be added. The instability of the current sequence databases is largely due to a substantial amount of work being carried out to improve their completeness and the quality of sequence annotation, a process that is likely to continue for a significant period of time. This has significant implications in that interpretation of the MS/MS-based proteomic data, *e.g.* assignment of peptides to entries in the protein sequence database and conclusions about the presence of a particular protein isoform in the sample, depends on the version of the protein sequence database used in the analysis. Frequent updating of the sequence databases by the database providers can complicate ongoing proteomic experiments. Researchers using these databases in the analysis of their data often have to reanalyze previously acquired and processed MS/MS spectra using a new version of the database or develop bioinformatic tools for automated mapping of peptide sequences, identified by searching MS/MS spectra against an older version of the database, to the latest version of that database. It is important to note that the data coming from MS-based proteomic experiments can itself be used to assist in the process of improving protein sequence databases provided a mechanism is developed for communicating the sequences of peptides identified by searching those databases back to the database developers and annotators (50, 51).

[9] A. I. Nesvizhskii *et al.*, manuscript in preparation.

IDENTIFICATION OF MATURE FORMS OF PROTEINS

The discussion so far has mostly focused on the problem of assigning peptides to proteins and distinguishing between different protein forms whose sequences are present in the protein sequence database. A closely related issue is the difficulty of using shotgun proteomic data to provide conclusive information regarding the mature form of the sample proteins. First, most existing protein sequence databases contain entries that are derived from full cDNAs encoding preprocessed forms. Thus, they do not typically contain the mature forms derived from various post-translational processing mechanisms, *e.g.* removal of the leading methionine, cleavage of the signal or transit peptide, etc. Second, even if all mature forms were annotated in the protein sequence database, distinguishing between different protein isoforms would be difficult. For example, a mere observation that none of the identified peptides are coming from the N-terminal region of the protein does not necessarily indicate the cleavage of the presequence. It can be explained by other factors, *e.g.* the absence of identifiable tryptic peptides in that region.

In some cases, post-translational processing events can be inferred using the knowledge regarding the specificity of the proteolytic enzyme used to digest proteins into peptides. For example, the enzyme trypsin cleaves after arginine and lysine residues. A peptide resulting from trypsin digestion should contain Lys or Arg at its C terminus (unless it is located at the C terminus of the protein), and in the sequences of its corresponding protein the residue immediately preceding the peptide should also be Lys or Arg (or the peptide is located at the N terminus). Thus, identification of a peptide whose sequence does not adhere to the enzymatic digestion constraint at one of its termini could indicate that the mature form of the protein is present in the sample. One such example is shown in Fig. 8*A* where identification of a "partially tryptic" peptide (not tryptic at its N terminus), assigned to the "Basigin precursor" database entry (Swiss-Prot accession number P35613), suggests that the mature form of that protein resulting from the proteolytic cleavage of the 22-residue-long signal peptide is present in the biological sample. In general, identifying peptides that are not tryptic (assuming no protein cleavage) and are located close to the N terminus of the protein can be a useful strategy for inferring signal peptide cleavage sites or other proteolytic cleavage events, thus confirming, refining, or adding to the annotations currently available in the protein databases such as Swiss-Prot. In some cases, it can also assist in discrimination between different protein isoforms resulting from alternative splicing. It should be noted, however, that some partially tryptic peptides can be observed due to in-source or in-solution fragmentation of the originally tryptic peptides. Thus, conclusions based on the observation of partially tryptic peptides require additional scrutiny.

Another example is shown in Fig. 8*B* where several peptides were identified and assigned to a single protein se-

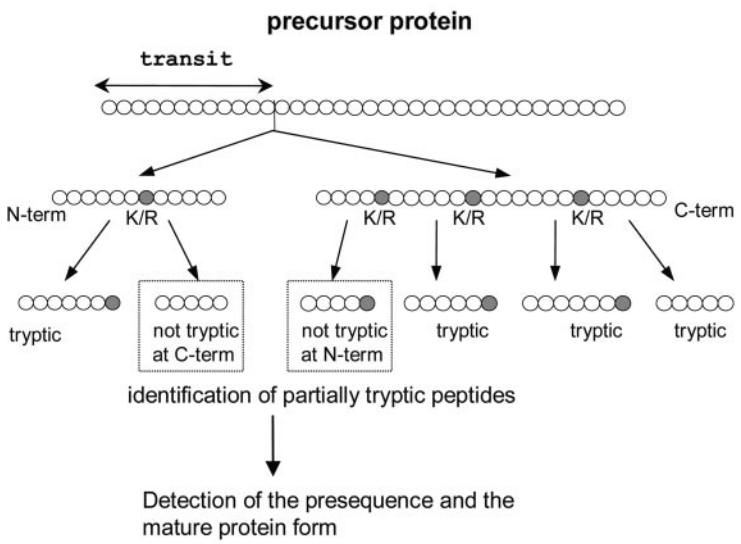
A

>IPI00019906.1|UniProt/Swiss-Prot:P35613 Basigin precursor

signal peptide

MAAALFVLLGFALLGTHGASG**AAGTVFTTVEDLGSK**ILLTCSLNDSATEVTGHR**WLKGGVVLKEDALPGQK**TEFKVDSDDQWGEYSCVFL
PEPMGTANIQLHGPPRVKAVK**SSEHINEGETAMLVCK**SESVPPVTDWAWYKITDSEDKALMNGSESRFFVSSSQGR**SELHIENLNMEADP**
GQYRCNGTSSKGSDQAIITLRVRSHLAALWPFLGIVAEVLVLVTIIFIYEKR**RKPEDVLDDDDAGSAPLK**SSGQHQNDKGKNVRQRNSS

B



>IPI00026964.1|UniProt/Swiss-Prot:P47985 Ubiquinol-cytochrome c reductase iron-sulfur subunit, mitochondrial precursor

RPLVASVGLNVPASVCY

transit peptide

MLSVAAR**SGPFAPVLSATSRGVAGALRPLVQATVPATPEQPVLDLK**RPFLSRESLSGQAVR**RPLVASVGLNVPASVCYSHTDIKVPDFSE**
YRRLEVLDSTKSSRESSEARKGFSYLVTGVTTVGVAYAAK**NAVTQFVSSMSASADVLALAK**IEIKLSDIPEGKNMAFKWRGKPLFVRHRT
QK**EIEQEAAVELSQLRDPQHDLDR**VKKPEWVILIGVCTHLGCVPIANAGDFGGYYCPCHGSHYDASGRIRLGPAPLNLEVPTYEFTSDDM
VIVG

SHTDIKVPDFSEYRRLEVLDSTK

FIG. 8. **Identification of mature forms of proteins in shotgun proteomics.** *A*, identification of a partially tryptic peptide, AAGTVFTTVEDLGSK, indicates the removal of the 22-residue-long signal peptide. *B*, identification of two partially tryptic peptides, RPLVASVGLNVPASVCY and SHTDIKVPDFSEYR, located adjacent to each other in the protein sequence indicates the cleavage of the 78-amino acid presequence of the ubiquinol-cytochrome *c* reductase iron-sulfur subunit. The processed presequence remains as a subunit of a protein complex.

quence database entry "ubiquinol-cytochrome *c* reductase iron-sulfur subunit, mitochondrial precursor" (Swiss-Prot accession number P47985, Rieske protein). Two of the peptides, RPLVASVGLNVPASVCY and SHTDIKVPDFSEYR, are partially tryptic and located adjacent to each other in the protein sequence, which suggests the cleavage of the 78-amino acid presequence (annotated in Swiss-Prot as "transit" peptide). Interestingly the identification of several peptides assigned to the N-terminal region of the protein indicates that the presequence has not been degraded. This observation is consistent[10] with the results of previous studies suggesting that the Rieske protein in the mammalian systems is processed in a single proteolytic step after it becomes associated with the cytochrome bc_1 complex and that the processed presequence remains as a subunit of the complex (52).

The strategy for the detection of proteolytic cleavage events, described above, relies on the identification of the N- and C-terminal peptides. However, in shotgun analysis of complex protein mixtures, the protein coverage (the number of identified peptides per protein) is typically low especially in the case of low abundance proteins. Thus, such events would

[10] Validation of the biological significance of this observation would require additional analysis to eliminate the possibility that the cleavage of the protein occurred at the digestion stage because of non-biological reasons, *e.g.* chymotrypsin-like secondary activity of trypsin.

only be detected for a fraction of all proteins, typically those of high abundance. The efficiency of the method can be improved by using targeted protein identification strategies designed to increase the likelihood of identifying N- and C-terminal peptides. One such strategy is based on isolation of N-terminal peptides from *in vivo* N-terminus-blocked proteins using fractional diagonal chromatography (53). The method can be further improved by optimizing the computational MS/MS data interpretation strategies to specifically look for peptides indicative of the proteolytic cleavage (54).

QUANTITATIVE PROTEOMICS

Mass spectrometry is increasingly used not only for the identification of proteins but also for their quantification (quantitative proteomics) (for recent reviews, see Refs. 1 and 55–57). The two problems are interdependent and in fact complementary, *e.g.* the quantitative information can be used to resolve some of the peptide grouping ambiguities.

