Mini Review

Crowdsourcing in proteomics: public resources lead to better experiments

Harald Barsnes[1] and Lennart Martens[2,3,*]

[1] Proteomics Unit, Department of Biomedicine, University of Bergen, Norway.

[2] Department of Medical Protein Research, VIB, Ghent, Belgium.

[3] Department of Biochemistry, Ghent University, Ghent, Belgium.

* Corresponding author:

Professor Lennart Martens, Department of Medical Protein Research and Biochemistry, VIB and Department of Biochemistry, Faculty of Medicine and Health Sciences, Ghent University, A. Baertsoenkaai 3, B-9000 Ghent, Belgium.

Email: lennart.martens@ugent.be

Tel: 32-92649458

Fax: 32-92649484

Running title: Public resources lead to better experiments

Abstract

With the growing interest in the field of proteomics, the amount of publicly available proteome resources has also increased dramatically. This means that there are many useful resources available for almost all aspects of a proteomics experiment. However, it remains vital to use the right resource, for the right purpose, at the right time. This review is therefore meant to aid the reader in obtaining an overview of the available resources and their application, thus providing the necessary background to choose the appropriate resources for the experiment at hand. Many of the resources are also taking advantage of so-called crowdsourcing to maximize the potential of the resource. What this means and how this can improve future experiments will also be discussed. The text roughly follows the steps involved in a proteomics experiment, starting with the planning of the experiment, *via* the processing of the data and the analysis of the results, to the community-wide sharing of the produced data.

Main Text

Background

The process of crowdsourcing refers to the outsourcing of tasks to a distributed group of people or communities, but unlike ordinary outsourcing the task or problem is here outsourced to an undefined public rather than to a specific group of people. Crowdsourcing is not a novel concept and has been employed in numerous types of situations, with perhaps the best known example being Wikipedia. The approach has become increasingly common with the global spread of the internet. While crowdsourcing may not be a commonly used term in the proteomics community, significant parts of key proteomics resources have already been crowdsourced from the scientific community for a long time. This review will highlight the best examples and show how the use of public resources can lead to better experiments.

The most commonly employed technique in proteomics today is mass spectrometry (MS). Usually, MS based proteomics is carried out according the bottom-up strategy, where the proteins to be analyzed are first proteolytically cleaved *in vitro* into smaller pieces called peptides, that are further fragmented by the MS instrument into fragment ions, resulting in MS/MS spectra (Aebersold and Mann 2003). There are two general flavors of this type of MS based proteomics experiment: i) discovery-oriented experiments that try to identify (and increasingly, try to quantify) all the proteins in a sample, and ii) targeted experiments that focus on the analysis of specific signature peptides for a limited number of proteins of interest in a sample.

Despite the overall similarity in methods employed between these two flavors of bottom-up proteomics, it is important to keep in mind that their properties are very different, thus requiring unique experimental workflows and bioinformatics pipelines for each. Fortunately, a vast amount of publicly available resources exist that can guide the user on his/her way. In this review we will cover the most important relevant resources and show when and how they ought to be applied in order to maximize the potential of the analysis, see Figure 1.

Note that we have decided to focus on resources that help the user with specific tasks required during an experiment. Software packages aiming at providing a complete pipeline are outside the scope of this review and will thus not be covered. The same is true for the long list of possible proteomics search engines. Finally, where possible, we will focus on the resources employing crowdsourcing, highlighting how these resources are especially important to get the most out of the large amount of proteomics data being produced every year.

From Ideas to Experiments

Perhaps the most important part of a proteomics experiment (or indeed any experiment) is the planning of the experiment. This step, which can also be referred to as experimental design, is paramount to obtaining a useful outcome. Shortcomings at this stage are very difficult, if not impossible, to correct downstream in the workflow, further emphasizing the critical relevance of this stage. Even though this simple truth ought to be obvious, it is not uncommon to encounter experiments where small changes in the design could have significantly increased the value of the experiment. In some cases a planning-phase analysis may even save one the trouble of doing

experiments, if it becomes clear *a priori* that they will have a very low probability of success, for instance if too few replicates are available to obtain a reasonable likelihood of achieving significant results. It is therefore essential to plan an experiment in advance, and it can be very helpful to consult relevant resources in this initial phase.