Although methods are being developed for the determination of absolute protein abundance levels (58–60), most current quantitative proteomic experiments are based on the determination of relative protein expression levels between two or more different pools of proteins. In the most straightforward application, the quantitative proteomics is used as the equivalent of the microarray gene expression profiling approach (61) except that the measurement is performed at the protein, rather than mRNA, level. The shotgun proteomic approach can be made quantitative by applying stable isotope labeling of proteins or peptides. This is illustrated in Fig. 9*A* using the most common case of a two-sample comparison. The compared samples can represent two different cell states (*e.g.* before and after a perturbation) or cells grown under different conditions. The proteins are labeled separately with either light (sample 1) or heavy (sample 2) stable isotopes. The labeling can be done in a number of ways, *e.g.* chemically (ICAT, iTRAQ, etc.) or metabolically (*e.g.* SILAC) (for reviews, see Refs. 1 and 55–57). Proteins from both samples are mixed and enzymatically digested into peptides. Labeled peptides are separated and subjected to sequencing and quantification using mass spectrometry. Peptides are identified from MS/MS spectra as described previously, and the quantitative information is extracted either from MS spectra (*e.g.* in ICAT- or SILAC-based quantitative methods) or directly from MS/MS spectra (iTRAQ) using software tools specifically developed for that purpose (62–65). Quantification is based on measuring relative ion intensity of heavy and light labeled peptide ions. Relative abundances of peptides between the two samples are then combined to compute the relative protein abundances. In addition to global protein profiling experiments, the same quantitative strategy can be used in a targeted way, *e.g.* for distinguishing members of macromolecular complexes or cell organelles from nonspecifically co-purifying proteins (66–69). It should also be mentioned that although the discussion here is centered on the quantitative proteomic approach based on isotopic labeling, it applies equally to semiquantitative methods based on simple peptide counts (70, 71) or on peptide ion current profiling (72, 73).[11]

The relative protein abundance ratios between the compared samples are computed based on the ratios of observed peptides. For a distinct peptide, its relative abundance ratio is a direct measure of the abundance ratio of its corresponding protein.[12] In contrast, the relative abundance ratio in the case of a shared peptide is a weighted average of the abundance ratios of all its corresponding proteins with the weighting factors being determined by the absolute abundance of those proteins in the samples. This is illustrated in Fig. 9*A* where two differentiable proteins, A and B, are inferred to be present in the samples based on the identification of three peptides (proteins C and D are discussed later in this section). In this example, one of the peptides (peptide 2) is shared between the two proteins, and the other two peptides (peptides 1 and 3) are unique to protein A or B, respectively. The relative protein abundance ratios of these proteins, R_A and R_B, can be measured using the relative abundance ratios of their distinct peptides, r_1 and r_3, respectively (see Fig. 9*B*). The relative abundance ratio of the shared peptide 2, r_2, can be anywhere between the protein ratios R_A and R_B depending on the absolute abundances of A and B in both samples that are being compared, N_A, N_B (sample 1) and N_A', N_B' (sample 2).

An example of this kind is shown in Fig. 10. In that experiment, lipid rafts were isolated from both control and stimulated Jurkat human T cells, and the protein samples were quantitatively compared using the ICAT method (27). A number of peptides were identified that are shared between several members of the guanine nucleotide-binding protein (G protein) family, including α inhibiting activity polypeptides 1, 2, and 3. Isotopically labeled peptides for which quantitative information is available (Cys-containing ICAT-labeled peptides) are shown in Fig. 10. The identification of G_i α_3 and α_2 proteins was also supported by several additional unlabeled distinct peptides for which no quantitative information is available (sequences not shown). The quantification of the protein G_i α_3 was based on one distinct ICAT-labeled peptide that was found to be present at higher abundance in the stimulated sample compared with the control sample (relative peptide abundance ratio close to 2:1). At the same time, quantification of G_i α_2 was based on five distinct ICAT-labeled peptides showing no significant difference in their abundances between the two samples (average relative abundance ratio close to 1:1). Thus, although protein quantification

[11] The issues discussed here are relevant to non-mass spectrometry-based protein quantification methods as well. For example, confirmation by Western blots can be equally misleading especially if anti-peptide antibodies are used and the peptide is shared.

[12] This is not entirely correct because peptides can be differentially modified (*e.g.* phosphorylated) under different conditions.

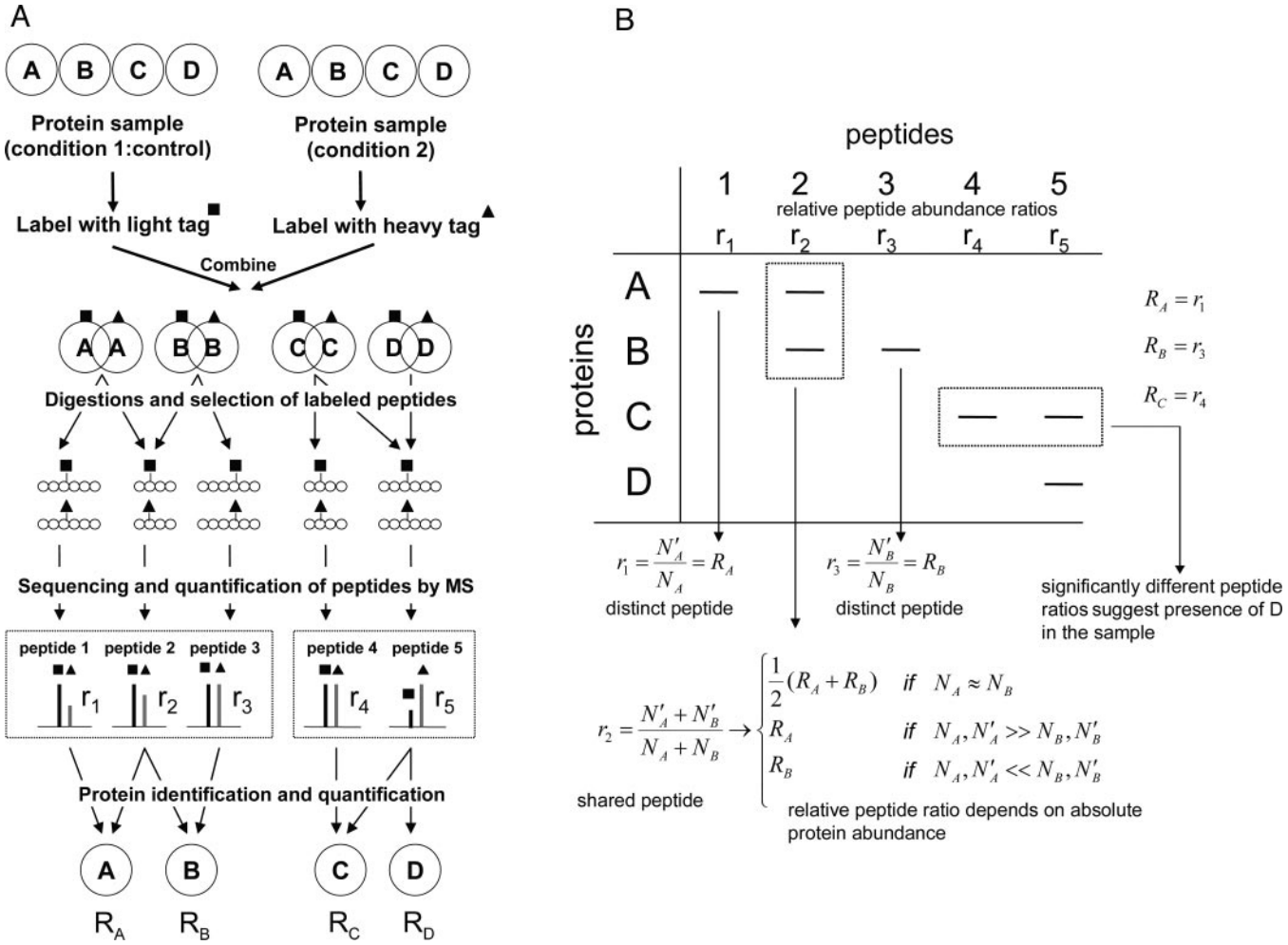


FIG. 9. **Quantitative shotgun proteomic analysis using stable isotopes.** *A*, in one quantitative method (ICAT), proteins are labeled using light or heavy mass tags and then digested into peptides. Labeled peptides are captured and sequenced using tandem mass spectrometry. Peptides are identified from MS/MS spectra using database searching and used to infer which proteins are present in the sample. Relative abundances of peptides between the compared samples are extracted from MS data, and then the relative protein abundance ratios are computed based on the ratios of observed peptides. The relative abundance ratio of a distinct peptide is a direct measure of the abundance ratio of its protein (for peptide 1, protein A; for peptide 3, protein B; and for peptide 4, protein C), whereas it is a weighted average of the abundance ratios of all its corresponding proteins in the case of a shared peptide (peptides 2 and 5). *B*, connection between the relative quantification observed at peptide and protein levels. Distinct peptides 1 and 3 directly measure the relative protein abundance ratios of their corresponding proteins A and B, R_A and R_B. The relative abundance ratio of the shared peptide 2, r_2, can be anywhere between the protein ratios R_A and R_B depending on the absolute abundances of A and B. Quantitative information can be used to resolve some cases of shared peptides. If peptides 4 and 5 have significantly different ratios r_4 and r_5, it can be explained by the presence of protein D is the sample.

based on a single distinct peptide should be interpreted with caution, it appears likely that these two members of the same gene family exhibited a different response to the external stimulation with G_i α_3 being up-regulated and G_i α_2 not changing. Interestingly the relative abundance ratios of the other two ICAT-labeled peptides that were shared between G_i α_3 and G_i α_2 were much closer to that of peptides unique to G_i α_2 (the shared peptides are also present in another member of the gene family, G_i α_1, but no distinct peptides were identified that would suggest the presence of that protein in the sample). This indicates that the absolute abundance level of the protein G_i α_2 was greater than that of G_i α_3 in agreement with a rough protein abundance measure such as the number of matched MS/MS spectra, 79 *versus* 27, determined for G_i α_2 and G_i α_3 proteins, respectively.

Quantitative information can therefore be used to resolve some cases of shared peptides or suggest the presence of multiple protein isoforms having a different biological function. This is again illustrated in Fig. 10 where protein C is identified by peptides 4 and 5 having relative peptide abundance ratios r_4 and r_5, respectively. Peptide 5 is also present in protein D. Because there are no distinct peptides in the dataset that correspond to protein D, it is not possible to conclude that this protein is present in the sample given the

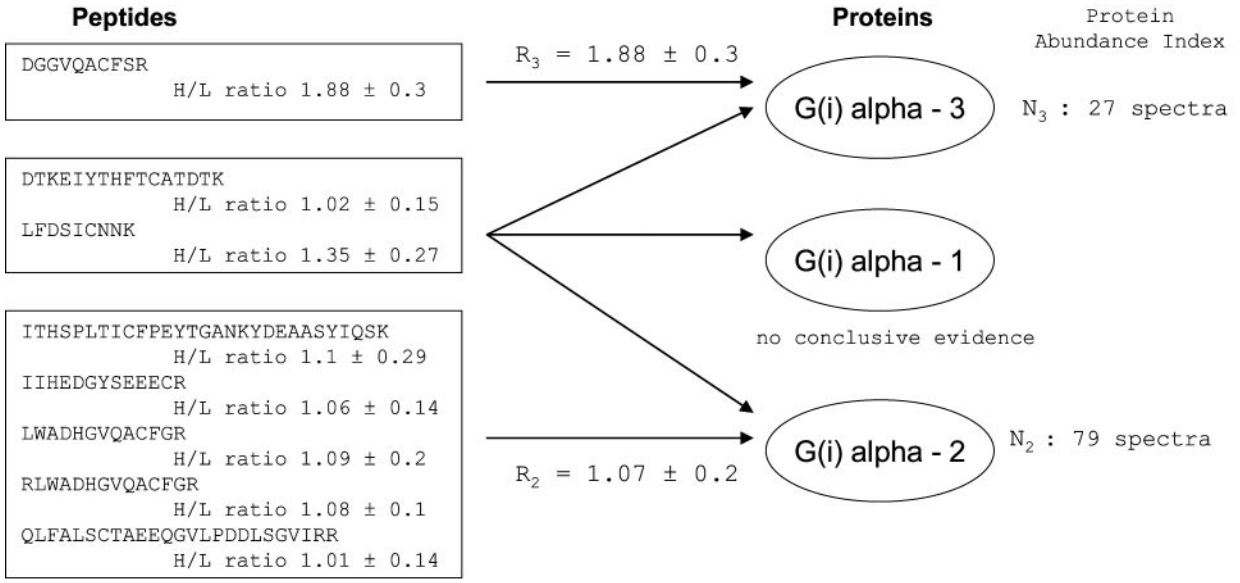


FIG. 10. **Identification and quantification of a group of peptides shared between several members of the guanine nucleotide-binding protein (G protein) family, α inhibiting activity polypeptides 1, 2, and 3 (Swiss-Prot accession numbers P04898, P04899, and P08754).**

sequences of identified peptides alone (subset protein identification). At the same time, if protein D is in the sample, its presence would be reflected in the relative abundance ratio of the shared peptide 5, whereas the relative abundance ratio of the distinct peptide 4 would always be determined solely by the relative abundance of protein C. Thus, significantly different ratios r_4 and r_5 would indicate the presence of protein D in the sample. Note that the reverse is not necessary true, *i.e.* observation of consistent peptide ratios does not rule out the presence of protein D because it can simply reflect a significantly lower abundance level of protein D compared with protein C.

Furthermore observation of peptides with inconsistent relative abundance ratios when all peptides appear to be distinct (according to the protein sequence database used in the analysis) can point to the presence of a novel biologically significant protein form (*e.g.* novel splice variant, product of protein degradation, etc.). One such interesting example has been noticed recently in a quantitative proteomic study concerned with the identification of a human transcription factor using an ICAT proteomic approach (74). Among six identified peptides that were assigned to cellular nucleic acid-binding protein (CNBP), three peptides from the N terminus of the protein had an average relative abundance ratio (enriched sample *versus* control) of less than 3:1, whereas the other three peptides derived from the C-terminal portion of the protein had ratios of more than 7:1. Thus, it has been suggested that two different forms of CNBP (or CNBP and its homologue) are present in the sample. In other cases, inconsistencies in the relative peptide abundance ratio can be due to post-translational modification of the protein, *e.g.* if one of the peptides is phosphorylated and its abundance (in the unmodified form) is different in the compared samples.

A close connection between the problem of assembling peptides into proteins and determining protein abundance ratios suggests a new integrated approach for dealing with quantitative proteomic data. At present, these two tasks are performed separately with the protein ratios computed using peptide ratios and the apportionment of shared peptides among their corresponding proteins performed independently of the quantitative data. Instead the apportionment of shared peptide and creation of the protein summary lists can be made dependent on the quantitative information observed at the peptide and protein level. This should enhance the interpretation of the data by resolving some of the ambiguities discussed above. However, such an approach would require high quality quantitative proteomic data. At present, the accuracy of relative peptide abundance ratios extracted from mass spectra using automated software tools often requires manual validation. This is especially true in the case of peptide "outliers," *i.e.* peptides whose relative abundance ratios are significantly different from the ratios observed for other peptides assigned to the same protein, which are of utmost interest in the context of this discussion. The development of such integrated tools is an imminent task for shotgun proteomics.

INTEGRATION OF PROTEOMIC AND TRANSCRIPTIONAL DATA

Quantitative MS/MS-based proteomic analysis and DNA microarray analysis are two complementary technologies that measure gene expression at the protein and RNA levels, respectively. Due to its technically more advanced stage, the microarray technology (61) allows monitoring of RNA expression levels for the number of genes that is significantly larger than the number of proteins that can be accurately identified and quantified in a typical proteomic experiment, and it can be effectively used for the analysis of alternative splicing and

genome annotation (75, 76). However, due to post-transcriptional regulatory mechanisms such as protein translation, post-translational modifications, and degradation, the microarray measurements of mRNA expression patterns alone are not sufficient for understanding protein expression and function (77, 78). Thus, by combining transcriptional and proteomic analysis of the same samples, it becomes possible to achieve a better understanding of complex biological systems. A number of integrative proteomic and transcriptional analyses have been recently performed, including studies on model organisms and mammalian cells and tissues (79–82). A recent review on the subject of integrating microarray and proteomic data can be found in Ref. 83, and the discussion here will be limited to the issues related to the protein inference problem.

Integration of different data types requires a good understanding of the underlying technologies and their limitations. Although a detailed review of the microarray technology goes beyond the scope of this article, it is interesting to note that many of the difficulties discussed here in the context of quantitative MS-based proteomic experiments are also present in the analysis of gene expression using microarrays. Unlike quantitative shotgun proteomics, in the case of oligonucleotide arrays the sequences of DNA probes present on the array are known in advance (the sequences of peptide "probes" in shotgun proteomics are determined from the spectra). Still ambiguities remain in connecting DNA probes to the target mRNAs (84). For example, multiples probes can map to the same gene; the same probe can map to different products of the same gene or even to multiple genes. Multiple probes mapping to the same gene can produce significantly different expression ratios; outliers might indicate the presence of several alternative splice forms, but they could also be a result of inaccurate quantification (75). Furthermore cross-hybridization, *i.e.* binding of the labeled RNA to non-target homologous probe sequence, introduces additional errors (85).

Integration of proteomic and transcriptional data is hindered by lack of relevant annotations and the use of different accessioning schemes. The information available for each probe present on an Affymetrix chip, for example, includes an arbitrary identification number, the GenBank™ accession number of the target RNA sequence, and brief functional annotation. In the case of MS-based proteomics, experimental MS/MS spectra are assigned peptides, and then peptides are assembled into proteins using a variety of protein sequence databases. Each protein sequence database has its unique accessioning scheme, and the degree of sequence annotation does not always allow easy cross-reference between different protein sequence databases or between protein and genomic sequence databases.

Correlating mRNA and protein data can be facilitated by selecting a well annotated database, *e.g.* UniGene, as a common reference (86) (Fig. 11). The UniGene database is created by an automated partitioning of GenBank™ sequences into a non-redundant set of gene-oriented clusters with each cluster containing sequences representing a unique gene. A number of tools have been described recently that can link the probes from Affymetrix arrays to the UniGene cluster identifiers (87). In turn, MS-derived protein identification datasets can be related to the UniGene clusters using the known connection between the RefSeq protein sequence database and UniGene. A tool for direct mapping of Affymetrix probes to RefSeq sequences has also been described (84).

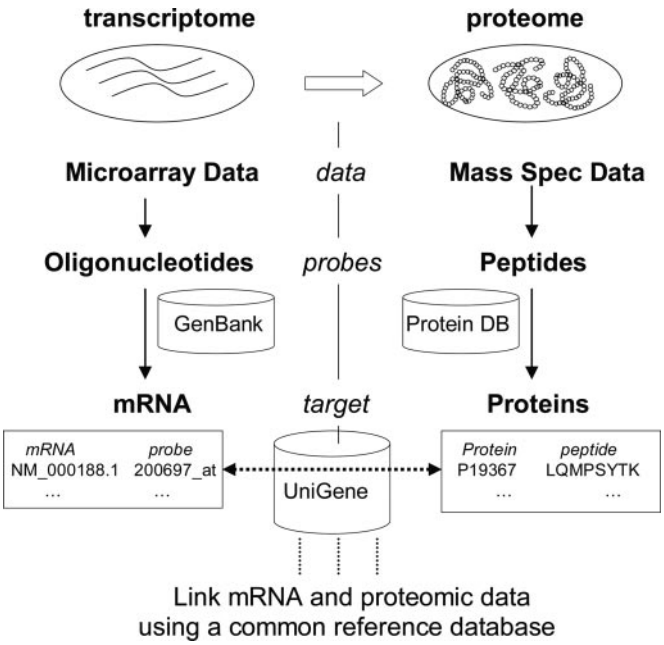


FIG. 11. **Integration of proteomic and transcriptional data.** mRNA and proteomic data can be linked using a common reference database such as UniGene.