Besides the obvious caveats connected to statistical considerations, such as power analysis (Levin 2011), replicate types and numbers (Oberg and Vitek 2009), and sample pooling (Karp and Lilley 2009) that affect discovery-oriented and targeted approaches alike, there are also more specific issues that relate predominantly to the planning of targeted proteomics experiments. These latter issues are the main focus of this section, since they can most readily be addressed through the judicious use of public resources. Examples include the determination of suitable peptides and transitions for a selected reaction monitoring (SRM) experiment, or the focused analysis of membrane proteins.

In targeted proteomics, as the name implies, a set of proteins of interest is typically known *a priori*, often as a result of a preceding discovery-oriented analysis (Gallien et al. 2011). But even if the protein set is known in advance it can still be very useful to perform a thorough pre-analysis. The simplest form of investigation in this case would be to perform an *in silic*o digest of the given protein, i.e., to theoretically cleave the protein into peptides using the cleavage pattern and specifics of the desired proteolytic enzyme, and list the detectable peptides as limited by one or more appropriate parameters such as sequence length or mass, sequence composition, number of missed cleavages, or theoretical iso-electric point (pI). Such *in silico*

digestion and in-line peptide filtering is readily offered by several freely available database processing tools and proteomics software libraries (Martens et al. 2005; Reisinger and Martens 2009; Colinge et al. 2006; Barsnes et al. 2011). This straightforward analysis yields indications of the detectable areas of the protein sequence, and the expected number of peptides that can be successfully targeted. Of course, if the objective of the study is the detection of a specific post-translational modification or *in vivo* cleavage, it can be immediately verified whether the sequence region of interest yields any detectable peptides, see Figure 2. If the chosen protease should not yield the desired peptide(s), it is then an exceedingly simple exercise to try another protease and evaluate the outcome. If multiple targets are considered, perhaps a combination of proteases (albeit employed in parallel) could even be considered (Swaney et al. 2010). Performing *in silico* digests is however not an exact science. First of all, cleavage may not always occur where expected (Hamady et al. 2005; Thiede et al. 2000; Yen et al. 2006), and may even occur where it is not expected (Rodriguez et al. 2008). Improved prediction of proteolytic behavior could therefore still benefit bottom-up proteomics approaches (Siepen et al. 2007). It should also be underlined that the detectability of a peptide is specific to the mass spectrometry protocol, and that the results are therefore not directly transferable between protocols.

Such an *in silico* digest can then be complemented by resources that provide so-called proteotypic peptides (Craig et al. 2005), such as PeptideAtlas (Desiere et al. 2006) or GPMDB (Craig et al. 2004), although the actual overlap in proteotypic peptides between these resources is incomplete. Computational methods have been proposed to predict prototypic peptides, e.g., (Mallick et al. 2007), but they have been shown to function quite poorly outside of their (typically limited) initial training conditions

(Mueller et al. 2008), indicating that caution should be used in relying only on purely predicted entries.

It is not just a predefined set of peptides that is selected for in targeted proteomics however, but rather a combination of fragment ion masses and peptide precursor mass for each targeted peptide. Such a precursor and fragment ion pair is commonly referred to as a transition, and such transitions form the core of the increasingly popular selected reaction monitoring (SRM) approach in targeted proteomics (Lange et al. 2008b). In this technique, targeting the correct transitions is of equal importance as targeting the right peptides. Various resources for predicting good transitions have been created, and these represent either predicted optimally detectable transitions (MacLean et al. 2010; Abbatiello et al. 2010; Brusniak et al. 2011), or unique transitions (Helsens et al. 2012), i.e., transitions that are unlikely to be contaminated by overlapping signals from other, non-targeted peptides (Sherman et al. 2009). Databases of known, experimentally observed transitions are also available (Domon and Aebersold 2006; Lange et al. 2008a; Deutsch et al. 2008; Craig et al. 2004), and are likely to become increasingly important over time as more of these analyses are ran. Importantly, work on standardizing an exchange format for transitions has also progressed(Deutsch et al. 2011), and an international initiative to find transitions for all human proteins has been initiated as a part of the Human Proteome Organization - Human Proteome Project (HUPO-HPP) (The call of the human proteome 2010; Nilsson et al. 2010). Interestingly, a rather complete library of synthetic peptides has already been completed for yeast (Picotti et al. 2010). It is important to consider that good transitions are often specific to an instrument and the experimental settings used, i.e., a good transition for one experimental setup might not be a good transition when

using a different experimental setup. To some extent this can be related to the observed instrument dependent peptide fragmentation (Barsnes et al. 2010; Sherwood et al. 2009), although it has been reported that top-ranking peaks are more readily transferable (Sherwood et al. 2009).