Although UniGene and RefSeq can provide a common reference for connecting proteomic and transcriptional data, a one-to-one correspondence will not always be possible. For example, some DNA probes cannot be linked to any UniGene clusters because their target sequences have been removed from the latest version of GenBank™ or deemed to be redundant and excluded from the UniGene build process. Furthermore in many proteomic studies, proteins are identified by searching MS/MS spectra against more complete protein sequence databases than RefSeq, *e.g.* IPI or Entrez Protein. Connecting protein sequences that are not annotated in RefSeq to the UniGene clusters is not straightforward. One of the main difficulties again comes from alternative splice forms. Without the ability to resolve different alternative splice forms, both on the part of proteomic and transcriptional analyses, the association between the two data types is not unique. As a result, the integration and correlation between proteomic and transcriptional data in some cases can be performed only at the gene level with mRNA and protein expression ratios averaged over multiple products of the same gene. Despite these difficulties, integrated analysis of mRNA and protein data can provide very valuable insights into complex biological systems.

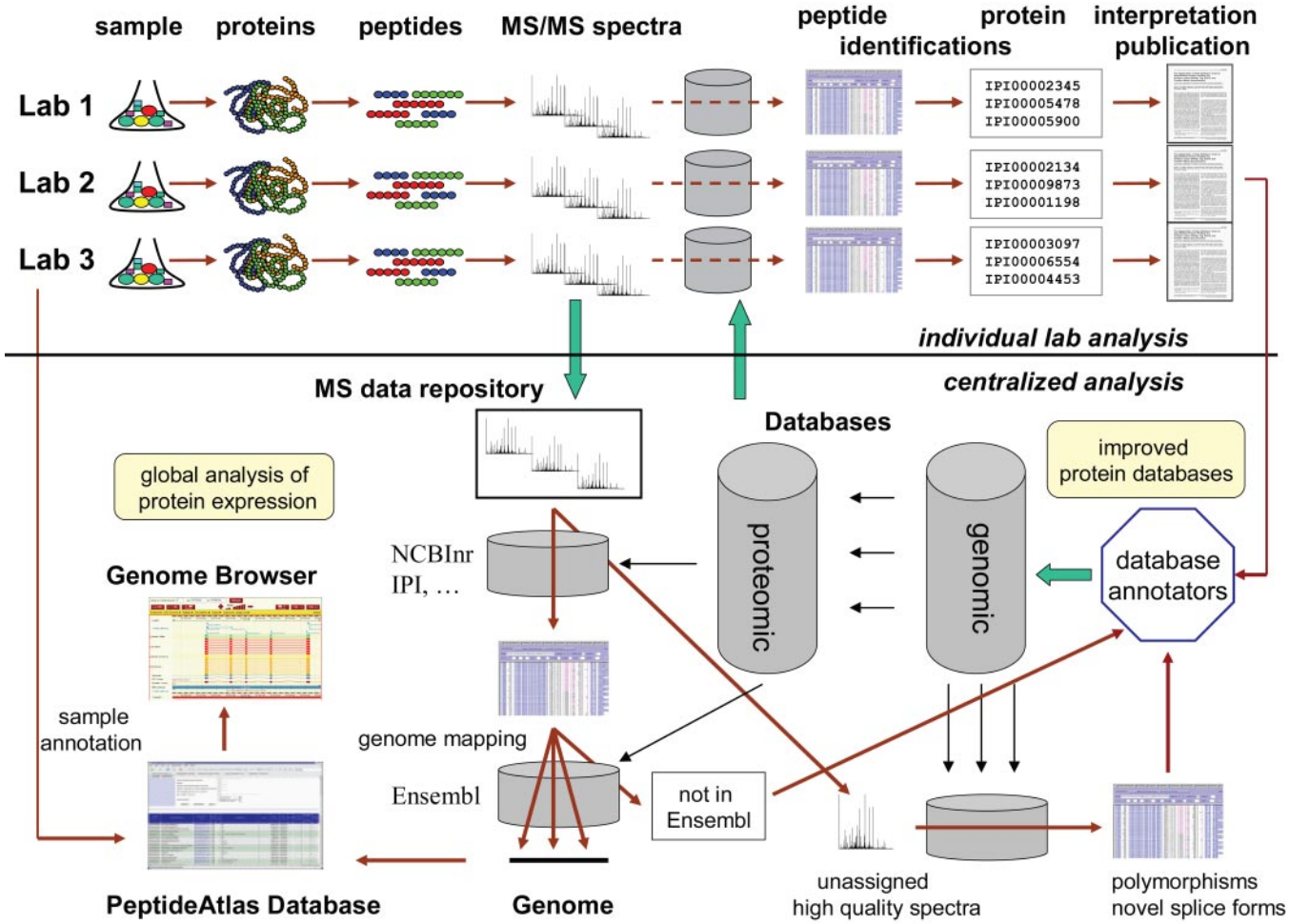


FIG. 12. **Submission of mass spectrometry data to public repositories allows extraction of additional valuable information that otherwise would be missed in the analysis of a single experiment by an individual laboratory.** More comprehensive characterization of the entire proteome can be achieved by combining the data from multiple diverse experiments (different tissues or cell types, enrichment schemes, etc.). In one such example, the PeptideAtlas project, MS/MS datasets from different laboratories are processed using the same high throughput pipeline. Identified peptides are mapped to the genome via the Ensembl gene index. Peptide sequences along with the chromosomal locations, sample annotation, and other information are stored in a relational database. The data can be visualized in the Ensembl genome browser, and the database itself can be mined to study global trends of protein expression. Peptide identification data, if communicated back to the database developers and annotators, can also be used to improve the quality of the protein sequence databases. Reanalysis of high quality MS/MS spectra that are left unassigned in a typical database search against a protein sequence database can lead to the identification of new open reading frames, novel splice forms, and sequence polymorphisms.

INTEGRATION OF MULTIPLE SHOTGUN PROTEOMIC DATASETS AND GENE-CENTERED DATA INTERPRETATION

The discussion so far has been limited to the analysis and interpretation of the data generated in a "single" experiment where all MS/MS data are acquired on a particular biological sample of interest. However, due to technical limitations of current proteomic technologies, in any given large scale proteomic experiment only a subset of the entire proteome is identified. In repeated analysis of the same type, the cumulative number of identified peptides and proteins quickly reaches a saturation point. A more comprehensive characterization of the entire proteome can be achieved by combining the data from multiple diverse experiments (different tissues or cell types, enrichment schemes, etc.) (50, 51, 88, 89). Furthermore performing secondary, centralized analysis of the datasets previously analyzed and published by individual laboratories can uncover interesting global trends not apparent in the analysis of any single dataset alone (Fig. 12).

The task of combining and comparing multiple large scale datasets generated using different biological samples (*e.g.* different cell states or tissues) requires the development of new approaches and computational tools. Due to the peptide-centric nature of shotgun proteomics, diverse datasets (from the same organism) can be best combined at the peptide level by linking the sequences of the identified peptides to a common gene index. One such approach, based on the mapping of peptides observed in a large group of proteomic experiments to the Ensembl genome, has been described recently

(51) and implemented in a public resource, PeptideAtlas (www.peptideatlas.org). In this approach, peptide identifications passing a certain probability threshold are matched to proteins in the Ensembl database. The chromosomal coordinates, or multiple sets of coordinates in the case of peptides matching to more than one gene, and Ensembl protein accession numbers are retrieved for all matched peptides. The results are stored in a relational database and can be visualized using the Ensembl genome browser Distributed Annotation System (DAS) (90).

By connecting the sequences of the identified peptides with the genome, PeptideAtlas allows gene-centered interpretation of the results of shotgun proteomic experiments in line with previous suggestions (26). Most of the identified peptides have a unique association with the genome, and in those cases it is possible to state with certainty that a product of a certain gene has been identified. Some peptides map to several different locations on the genome due to the presence of gene paralogues, repeated protein domains, or simple sequence redundancy. PeptideAtlas database, and other emerging repositories of this kind (91), should be useful for validation and improved annotation of the human genome by complementing other types of data currently used for that purpose, such as mRNA and EST data, with large scale proteomic data.

In turn, peptide identification data can be used to improve the quality of the protein sequence databases by making them more complete and accurate. For example, the identification of a peptide from a certain protein by searching MS/MS spectra against a protein sequence database would ensure that the sequence of that protein does not disappear from the database in the future (a situation not that uncommon at present). The use of mass spectrometry data can be further extended to go beyond genome validation (confirming the proteins already present in the current sequence databases) to the discovery of novel gene products and variants. For example, high quality MS/MS spectra that are left unassigned when searched against protein sequence databases such as IPI could be reanalyzed more comprehensively by searching genomic databases with the purpose to discover open reading frames missed by the current gene prediction programs, novel splice forms, or sequence polymorphisms.

The information stored in PeptideAtlas, which includes experimental conditions and the type of cell or tissue analyzed, could also be used to statistically explore global trends of differential protein expression. Similar to the method of measuring gene expression using the number of corresponding expressed sequence tags in EST databases (92–94), the correlation between splice forms and disease states or tissue types can potentially be investigated at the level of proteins using, *e.g.* MS/MS spectrum counts as a rough measure of protein abundance. It can also be used to study the correlation between the physical-chemical properties of peptides and the likelihood of them being detected by a mass spectrometer (60). It can be anticipated that eventually the computational tools (*e.g.* MS/MS database search tools or the tools for assembling peptides into proteins) will not treat all peptides equally but will use a weighting scheme to account for the probability of detecting a peptide. It will also be useful for selecting synthetic peptides for the absolute protein quantification using mass spectrometry or peptide arrays. The ability to perform these different analyses could make a significant contribution to our understanding of complex biological systems, thus significantly enhancing the overall value of shotgun proteomics.

CONCLUDING REMARKS

Shotgun proteomic technology has matured to a point where it can be used for routine identification and, when coupled with stable isotope labeling, accurate relative quantification of thousands of peptides in a single experiment. A significant effort has been made in recent years to improve various aspects of the technology, including extensive work on developing computational tools for identifying peptides from MS/MS spectra. It has also been recognized that the analysis of large scale shotgun proteomic datasets requires the application of transparent and tested statistical tools to estimate the confidence measures of peptide and protein identifications and to estimate false identification error rates in the published data. At the same time, significant inconsistencies still exist in how the information derived at the peptide level can be used to draw conclusions regarding the identities and quantities of the sample proteins and how the resulting protein identifications are interpreted in a biological context and published in the literature.

The peptide-centric nature of shotgun proteomics becomes apparent in the analysis of data acquired on higher eukaryote organisms where a significant fraction of identified peptides can be assigned to more than one entry in the protein sequence database. In the best case scenario, seldom observed in shotgun proteomics, the sequences of the identified peptides would allow fairly complete characterization of the corresponding mature protein form expressed in the sample. This necessarily requires very high protein sequence coverage, including identification of the N-terminal peptide, and determination of the type and location of any post-translational modification. However, the identification of N-terminal peptides is not always possible (especially without specific enrichment for those peptides), and identification of post-translational modifications or sequence polymorphisms is also difficult. Thus, more often, observed peptide data allow identification of a certain protein but not accurate characterization of its mature form. In many cases, the sequences of the identified peptides would not be sufficient to allow differentiation between two or more splice forms of a particular gene. Furthermore in some cases it would only be possible to state with certainty that one or more members of a particular protein family are identified but not to single out any of them. The examples and discussion presented in this article should as-

sist in the data interpretation process by providing general guidelines and a nomenclature for describing all these various protein identification scenarios. It is also hoped that this discussion will contribute to the development of more formal guidelines for publishing protein identification datasets obtained using shotgun proteomic strategy in the literature. Furthermore efforts are currently underway to develop common standards and schema for the representation, interchange, and storage of the results of proteomic experiments (95–97). The issues discussed here should be taken into consideration in developing such standards.

Understanding of these data interpretation difficulties is helpful for deciding upon what experimental strategy is most appropriate given the aims of each particular experiment as well as for the development of new experimental and computational approaches. Higher protein coverage, which leads to improved ability to differentiate between protein isoforms and to identify sites of post-translational modifications, in some cases can be achieved by using multiple digestion enzymes (98, 99). Another strategy is based on generation of synthetic peptides that are unique to a particular isoform of interest (59, 60, 100). Such peptides can be selected (using computational approaches maximizing the likelihood of those peptides being detected by a mass spectrometer), synthesized, isotopically labeled, and spiked into an isotopically labeled biological sample. This should allow selective sequencing and quantification of peptides that are unique to the proteins of interest. It has also been suggested that peptide sequencing can be performed in two stages. After peptides are identified from the MS/MS spectra acquired at the first stage, indistinguishable sequence database entries are aligned, and peptides that discriminate between different isoforms are predicted from unique stretches. The second stage of the MS/MS sequencing process would then be directed toward the analysis of those predicted distinct peptides (26).

The shotgun proteomic approach alone does not appear to be sufficient to comprehensively and unambiguously characterize the proteome. Complementary to shotgun proteomics (often called the "bottom-up" proteomic approach), the "top-down" proteomic approaches that deal with intact proteins (or involve extensive protein separation prior to digestion) offer certain advantages with regard to the discrimination between protein isoforms and characterization of post-translational modifications. The most established of these methods are based on 2D gels (7, 31–33). However, gel-based methods have known limitations such as low detection sensitivity, bias toward high abundance proteins, and difficulty in resolving internal membrane or basic proteins. Non-gel-based multidimensional protein separation methods are being developed and can circumvent some of these limitations (101–104). Another promising top-down protein characterization technique is based on MS/MS sequencing of intact proteins (5, 6, 105). Although still not at the level of automation and data throughput currently achievable in shotgun proteomics, this technology has experienced significant advances in the last few years. An attractive approach is to integrate the measurements performed on the same systems both at the level of peptides and intact proteins (101, 102, 106–109). In this method, the molecular weights of intact proteins are measured using high mass accuracy instruments such as ESI-FTMS or ESI-TOF. In parallel, proteins are digested, and peptides are sequenced using a typical shotgun proteomics set-up and/or using a peptide fingerprinting method. The advantage of this approach (in the context of the protein inference problem) is that the process of assembling peptides into proteins and discrimination between protein isoforms can be assisted by the knowledge of the molecular weights of the sample proteins. Other proposed approaches include generation of isoform-specific affinity ligands such as antibodies or peptides for selective targeting of proteins of interest (110, 111). Additional insights can be obtained by integrating measurements performed on the same biological systems but at different levels, *e.g.* proteomic and transcriptional measurements. The knowledge available from microarray experiments regarding the presence or absence of a certain mRNA transcript can assist in the process of assigning peptides to the corresponding protein isoforms observed in proteomic experiments.

It has been stressed already that shotgun proteomic datasets should be analyzed using transparent computational tools that are well documented and made generally available to the scientific community. Even when this is the case, however, publication of long lists of protein and peptide identifications by itself has only a limited value. As databases become updated, new protein sequences are added, some sequences are removed, and annotations or accession schemes change, those lists become obsolete and can no longer be easily interpreted or correlated with other data. Thus, the authors should be encouraged to provide access to all raw data, or at least to MS/MS spectra, as a part of the publication. This would allow re-evaluation of the primary MS data using the most up-to-date protein sequence databases. Coupled with the development of open MS data formats (96, 97), centralized data repositories (41, 51, 91, 112), and infrastructure for processing and integrating datasets from different experiments (41, 51, 91), this would allow new uses of proteomic data such as validation of genes that are expressed on the protein level or elucidation of global protein expression patterns that would otherwise be missed in an analysis of a single experiment. Finally if communicated back to the database developers and annotators, MS-derived proteomic data could become a useful resource in the process of annotating the genomes of the corresponding organisms.

Acknowledgments—We acknowledge fruitful discussions with Karl Clauser, Frank Desiere, Eric Deutsch, Jimmy Eng, Anne-Claude Gingras, Andrew Keller, Jeff Kowalack, Xiao-jun Li, Parag Mallick, Jeff Ranish, Katheryn Resing, Julian Watts, and Bernd Wollscheid. We are particularly grateful to Anne-Claude Gingras, Jeff Ranish, and Bernd

Wollscheid for reading the manuscript and to Nichole King and James Eddes for help with Table I and Fig. 7.

* This work was funded in part with federal funds from the NHLBI, National Institutes of Health under Contract Number N01-HV-28179. 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.

§ To whom correspondence should be addressed: Inst. for Systems Biology, 1441 N. 34th St., Seattle, WA 98103. Tel.: 206-732-1245; Fax: 206-732-1299; E-mail: nesvi@systemsbiology.org.