In order to efficiently reuse experimental data it is important to filter the data according to requirements (Foster et al. 2011), and to understand the experimental setup in sufficient detail. This is not always straightforward however, since even the protocol descriptions in the metadata-rich PRIDE database (Vizcaíno et al. 2010) are often lacking in detail. Furthermore, data quality is not necessarily correlated with annotation quality. This situation highlights the importance of adhering to minimum information guidelines such as MIAPE (Taylor 2006), one of the main goals of both the ProDaC project (Eisenacher et al. 2009) and the HUPO-PSI (Martens et al. 2007; Vizcaino et al. 2007).

It is of course always possible that a particular protein of interest turns out to be very hard to target, despite the best efforts at finding suitable candidate peptides and transitions. If the study objective is not the specific characterization of a cleavage or post-translational modification on that particular protein, a possible alternative may be to find related proteins and target these instead. This can be achieved by searching for interaction partners in IntAct (Kerrien et al. 2012), STRING (Szklarczyk et al. 2011) or DAVID (Huang da et al. 2009), or by looking for proteins in the same pathway in Reactome (Matthews et al. 2009) or KEGG (Kanehisa et al. 2012).

From Acquired Data to Processed Results

Once the experiment has been performed, the data needs to be processed into peptide identifications (and quantifications). Since the actual output of an MS experiment is a set of MS/MS spectra, the first step typically consists of identifying the origin of these spectra, i.e., match spectra to peptides. While various options exist for this goal (including *de novo* sequencing (Frank et al. 2007; Ma and Johnson 2012) and tag-based approaches (Tabb et al. 2008; Tabb et al. 2003)), database searching is by far the most common spectrum identification method. Here, acquired spectra are matched to a set of known spectra to find the best match and thereby transitively assign identification. There are two types of databases that can be used for this purpose: protein sequence databases and spectral databases. The former are used to derive relatively unsophisticated *in silico* theoretical fragmentation spectra (Cottrell 1994; Eng et al. 1994; Yates et al. 1995; Geer et al. 2004; Fenyo and Beavis 2003), while the latter contain previously identified experimental MS/MS spectra (Lam 2011)(NIST: http://peptide.nist.gov) and have been shown to outperform sequence database searching under optimal conditions (Zhang et al. 2011). One can also generate a realistic spectral library from a sequence database using detailed MS/MS spectrum prediction (Zhang 2004, 2005; Yen et al. 2011; Yen et al. 2009), but this is a more complicated process.

When searching a database, it is of course exceedingly important to use the right database. Selecting a database that does not contain the protein in question will result in the protein going undetected, while choosing a database that contains too much data can result in a large amount of false positives (Colaert et al. 2011; Everett et al. 2010; Na et al. 2012; Bern and Kil 2011), similar to the effect encountered when using error-tolerant searches (Creasy and Cottrell 2002). Numerous sequence

databases are available, ranging from small home-grown databases to the centralized and well-annotated species-wide databases like UniProt (UniProt Consortium 2010), Ensembl (Flicek et al. 2011) and NCBI nr (Pruitt et al. 2007).

In choosing a database, a choice has therefore to be made between completeness, reliability and redundancy of the information contained in these systems. It is however possible to compare search databases on this basis using a very simple set of metrics: define the number of identifiable tryptic peptides with one allowed missed cleavage in each database as the database size, define the number of unique such peptides as the database information, and then calculate the information ratio as the ratio of database information over database size. The results are given for several popular human sequence databases in Figure 3. Note that bigger databases typically have larger information content, but that the information ratio does not scale very well. This indicates that smaller databases tend to be more compact and efficient, and that larger databases are more redundant. This is expected, since the larger databases (for human sequences at least) achieve their size growth by including splice isoforms that share a substantial part of their sequence, and therefore many of their tryptic peptides.

It is of note that the International Protein Index (IPI) (Kersey et al. 2004), one of the most popular proteomics sequence databases, has been discontinued in September 2011. The effect that this has to the proteomics community has been examined in some detail in two recent papers (Griss et al. 2011a; Griss et al. 2011b) where it was also shown that UniProt now provides directly comparable alternatives, an

observation that is also evident for human sequences from Figure 3, where both UniProt and Ensembl are seen to closely resemble IPI.

The information in a sequence database need not necessarily be taken as is, however. Indeed, further processing can dramatically extend the reach of a sequence database and correspondingly increase the number of relevant identifications, by performing serial differential enzymatic digestion (Van Damme et al. 2005), amino- or carboxy-terminal ragging (Gevaert et al. 2003), sequence-based subset selection (Gevaert et al. 2004; Gevaert et al. 2002), or to create and add decoy sequences (Vaudel et al. 2011). Various tools exist that greatly facilitate such operations (Martens et al. 2005; Reisinger and Martens 2009; Reidegeld et al. 2008).