REFERENCES

1. Aebersold, R., and Mann, M. (2003) Mass spectrometry-based proteomics. *Nature* **422,** 198–207
2. Link, A. J., Eng, J., Schieltz, D. M., Carmack, E., Mize, G. J., Morris, D. R., Garvik, B. M., and Yates, J. R. (1999) Direct analysis of protein complexes using mass spectrometry. *Nat. Biotechnol.* **17,** 676–682
3. 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
4. Washburn, M. P., Wolters, D., and Yates, J. R. (2001) Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat. Biotechnol.* **19,** 242–247
5. Reid, G. E., and McLuckey, S. A. (2002) "Top down" protein characterization via tandem mass spectrometry. *J. Mass Spectrom.* **37,** 663–675
6. Meng, F., Forbes, A. J., Miller, L. M., and Kelleher, N. L. (2005) Detection and localization of protein modifications by high resolution tandem mass spectrometry. *Mass Spectrom. Rev.* **24,** 126–134
7. Gorg, A., Weiss, W., and Dunn, M. J. (2004) Current two-dimensional electrophoresis technology for proteomics. *Proteomics* **4,** 3665–3685
8. Patterson, S. D. (2003) Data analysis—the Achilles heel of proteomics. *Nat. Biotechnol.* **21,** 221–222
9. Boguski, M. S., and McIntosh, M. W. (2003) Biomedical informatics for proteomics. *Nature* **422,** 233–237
10. Nesvizhskii, A. I., and Aebersold, R. (2004) Analysis, statistical validation and dissemination of large-scale proteomics datasets generated by tandem MS. *Drug Discov. Today* **9,** 173–181
11. Johnson, R. S., Davis, M. T., Taylor, J. A., and Patterson, S. D. (2005) Informatics for protein identification by mass spectrometry. *Methods* **35,** 223–236
12. Russell, S. A., Old, W., Resing, K. A., and Hunter, L. (2004) Proteomic informatics. *Int. Rev. Neurobiol.* **61,** 129–157
13. Baldwin, M. A. (2004) Protein identification by mass spectrometry: issues to be considered. *Mol. Cell. Proteomics* **3,** 1–9
14. Eng, J. K., McCormack, A. L., and Yates, J. R. (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
15. Mann, M., and Wilm, M. (1994) Error-tolerant identification of peptides in sequence databases by peptide sequence tags. *Anal. Chem.* **66,** 4390–4399
16. Perkins, D. N., Pappin, D. J., Creasy D. M., and Cottrell, J.C. (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20,** 3551–3567
17. Clauser, K. R., Baker, P., and Burlingame, A. L. (1999) Role of accurate mass measurement (±10 ppm) in protein identification strategies employing MS or MS/MS and database searching. *Anal. Chem.* **71,** 2871–2882
18. Field, H. I., Fenyo, D., and Beavis, R. C. (2002) RADARS, a bioinformatics solution that automates proteome mass spectral analysis, optimizes protein identification, and archives data in a relational database. *Proteomics* **2,** 36–47
19. Craig, R., and Beavis R. C. (2004) TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* **20,** 1466–1467
20. Geer, L. Y., Markey, S. P., Kowalak, J. A., Wagner, L., Xu, M., Maynard, D. M., Yang, X., Shi, W., and Bryant, S. H. (2004) Open mass spectrometry search algorithm. *J. Proteome Res.* **3,** 958–964
21. Keller, A., Nesvizhskii, A. I., Kolker, E., and Aebersold, R. (2002) Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal. Chem.* **74,** 5383–5392
22. 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
23. Fenyo, D., and Beavis, R. C. (2003) A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes. *Anal Chem.* **75,** 768–774
24. Peng, J., Elias, J. E., Thoreen, C. C., Licklider, L. J., and Gygi, S. P. (2003) Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large scale protein analysis: the yeast proteome. *J. Proteome Res.* **2,** 43–50
25. Sadygov, R. G., and Yates, J. R. (2003) A hypergeometric probability model for protein identification and validation using tandem mass spectral data and protein sequence databases. *Anal. Chem.* **75,** 3792–3798
26. Rappsilber, J., and Mann, M. (2002) What does it mean to identify a protein in proteomics? *Trends Biochem. Sci.* **27,** 74–78
27. Von Haller, P. D., Yi, E., Donohoe, S., Vaughn, K., Keller, A., Nesvizhskii, A. I., Eng, J., Li, X. J., Goodlett, D. R., Aebersold, R., and Watts, J. D. (2003) The application of new software tools to quantitative protein profiling via ICAT and tandem mass spectrometry: II. Evaluation of tandem mass spectrometry methodologies for large-scale protein analysis and the application of statistical tools for data analysis and interpretation. *Mol. Cell. Proteomics* **2,** 428–442
28. Resing, K. A., Meyer-Arendt, K., Mendoza, A. M., Aveline-Wolf, L. D., Jonscher, K. R., Pierce, K. G., Old, W. M., Cheung, H. T., Russell, S., Wattawa, J. L., Goehle, G. R., Knight, R. D., and Ahn, N. G. (2004) Improving reproducibility and sensitivity in identifying human proteins by shotgun proteomics. *Anal. Chem.* **76,** 3556–3568
29. Carr, S., Aebersold, R., Baldwin, M., Burlingame, A., Clauser, K., and Nesvizhskii, A. (2004) The need for guidelines in publication of peptide and protein identification data. *Mol. Cell. Proteomics* **3,** 531–533
30. Yang, X., Dondeti, V., Dezube, R., Maynard, D. M., Geer, L. Y., Epstein, J., Chen, X., Markey, S. P., Kowalak, J. A. (2004) DBParser: web-based software for shotgun proteomic data analyses. *J. Proteome Res.* **3,** 1002–1008
31. Pedersen, S. K., Harry, J. L., Sebastian, L., Baker, J., Traini, M. D., McCarthy, J. T., Manoharan, A., Wilkins, M. R., Gooley, A. A., Righetti, P. G., Packer, N. H., Williams, K. L., and Herbert, B. R. (2003) Unseen proteome: mining below the tip of the iceberg to find low abundance and membrane proteins. *J. Proteome Res.* **2,** 303–311
32. Fung, K. Y., Glode, L. M., Green, S., and Duncan, M. W. (2004) A comprehensive characterization of the peptide and protein constituents of human seminal fluid. *Prostate* **61,** 171–181
33. Godovac-Zimmermann, J., Kleiner, O., Brown, L. L., and Drukier, A. L. (2005) Perspectives in splicing up proteomics with splicing. *Proteomics* **5,** 699–709
34. Black, D. L. (2000) Protein diversity from alternative splicing: a challenge for bioinformatics and post-genome biology. *Cell* **103,** 367–370
35. Delalande, F., Carapito, C., Brizard, J. P., Brigidou, C., and Dorsselaer, A. V. (2005) Multigenic families and proteomics: Extended protein characterization as a tool for paralog gene identification. *Proteomics* **5,** 450–460
36. Sam-Yellowe, T. Y., Florens, L., Johnson, J. R., Wang, T., Drazba, J. A., Le Roch, K. G., Zhou, Y., Batalov, S., Carucci, D. J., Winzeler, E. A., and Yates, J. R. (2004) A Plasmodium gene family encoding Maurer's cleft membrane proteins: structural properties and expression profiling. *Genome Res.* **14,** 1052–1059
37. Kislinger, T., Rahman, K., Radulovic, D., Cox, B., Rossant, J., and Emili, A. (2003) PRISM, a generic large scale proteomic investigation strategy for mammals. *Mol. Cell. Proteomics* **2,** 96–106
38. Kristensen, D. B., Brond, J. C., Nielsen, P. A., Andersen, J. R., Sorensen, O. T., Jorgensen, V., Budin, K., Matthiesen, J., Veno, P., Jespersen, H. M., Ahrens, C. H., Schandorff, S., Ruhoff, P. T., Wisniewski, J. R., Bennett, K. L., and Podtelejnikov, A. V. (2004) Experimental Peptide Identification Repository (EPIR): an integrated peptide-centric platform for validation and mining of tandem mass spectrometry data. *Mol. Cell. Proteomics.* **3,** 1023–1038
39. Tabb, D. L., McDonald, W. H., and Yates, J. R. (2002) DTASelect and Contrast: tools for assembling and comparing protein identifications from shotgun proteomics. *J. Proteome Res.* **1,** 21–26
40. Allet, N., Barrillat, N., Baussant, T., Boiteau, C., Botti, P., Bougueleret, L.,

Budin, N., Canet, D., Carraud, S., Chiappe, D., Christmann, N., Colinge, J., Cusin, I., Dafflon, N., Depresle, B., Fasso, I., Frauchiger, P., Gaertner, H., Gleizes, A., Gonzalez-Couto, E., Jeandenans, C., Karmime, A., Kowall, T., Lagache, S., Mahe, E., Masselot, A., Mattou, H., Moniatte, M., Niknejad, A., Paolini, M., Perret, F., Pinaud, N., Ranno, F., Raimondi, S., Reffas, S., Regamey, P. O., Rey, P. A., Rodriguez-Tome, P., Rose, K., Rossellat, G., Saudrais, C., Schmidt, C., Villain, M., and Zwahlen, C. (2004) *In vitro* and *in silico* processes to identify differentially expressed proteins. *Proteomics* **4,** 2333–2351