As mentioned above however, the use of arbitrarily large databases (or simulating them, through second-pass, error-tolerant, or multi-stage searches (Tharakan et al. 2010)) is wrought with peril. It becomes relatively easy to severely underestimate false positive rates (Creasy and Cottrell 2002; Everett et al. 2010; Colaert et al. 2011) and it can strongly complicate the assignment of peptides to proteins (Martens and Hermjakob 2007; Nesvizhskii and Aebersold 2005). The best possible database should therefore be both comprehensive as well as compact, and should be *in silico* adapted to the specifics of the protocol.

From Proteins to Knowledge

Once a set of peptides is identified, and a set of proteins is successfully inferred from these, it is often useful to obtain as much information as possible about these proteins. This transition from results to knowledge is of course extensively reliant on the large

amount of available annotation resources. These include, but are not limited to, gene ontology annotations (Binns et al. 2009), pathways (Matthews et al. 2009), interactions (Szklarczyk et al. 2011), functional annotations (Huang da et al. 2009), 3D structures (Villaveces et al. 2011; Prlic et al. 2005; Berman et al. 2000) and protein domain signatures (Hunter et al. 2012). A comprehensive overview of available resources for the annotation of protein sets is provided in (Vizcaino et al. 2009).

In order to use most of these resources however, one first has to be able to provide a protein identifier or accession number that the resource in question can understand. Note that this means that using home-made databases for peptide identification of protein inference may severely limit the downstream annotation options. Such problems are largely circumvented when a well-annotated database like UniProt is used. Incidentally, UniProt itself serves as a powerful information hub that links each protein to a plethora of external databases that span the whole spectrum of the life sciences. In those cases where the available protein identifiers are different from the ones requested by a given resource, the Protein Identifier Cross-Reference Service (PICR) can be used to perform a seamless accession number conversion (Côté et al. 2007). PICR provides protein identifier mappings based on 100% sequence identity to proteins from a large number of popular databases. The mappings can be limited by database, taxonomy and activity status in the databases.

Using a single resource to annotate your data set can be very valuable, but the real value comes from combining information from different data resources. In the past, this could often be a time consuming and laborious task, but recent developments

have made this task much easier. Perhaps the most powerful tool in this context is provided by BioMart (Guberman et al. 2011; Haider et al. 2009; Kasprzyk 2011; Smedley et al. 2009; Woollard 2010), a distributed system that includes a large amount of databases and resources in the same, easily accessible framework. With the ability to perform queries spanning two repositories, it becomes straightforward to combine data from multiple, unrelated databases. Thanks to the inclusion in BioMart of hub databases such as UniProt and Ensembl that provide extensive cross-references, and in the case of Ensembl even span multiple levels of biological information from genome over transcriptome to proteome, BioMart queries can be extremely powerful ways to collate detailed knowledge about thousands of proteins in a few mouse clicks, see Figure 4.

From Private Local Data to Public Online Data

At this stage, the workflow ought to have resulted in well-annotated protein identifications and new knowledge. The next and final step is then to make this information available to the proteomics community at large. With an increased focus on the benefits of data sharing (Editors 2007, 2008), the development of standardized data formats such as mzML (Martens et al. 2011) and mzIdentML (Eisenacher 2011), and the development of easy to use conversion tools like PRIDE Converter (Barsnes et al. 2009), the sharing of proteomics data has become common practice. Meanwhile, the field continues to develop towards even more efficient data sharing. Notably, the ProteomeXchange consortium (Martens 2011; Orchard et al. 2011) was founded with the specific aim of providing a single point of submission for MS-based proteomics data, and the objective to distribute these data across all current repositories for optimal re-use and dissemination. Indeed, depositing data into a public repository is

not enough; the data has to be easy to retrieve, to search, to evaluate, and to view. The PRIDE team for instance, developed the PRIDE Inspector application (Wang et al. 2012) specifically for this purpose. A desktop application for visualizing and performing quality assessment on mass spectrometry data contained in the PRIDE database, PRIDE Inspector greatly improves on the old, web-based interaction with the PRIDE database, making it much more straightforward to work with the data in the repository.

The remaining challenge for the field is to provide high-quality, well-annotated data to the repositories, allowing today's proteomics experiments to become part of a priceless resource for the planning of future experiments. Interestingly, there is thus a strong link back to the start of this review article, where the design and planning of experiments was discussed; the public availability of quality controlled and well-annotated proteomics data will result in ever better planned experiments, in turn leading to improved insight and knowledge from these experiments. This self-enhancing feedback loop will allow data itself to become proteomics' biggest resource, and will allow the field to become a center-stage participant in the broader life sciences.

Acknowledgments

H.B. is supported by the Research Council of Norway, and L.M. acknowledges the support of Ghent University (Multidisciplinary Research Partnership “Bioinformatics: from nucleotides to networks”), and the PRIME-XS and ProteomeXchange projects, grant agreement numbers 262067 and 260558, both funded by the European Union 7th Framework Program.

References

Abbatiello SE, Mani DR, Keshishian H, Carr SA (2010) Automated detection of inaccurate and imprecise transitions in peptide quantification by multiple reaction monitoring mass spectrometry. Clin Chem 56 (2):291-305. doi:clinchem.2009.138420 [pii] 10.1373/clinchem.2009.138420 [doi]

Aebersold R, Mann M (2003) Mass spectrometry-based proteomics. Nature 422:198-207

Barsnes H, Eidhammer I, Martens L (2010) FragmentationAnalyzer: An open-source tool to analyze MS/MS fragmentation data. Proteomics 10 (5):1087-1090

Barsnes H, Vaudel M, Colaert N, Helsens K, Sickmann A, Berven FS, Martens L (2011) compomics-utilities: an open-source Java library for computational proteomics. BMC Bioinformatics 12:70. doi:1471-2105-12-70 [pii] 10.1186/1471-2105-12-70 [doi]

Barsnes H, Vizcaíno JA, Eidhammer I, Martens L (2009) PRIDE Converter: making proteomics data-sharing easy. Nat Biotechnol 27 (7):598-599

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The Protein Data Bank. Nucleic Acids Res 28 (1):235-242. doi:gkd090 [pii]

Bern M, Kil YJ (2011) Comment on "Unbiased statistical analysis for multi-stage proteomic search strategies". J Proteome Res 10 (4):2123-2127. doi:10.1021/pr101143m [doi]

Binns D, Dimmer E, Huntley R, Barrell D, O'Donovan C, Apweiler R (2009) QuickGO: a web-based tool for Gene Ontology searching. Bioinformatics 25 (22):3045-3046. doi:btp536 [pii] 10.1093/bioinformatics/btp536 [doi]

Brusniak MY, Kwok ST, Christiansen M, Campbell D, Reiter L, Picotti P, Kusebauch U, Ramos H, Deutsch EW, Chen J, Moritz RL, Aebersold R (2011) ATAQS: A computational software tool for high throughput transition optimization and validation for selected reaction monitoring mass spectrometry. BMC Bioinformatics 12:78. doi:1471-2105-12-78 [pii] 10.1186/1471-2105-12-78 [doi]

The call of the human proteome (2010). Nat Methods 7 (9):661

Colaert N, Degroeve S, Helsens K, Martens L (2011) Analysis of the resolution limitations of peptide identification algorithms. J Proteome Res 10 (12):5555-5561. doi:10.1021/pr200913a [doi]

Colinge J, Masselot A, Carbonell P, Appel RD (2006) InSilicoSpectro: an open-source proteomics library. J Proteome Res 5 (3):619-624. doi:10.1021/pr0504236 [doi]

Côté RG, Jones P, Martens L, Kerrien S, Reisinger F, Lin Q, Leinonen R, Apweiler R, Hermjakob H (2007) The Protein Identifier Cross-Referencing (PICR) service: reconciling protein identifiers across multiple source databases. BMC Bioinformatics 8 (401)

Cottrell JS (1994) Protein identification by peptide mass fingerprinting. Pept Res 7 (3):115-124

Craig R, Cortens JP, Beavis RC (2004) Open source system for analyzing, validating, and storing protein identification data. J Proteome Res 3 (6):1234-1242. doi:10.1021/pr049882h [doi]

Craig R, Cortens JP, Beavis RC (2005) The use of proteotypic peptide libraries for protein identification. Rapid Commun Mass Spectrom 19 (13):1844-1850. doi:10.1002/rcm.1992 [doi]

Creasy DM, Cottrell JS (2002) Error tolerant searching of uninterpreted tandem mass spectrometry data. Proteomics 2 (10):1426-1434. doi:10.1002/1615-9861(200210)2:10<1426::AID-PROT1426>3.0.CO;2-5 [doi]