41. Martens, L., Hermjakob, H., Jones, P., Adamski, M., Taylor, C. F., States, D., Gevaert, K., Vandekerckhove, J., and Apweiler, R. (2005) PRIDE: the proteomics identifications database. *Proteomics*, in press
42. Apweiler, R., Bairoch, A., and Wu, C. H. (2004) Protein sequence databases. *Curr. Opin. Chem. Biol.* **8,** 76–80
43. Wheeler, D. L., Church, D. M., Edgar, R., Federhem, S., Helmberg, W., Madden, T. L., Pontius, J. U., Schuler, G. D., Schriml, L. M., Sequeira, E., Suzek, T. O., Tatusova, T. A., and Wagner, L. (2004) Database resources of the National Center for Biotechnology Information: update. *Nucleic Acids Res.* **32,** D35–D40
44. Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M., Estreicher, A., Gasteiger, E., Martin, M. J., Michoud, K., O'Donovan, C., Phan, I., Pilbout, S., and Schneider, M. (2003) The Swiss-Prot protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.* **31,** 365–370
45. Pruitt, K. D., Tatusova, T., and Maglott, D. R. (2005) NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* **33,** D501–D504
46. Kersey, P. J., Duarte, J., Williams, A., Karavidopoulou, Y., Birney, E., and Apweiler, R. (2004) The International Protein Index: an integrated database for proteomics experiments. *Proteomics* **4,** 1985–1988
47. Birney, E., Andrews, D., Bevan, P., Caccamo, M., Cameron, G., Chen, Y., Clarke, L., Coates, G., Cox, T., Cuff, J., Curwen, V., Cutts, T., Down, T., Durbin, R., Eyras, E., Fernandez-Suarez, X. M., Gane, P., Gibbins, B., Gilbert, J., Hammond, M., Hotz, H., Iyer, V., Kahari, A., Jekosch, K., Kasprzyk, A., Keefe, D., Keenan, S., Lehvaslaiho, H., McVicker, G., Melsopp, C., Meidl, P., Mongin, E., Pettett, R., Potter, S., Proctor, G., Rae, M., Searle, S., Slater, G., Smedley, D., Smith, J., Spooner, W., Stabenau, A., Stalker, J., Storey, R., Ureta-Vidal, A., Woodwark, C., Clamp, M., and Hubbard, T. (2004) Ensembl 2004. *Nucleic Acids Res.* **32,** D468–D470
48. Kuster, B., Mortensen, P., Andersen, J. S., and Mann, M. (2001) Mass spectrometry allows direct identification of proteins in large genomes. *Proteomics* **1,** 641–650
49. Choudhary, J. S., Blackstock, W. P., Creasy, D. M., and Cottrell, J. S. (2001) Interrogating the human genome using uninterpreted mass spectrometry data. *Proteomics* **1,** 651–667
50. Mann, M., and Pandey, A. (2001) Use of mass spectrometry-derived data to annotate nucleotide and protein sequence databases. *Trends Biochem. Sci.* **26,** 54–60
51. Desiere, F., Deutsch, E. W., Nesvizhskii, A. I., Mallick, P., King, N. L., Eng, J. K., Aderem, A., Boyle, R., Brunner, E., Donohoe, S., Fausto, N., Hafen, E., Hood, L., Katze, M. G., Kennedy, K. A., Kregenow, F., Lee, H., Lin, B., Martin, D., Ranish, J. A., Rawlings, D. J., Samelson, L. E., Shiio, Y., Watts, J. D., Wollscheid, B., Wright, M. E., Yan, W., Yang, L., Yi, E. C., Zhang, H., and Aebersold, R. (2005) Integration with the human genome of peptide sequences obtained by high-throughput mass spectrometry. *Genome Biol.* **6,** R5
52. Brandt, U., Yu, L., Yu, C. A., and Trumpower, B. L. (1993) The mitochondrial targeting presequence of the Rieske iron-sulfur protein is processed in a single step after insertion into the cytochrome bc_1 complex in mammals and retained as a subunit in the complex. *J. Biol. Chem.* **268,** 8387–8390
53. Gevaert, K., Goethais, M., Martens, L., Van Damme, J., Staes, A., Thomas, G. R., and Vandekerckhove, J. (2003) Exploring proteomes and analyzing protein processing by mass spectrometric identification of sorted N-terminal peptides. *Nat. Biotechnol.* **21,** 566–569
54. Song, H., Hecimovic, S., Goate, A., Hsu, F. F., Bao, S., Vidavsky, I., Ramanadham, S., and Turk, J. (2004) Characterization of N-terminal processing of group VIA phospholipase A2 and of potential cleavage sites of amyloid precursor protein constructs by automated identification of signature peptides in LC/MS/MS analyses of proteolytic digests. *J. Am. Soc. Mass Spectrom.* **15,** 1780–1793
55. Zhang, H., Yan, W., and Aebersold, R. (2004) Chemical probes and tandem mass spectrometry: a strategy for the quantitative analysis of proteomes and subproteomes. *Curr. Opin. Chem. Biol.* **8,** 66–75
56. Julka, S., and Regnier, F. (2004) Quantification in proteomics through stable isotope coding: a review. *J. Proteome Res.* **3,** 350–363
57. Goshe, M. B., and Smith, R. D. (2003) Stable isotope-coded proteomic mass spectrometry. *Curr. Opin. Biotechnol.* **14,** 101–109
58. Gerber, S. A., Rush, J., Stemman, O., Kirschner, M. W., and Gygi, S. P. (2003) Absolute quantification of proteins and phosphoproteins from cell lysates by tandem MS. *Proc. Natl. Acad. Sci. U. S. A.* **100,** 6940–6945
59. Aebersold, R. (2003) Constellations in a cellular universe. *Nature* **422,** 115–116
60. Kuster, B., Schirle, M., Mallick, P., and Aebersold, R. (2005) *Nat. Rev. Mol. Cell. Biol.* **6,** 577–583
61. Schena, M. (2003) *Microarray Analysis*, Wiley-Liss, Hoboken, NJ
62. Han, D. K., Eng, J., Zhou, H., and Aebersold, R. (2001) Quantitative profiling of differentiation-induced microsomal proteins using isotope-coded affinity tags and mass spectrometry. *Nat. Biotechnol.* **19,** 946–951
63. Li, X. J., Zhang, H., Ranish, J. A., and Aebersold, R. (2003) Automated statistical analysis of protein abundance ratios from data generated by stable-isotope dilution and tandem mass spectrometry. *Anal Chem.* **75,** 6648–6657
64. MacCoss, M. J., Wu, C. C., Liu, H., Sadygov, R., and Yates, J. R. (2003) A correlation algorithm for the automated quantitative analysis of shotgun proteomics data. *Anal. Chem.* **75,** 6912–6921
65. Halligan, B. D., Slyper, R. Y., Twigger, S. N., Hicks, W., Olivier, M., and Greene, A. S. (2005) ZoomQuant: an application for the quantitation of stable isotope labeled peptides. *J. Am. Soc. Mass Spectrom.* **16,** 302–306
66. Ranish, J. A., Yi, E. C., Leslie, D. M., Purvine, S. O., Goodlett, D. R., Eng, J., and Aebersold, R. (2003) The study of macromolecular complexes by quantitative proteomics. *Nat. Genet.* **33,** 349–355
67. Foster, L. J., De Hoog, C. L., and Mann, M. (2003) Unbiased quantitative proteomics of lipid rafts reveals high specificity for signaling factors. *Proc. Natl. Acad. Sci. U. S. A.* **100,** 5813–5818
68. Marelli, M., Smith, J. J., Jung, S., Yi, E., Nesvizhskii, A. I., Christmas, R. H., Saleem, R. A., Tam, Y. Y. C., Faragasanu, A., Goodlett, D. R., Aebersold, R., Rachubinski, R. A., and Aitchison, J. D. (2004) Quantitative mass spectrometry reveals a role for the GTPase Rho1p in actin organization on the peroxisome membrane. *J. Cell Biol.* **167,** 1099–1112
69. Gingras, A. C., Aebersold, R., and Raught, B. (2005) Advances in protein complex analysis using mass spectrometry. *J. Physiol.* **563,** 11–21
70. Liu, H., Sadygov, R. G., and Yates, J. R. (2004) A model for random sampling and estimation of relative protein abundances in shotgun proteomics. *Anal. Chem.* **76,** 4193–4201
71. Blondeau, F., Ritter, B., Allaire, P. D., Wasiak, S., Girard, M., Hussain, N. K., Angers, A., Legendre-Guillemin, V., Roy, L., Boismenu, D., Kearney, R. E., Bell, A. W., Bergeron, J. J., and McPherson, P. S. (2004) Tandem MS analysis of brain clathrin-coated vesicles reveals their critical involvement in synaptic vesicle recycling. *Proc. Natl. Acad. Sci. U. S. A.* **101,** 3833–3838
72. Chelius, D., and Bondarenko, P. V. (2002) Quantitative profiling of proteins in complex mixtures using liquid chromatography and mass spectrometry. *J. Proteome Res.* **1,** 317–323
73. Wang, W., Zhou, H., Lin, H., Roy, S., Shaler, T. A., Hill, L. R., Norton, S., Kumar, P., Anderle, M., and Becker, C. H. (2003) Quantification of proteins and metabolites by mass spectrometry without isotopic labeling or spiked standards. *Anal. Chem.* **75,** 4818–4826
74. Himeda, C. L., Ranish, J. A., Angello, J. C., Maire, P., Aebersold, R., and Hauschka, S. D. (2004) Quantitative proteomics identification of Six4 as the Trex-binding factor in the muscle creatine kinase enhancer. *Mol. Cell. Biol.* **24,** 2132–2143
75. Lee, C., and Roy, M. (2004) Analysis of alternative splicing with microarrays: successes and challenges. *Genome Biol.* **5,** 231
76. Johnson, J. M., Edwards, S., Shoemaker, D., and Schadt, E. E. (2005) Dark matter in the genome: evidence of widespread transcription detected by microarray tiling experiments. *Trends Genet.* **21,** 93–102
77. Gygi, S. P., Rochon, Y., Franza, B. R., and Aebersold, R. (1999) Correlation