Desiere F, Deutsch EW, King NL, Nesvizhskii AI, Mallick P, Eng J, Chen S, Eddes J, Loevenich SN, Aebersold R (2006) The PeptideAtlas project. Nucleic Acids Res 34 (Database issue):D655-658. doi:34/suppl_1/D655 [pii] 10.1093/nar/gkj040 [doi]

Deutsch EW, Chambers M, Neumann S, Levander F, Binz PA, Shofstahl J, Campbell DS, Mendoza L, Ovelleiro D, Helsens K, Martens L, Aebersold R, Moritz RL, Brusniak MY (2011) TraML: a standard format for exchange of selected reaction monitoring transition lists. Mol Cell Proteomics. doi:R111.015040 [pii] 10.1074/mcp.R111.015040 [doi]

Deutsch EW, Lam H, Aebersold R (2008) PeptideAtlas: a resource for target selection for emerging targeted proteomics workflows. EMBO Rep 9 (5):429-434. doi:embor200856 [pii] 10.1038/embor.2008.56 [doi]

Domon B, Aebersold R (2006) Mass spectrometry and protein analysis. Science 312 (5771):212-217. doi:312/5771/212 [pii] 10.1126/science.1124619 [doi]

Editors (2007) Democratizing proteomics data. Nat Biotechnol 25 (3):262

Editors (2008) Thou shalt share your data. Nat Methods 5 (3):209-209

Eisenacher M (2011) mzIdentML: an open community-built standard format for the results of proteomics spectrum identification algorithms. Methods Mol Biol 696:161-177. doi:10.1007/978-1-60761-987-1_10 [doi]

Eisenacher M, Martens L, Hardt T, Kohl M, Barsnes H, Helsens K, Häkkinen J, Levander F, Aebersold R, Vandekerckhove J, Dunn MJ, Lisacek F, Siepen JA, Hubbard SJ, Binz PA, Blüggel M, Thiele H, Cottrell J, Meyer HE, Apweiler R, Stephan C (2009) Getting a grip on proteomics data - Proteomics Data Collection (ProDaC). Proteomics 9 (15):3928-3933.

Eng J, McCormack AL, Yates JR, III (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 (11):976–989

Everett LJ, Bierl C, Master SR (2010) Unbiased statistical analysis for multi-stage proteomic search strategies. J Proteome Res 9 (2):700-707. doi:10.1021/pr900256v [doi]

Fenyo D, Beavis RC (2003) A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes. Anal Chem 75 (4):768-774

Flicek P, Amode MR, Barrell D, Beal K, Brent S, Chen Y, Clapham P, Coates G, Fairley S, Fitzgerald S, Gordon L, Hendrix M, Hourlier T, Johnson N, Kahari A, Keefe D, Keenan S, Kinsella R, Kokocinski F, Kulesha E, Larsson P, Longden I, McLaren W, Overduin B, Pritchard B, Riat HS, Rios D, Ritchie GR, Ruffier M, Schuster M, Sobral D, Spudich G, Tang YA, Trevanion S, Vandrovcova J, Vilella AJ, White S, Wilder SP, Zadissa A, Zamora J, Aken BL, Birney E, Cunningham F, Dunham I, Durbin R, Fernandez-Suarez XM, Herrero J, Hubbard TJ, Parker A, Proctor G, Vogel J, Searle SM (2011)

Ensembl 2011. Nucleic Acids Res 39 (Database issue):D800-806. doi:gkq1064 [pii]
10.1093/nar/gkq1064 [doi]

Foster JM, Degroeve S, Gatto L, Visser M, Wang R, Griss J, Apweiler R, Martens L (2011) A posteriori quality control for the curation and reuse of public proteomics data. Proteomics 11 (11):2182-2194. doi:10.1002/pmic.201000602 [doi]

Frank AM, Savitski MM, Nielsen ML, Zubarev RA, Pevzner PA (2007) De novo peptide sequencing and identification with precision mass spectrometry. J Proteome Res 6 (1):114-123

Gallien S, Duriez E, Domon B (2011) Selected reaction monitoring applied to proteomics. J Mass Spectrom 46 (3):298-312. doi:10.1002/jms.1895 [doi]

Geer LY, Markey SP, Kowalak JA, Wagner L, Xu M, Maynard DM, Yang X, Shi W, Bryant SH (2004) Open mass spectrometry search algorithm. J Proteome Res 3 (5):958-964