between protein and mRNA abundance in yeast. *Mol. Cell. Biol.* **19,** 1720–1730
78. Chen, G., Gharib, T. G., Huang, C. C., Taylor, J. M., Misek, D. E., Kardia, S. L. R., Giodano, T. J., Iannettoni, M. D., Orringer, M. B., Hanash, S. M., and Beer, D. G. (2002) Discordant protein and mRNA expression in lung adenocarcinomas. *Mol. Cell. Proteomics* **1,** 304–313
79. Griffin, T. J., Gygi, S. P., Ideker, T., Rist, B., Eng, J., Hood, L., and Aebersold, R. (2002) Complementary profiling of gene expression at the transcriptome and proteome levels in *Saccharomyces cerevisiae*. *Mol. Cell. Proteomics* **1,** 323–333
80. Tian, Q., Stepaniants, S., Mao, M., Weng, L., Feetham, M. C., Doyle, M. J., Yi, Y. C., Dai, H., Thorsson, V., Eng, J., Goodlett, D., Berger, J. P., Gunter, B., Linseley, P. S., Stoughton, R. B., Aebersold, R., Collins, S. J., Hanlon, W. A., and Hood, L. E. (2004) Integrated genomic and proteomics analyses of gene expression in mammalian cells. *Mol. Cell. Proteomics* **3,** 960–969
81. McRedmond, J. P., Park, S. D., Reilly, D. F., Coppinger, J. A., Maguire, P. B., Shields, D. C., and Fitzgerald, D. J. (2003) Integration of proteomics and genomics in platelets: a profile of platelet proteins and platelet-specific genes. *Mol. Cell. Proteomics* **3,** 133–144
82. Maziarz, M., Chung, C., Drucker, D. J., and Emili, A. (2005) Integrating global proteomics and genomic expression profiles generated from islet α cells: opportunities and challenges to deriving reliable biological inferences. *Mol. Cell. Proteomics* **4,** 458–474
83. Cox, B., Kislinger, T., and Emili, A. (2005) Integrating gene and protein expression data: pattern analysis and profile mining. *Methods* **35,** 303–314
84. Gautier, L., Mooller, M., Friis-Hansen, L., and Knudsen, S. (2004) Alternative mapping of probes to genes for Affymetrix chips. *BMC Bioinformatics* **5:**111
85. Flikka, K., Yadetie, F., Laegreid, A., and Jonassen, I. (2004) XHM: a system for detection of potential cross-hybridizations in DNA microarrays. *BMC Bioinformatics* **5:**117
86. Pontius, J. U., Wagner, L., and Schuler, G. D. (2003) UniGene: a unified view of the transcriptome, in *The NCBI Handbook*, pp. 1–12, National Center for Biotechnology Information, Bethesda, MD
87. Liu, G., Loraine, A. E., Shigeta, R., Cline, M., Cheng, J., Valmeekam, V., Sun, S., Kulp, D., and Siani-Rose, M. A. (2003) NetAffx: Affymetrix probesets and annotations. *Nucleic Acids Res.* **31,** 82–86
88. McGowan, S. J., Terrett, J., Brown, C. G., Adam, P. J., Aldridge, L., Allen, J. C., Amess, B., Andrews, K. A., Barnes, M., Barnwell, D. E., Berry, J., Bird, H., Boyd, R. S., Broughton, M. J., Brown, A., Bruce, J. A., Brusten, L. C. M., Draper, N. J., Elsmore, B. M., Freeman, C. D., Giles, D. M., Gong, H., Gormley, D., Griffiths, M. R., Hawkes, T. D. R., Haynes, P. S., Heesom, K. J., Herath, A., Hollis, K., Hudsen, L. J., Inman, J., Jacobs, M., Jarman, D., Kibria, I., Kilgour, J. J., Kinuthia, S. K., Lane, K. E., Lees, M. L., Loader, J., Longmore, A., McEwan, M., Middleton, A., Moore, S., Murray, C., Murray, H. M., Myatt, C. P., Ng, S. S., O'Neil, A., Parekh, R. B., Patel, A., Patel, K. B., Patel, S., Patel, T. P., Philp, R. J., Platt, A. E., Poyser, H., Prendergast, C., Prime, S., Redpath, N., Reeves, M., Robinson, A. W., Rohlff, C., Rosenbaum, J. M., Schenker, M., Scrivener, E., Shipston, N., Siddiq, S., Southan, C., Spencer, D. I. R., Stamps, A., Steffens, M. A., Stevenson, D., Sweetman, G. M. A., Taylor, S., Townsend, R., Ventom, A. M., Waller, M. N. H., Weresch, C., Williams, A. M., Woolliscroft, R. J., Yu, X., and Lyall, A. (2004) Annotation of the human genome by high-throughput sequence analysis of naturally occurring proteins. *Curr. Proteomics* **1,** 41–48
89. Rohlff, C. (2004) New approaches towards integrated proteomic databases and depositories. *Expert Rev. Proteomics* **1,** 267–274
90. Dowel, R. D., Jokerst, R. M., Day, A., Eddy, S. R., and Stein, L. (2001) The distributed annotation system. *BMC Bioinformatics* **2:**7
91. Craig, R., Cortens, J. P., and Beavis, R. C. (2004) Open source system for analyzing, validating, and storing protein identification data. *J. Proteome Res.* **3,** 1234–1242
92. Skrabanek, L., and Campagne, F. (2001) TissueInfo: high-throughput identification of tissue expression profiles and specificity. *Nucleic Acids Res.* **29,** e102
93. Mu, X., Zhao, S., Pershad, R., Hsieh, T. F., Scarpa, A., Wang, S. W., White, R. A., Beremand, P. D., Thomas, T. L., Gan, L., and Klein, W. H. (2001) Gene expression in the developing mouse retina by EST sequencing and microarray analysis. *Nucleic Acids Res.* **29,** 4983–4993
94. Yeo, G., Holste, D., Kreiman, G., and Burge, C. B. (2004) Variation in alternative splicing across human tissues. *Genome Biol.* **5,** R74
95. Taylor, C. F., Paton, N. W., Garwood, K. L., Kirby, P. D., Stead, D. A., Yin, Z., Deutsch, E. W., Selway, L., Walker, J., Riba-Garcia, I., Mohammed, S., Deery, M. J., Howard, J. A., Dunkley, T., Aebersold, R., Kell, D. B., Lilley, K. S., Roepstorff, P., Yates, J. R., III, Brass, A., Brown, A. J., Cash, P., Gaskell, S. J., Hubbard, S. J., and Oliver, S. G. (2003) A systematic approach to modeling capturing and disseminating proteomics experimental data. *Nat. Biotechnol.* **21,** 247–254
96. Pedrioli, P. G., Eng, J. K., Hubley, R., Vogelzang, M., Deutsch, E. W., Raught, B., Pratt, B., Nilsson, E., Angeletti, R. H., Apweiler, R., Cheung, K., Costello, C. E., Hermjakob, H., Huang, S., Julian, R. K., Kapp, E., McComb, M. E., Oliver, S. G., Omenn, G., Paton, N. W., Simpson, R., Smith, R., Taylor, C. F., Zhu, W., and Aebersold, R. (2004) A common open representation of mass spectrometry data and its application to proteomics research. *Nat. Biotechnol.* **22,** 1459–1466
97. Orchard, S., Zhu, W., Julian, R. K., Hermjakob, H., and Apweiler, R. (2003) Further advances in the development of a data interchange standard for proteomics data. *Proteomics* **3,** 2065–2066
98. MacCoss, M. J., McDonald, W. H., Saraf, A., Sadygov, R., Clark, J. M., Tasto, J. J., Gould, K. L., Wolters, D., Washburn, M., Weiss, A., Clark, J. I., and Yates, J. R. (2002) Shotgun identification of protein modifications from protein complexes and lens tissue. *Proc. Natl. Acad. Sci. U. S. A.* **99,** 7900–7905
99. Choudhary, G., Wu, S. L., Shieh, P., and Hancock, W. S. (2003) Multiple enzymatic digestion for enhanced sequence coverage of proteins in complex proteomic mixtures using capillary LC with ion trap MS/MS. *J. Proteome Res.* **2,** 59–67
100. Pan, S., Zhang, H. Rush, J., Eng, J., Zhang, N., Patterson, D., Comb, M. J., and Aebersold, R. (2005) High throughput proteome-screening for biomarker detection. (2005) *Mol. Cell. Proteomics* **4,** 182–190
101. Wall, D. B., Kachman, M. T., Gong, S. S., Parus, S. J., Long, M. W., and Lubman, D. M. (2001) Isoelectric focusing nonporous silica reversed-phase high-performance liquid chromatography/electrospray ionization time-of-flight mass spectrometry: a three-dimensional liquid-phase protein separation method as applied to the human erythroleukemia cell-line. *Rapid Commun. Mass Spectrom.* **15,** 1649–1661
102. Liu, H., Berger, S. J., Chakraborty, A. B., Plumb, R. S., and Cohen, S. A. (2002) Multidimensional chromatography coupled to electrospray ionization time-of-flight mass spectrometry as an alternative to two-dimensional gels for the identification and analysis of complex mixtures of intact proteins. *J. Chromatogr. B Anal. Technol. Biomed. Life Sci.* **782,** 267–289
103. Wienkoop, S., Glinski, M., Tanaka, N., Tolstikov, V., Fiehn, O., and Weckwerth, W. (2004) Linking protein fractionation with multidimensional monolithic reversed-phase peptide chromatography/ mass spectrometry enhances protein identification from complex mixtures even in the presence of abundant proteins. *Rapid Commun. Mass Spectrom.* **18,** 643–650
104. Moritz, R. L., Ji, H., Schutz, F., Connolly, L. M., Kapp, E. A., Speed, T. P., and Simpson, R. J. (2004) A proteome strategy for fractionating proteins and peptides using continuous free-flow electrophoresis coupled off-line to reversed-phase high-performance liquid chromatography. *Anal. Chem.* **76,** 4811–4824
105. Lee, S. W., Berger, S. J., Martinovic, S., Pasa-Tolic, L., Anderson, G. A., Shen, Y., Zhao, R., and Smith, R. D. (2002) Direct mass spectrometric analysis of intact proteins of the yeast large ribosomal subunit using capillary LC/FTICR. *Proc. Natl. Acad. Sci. U. S. A.* **99,** 5942–5947
106. VerBerkmoes, N. C., Bundy, J. L., Hauser, L., Asano, K. G., Razumovskaya, J., Larimer, F., Hettich, R. L., and Stephenson, J. L., Jr. (2002) Integrating "top-down" and "bottom-up" mass spectrometric approaches for proteomic analysis of *Shewanella oneidensis*. *J. Proteome Res.* **1,** 239–252
107. Strader, M. B., VerBerkmoes, N. C., Tabb, D. L., Connelly, H. M., Barton, J. W., Bruce, B. D., Pelletier, D. A., Davison, B. H., Hettich, R. L., Larimer, F. W., and Hurst, G. B. (2004) Characterization of the 70S ribosome from *Rhodopseudomonas palustris* using an integrated "top-down" and "bottom-up" mass spectrometric approach. *J. Proteome Res.* **3,** 965–978
108. Nemeth-Cawley, J. F., Tangarone, B. S., and Rouse, J. C. (2003) "Top down" characterization is a complementary technique to peptide se-

quencing for identifying protein species in complex mixtures. *J. Proteome Res.* **2,** 495–505
109. Wang, H., Kachman, M. T., Schwartz, D. R., Cho, K. R., and Lubman, D. M. (2004) Comprehensive proteome analysis of ovarian cancers using liquid phase separation, mass mapping and tandem mass spectrometry: a strategy for identification of candidate cancer biomarkers. *Proteomics* **4,** 2476–2495
110. Humphery-Smith, I. (2004) A human proteome project with a beginning and an end. *Proteomics* **4,** 2519–2521
111. Uhlen, M., and Ponten, F. (2005) Antibody-based proteomics for human tissue profiling. *Mol. Cell. Proteomics* **4,** 384–393
112. Prince, J. T., Carlson, M. W., Wang, R., Lu, P., and Marcotte, E. M. (2004) The need for a public proteomics repository. *Nat. Biotechnol.* **22,** 471–472