Gevaert K, Ghesquiere B, Staes A, Martens L, Van Damme J, Thomas GR, Vandekerckhove J (2004) Reversible labeling of cysteine-containing peptides allows their specific chromatographic isolation for non-gel proteome studies. Proteomics 4 (4):897-908. doi:10.1002/pmic.200300641 [doi]

Gevaert K, Goethals M, Martens L, Van Damme J, Staes A, Thomas GR, Vandekerckhove J (2003) Exploring proteomes and analyzing protein processing by mass spectrometric identification of sorted N-terminal peptides. Nat Biotechnol 21 (5):566-569. doi:10.1038/nbt810 [doi]
nbt810 [pii]

Gevaert K, Van Damme J, Goethals M, Thomas GR, Hoorelbeke B, Demol H, Martens L, Puype M, Staes A, Vandekerckhove J (2002) Chromatographic isolation of methionine-containing peptides for gel-free proteome analysis: identification of more than 800 Escherichia coli proteins. Mol Cell Proteomics 1 (11):896-903

Griss J, Cote RG, Gerner C, Hermjakob H, Vizcaino JA (2011a) Published and perished? The influence of the searched protein database on the long-term storage of proteomics data. Mol Cell Proteomics 10 (9):M111 008490. doi:M111.008490 [pii]
10.1074/mcp.M111.008490 [doi]

Griss J, Martin M, O'Donovan C, Apweiler R, Hermjakob H, Vizcaino JA (2011b) Consequences of the discontinuation of the International Protein Index (IPI) database and its substitution by the UniProtKB "complete proteome" sets. Proteomics 11 (22):4434-4438. doi:10.1002/pmic.201100363 [doi]

Guberman JM, Ai J, Arnaiz O, Baran J, Blake A, Baldock R, Chelala C, Croft D, Cros A, Cutts RJ, Di Genova A, Forbes S, Fujisawa T, Gadaleta E, Goodstein DM, Gundem G, Haggarty B, Haider S, Hall M, Harris T, Haw R, Hu S, Hubbard S, Hsu J, Iyer V, Jones P, Katayama T, Kinsella R, Kong L, Lawson D, Liang Y, Lopez-Bigas N, Luo J, Lush M, Mason J, Moreews F, Ndegwa N, Oakley D, Perez-Llamas C, Primig M, Rivkin E, Rosanoff S, Shepherd R, Simon R, Skarnes B, Smedley D, Sperling L, Spooner W, Stevenson P, Stone K, Teague J, Wang J, Whitty B, Wong DT, Wong-Erasmus M, Yao L, Youens-Clark K, Yung C, Zhang J, Kasprzyk A (2011) BioMart Central Portal: an open database network for the biological community. Database (Oxford) 2011:bar041. doi:bar041 [pii]
10.1093/database/bar041 [doi]

Haider S, Ballester B, Smedley D, Zhang J, Rice P, Kasprzyk A (2009) BioMart Central Portal--unified access to biological data. Nucleic Acids Res 37 (Web Server issue):W23-27. doi:gkp265 [pii]
10.1093/nar/gkp265 [doi]

Hamady M, Cheung TH, Tufo H, Knight R (2005) Does protein structure influence trypsin miscleavage? Using structural properties to predict the behavior of related proteins. IEEE Eng Med Biol Mag 24 (3):58-66

Helsens K, Mueller M, Hulstaert N, Martens L (2012) Sigpep: Calculating unique peptide signature transition sets in a complete proteome background. Proteomics (in press)

Huang da W, Sherman BT, Lempicki RA (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 4 (1):44-57. doi:nprot.2008.211 [pii]
10.1038/nprot.2008.211 [doi]

Hunter S, Jones P, Mitchell A, Apweiler R, Attwood TK, Bateman A, Bernard T, Binns D, Bork P, Burge S, de Castro E, Coggill P, Corbett M, Das U, Daugherty L, Duquenne L, Finn RD, Fraser M, Gough J, Haft D, Hulo N, Kahn D, Kelly E, Letunic I, Lonsdale D, Lopez R, Madera M, Maslen J, McAnulla C, McDowall J, McMenamin C, Mi H, Mutowo-Muellenet P, Mulder N, Natale D, Orengo C, Pesseat S, Punta M, Quinn AF, Rivoire C, Sangrador-Vegas A, Selengut JD, Sigrist CJ, Scheremetjew M, Tate J, Thimmajanarthanan M, Thomas PD, Wu CH, Yeats C, Yong SY (2012) InterPro in 2011: new developments in the family and domain prediction database. Nucleic Acids Res 40 (Database issue):D306-312. doi:gkr948 [pii]
10.1093/nar/gkr948 [doi]

Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M (2012) KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Res 40 (Database issue):D109-114. doi:gkr988 [pii]
10.1093/nar/gkr988 [doi]

Karp NA, Lilley KS (2009) Investigating sample pooling strategies for DIGE experiments to address biological variability. Proteomics 9 (2):388-397. doi:10.1002/pmic.200800485 [doi]

Kasprzyk A (2011) BioMart: driving a paradigm change in biological data management. Database (Oxford) 2011:bar049. doi:bar049 [pii]
10.1093/database/bar049 [doi]

Kerrien S, Aranda B, Breuza L, Bridge A, Broackes-Carter F, Chen C, Duesbury M, Dumousseau M, Feuermann M, Hinz U, Jandrasits C, Jimenez RC, Khadake J, Mahadevan U, Masson P, Pedruzzi I, Pfeiffenberger E, Porras P, Raghunath A, Roechert B, Orchard S, Hermjakob H (2012) The IntAct molecular interaction database in 2012. Nucleic Acids Res 40 (Database issue):D841-846. doi:gkr1088 [pii]
10.1093/nar/gkr1088 [doi]

Kersey PJ, Duarte J, Williams A, Karavidopoulou Y, Birney E, Apweiler R (2004) The International Protein Index: an integrated database for proteomics experiments. Proteomics 4 (7):1985-1988. doi:10.1002/pmic.200300721 [doi]

Lam H (2011) Building and searching tandem mass spectral libraries for peptide identification. Mol Cell Proteomics 10 (12):R111 008565. doi:R111.008565 [pii]
10.1074/mcp.R111.008565 [doi]

Lange V, Malmstrom JA, Didion J, King NL, Johansson BP, Schafer J, Rameseder J, Wong CH, Deutsch EW, Brusniak MY, Buhlmann P, Bjorck L, Domon B, Aebersold R (2008a) Targeted quantitative analysis of Streptococcus pyogenes virulence factors by multiple reaction monitoring. Mol Cell Proteomics 7 (8):1489-1500. doi:M800032-MCP200 [pii] 10.1074/mcp.M800032-MCP200 [doi]

Lange V, Picotti P, Domon B, Aebersold R (2008b) Selected reaction monitoring for quantitative proteomics: a tutorial. Mol Syst Biol 4 (222):Epub

Levin Y (2011) The role of statistical power analysis in quantitative proteomics. Proteomics 11 (12):2565-2567. doi:10.1002/pmic.201100033 [doi]

Ma B, Johnson R (2012) De novo sequencing and homology searching. Mol Cell Proteomics 11 (2):O111 014902. doi:O111.014902 [pii] 10.1074/mcp.O111.014902 [doi]

MacLean B, Tomazela DM, Shulman N, Chambers M, Finney GL, Frewen B, Kern R, Tabb DL, Liebler DC, MacCoss MJ (2010) Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. Bioinformatics 26 (7):966-968. doi:btq054 [pii] 10.1093/bioinformatics/btq054 [doi]

Mallick P, Schirle M, Chen SS, Flory MR, Lee H, Martin D, Ranish J, Raught B, Schmitt R, Werner T, Kuster B, Aebersold R (2007) Computational prediction of proteotypic peptides for quantitative proteomics. Nat Biotechnol 25 (1):125-131. doi:nbt1275 [pii] 10.1038/nbt1275 [doi]

Martens L (2011) Proteomics databases and repositories. Methods Mol Biol 694:213-227. doi:10.1007/978-1-60761-977-2_14 [doi]

Martens L, Chambers M, Sturm M, Kessner D, Levander F, Shofstahl J, Tang WH, Rompp A, Neumann S, Pizarro AD, Montecchi-Palazzi L, Tasman N, Coleman M, Reisinger F, Souda P, Hermjakob H, Binz PA, Deutsch EW (2011) mzML--a community standard for mass spectrometry data. Mol Cell Proteomics 10 (1):R110 000133. doi:R110.000133 [pii] 10.1074/mcp.R110.000133 [doi]

Martens L, Hermjakob H (2007) Proteomics data validation: why all must provide data. Mol Biosyst 3 (8):518-522. doi:10.1039/b705178f [doi]

Martens L, Orchard S, Apweiler R, Hermjakob H (2007) Human Proteome Organization Proteomics Standards Initiative: data standardization, a view on developments and policy. Mol Cell Proteomics 6 (9):1666-1667. doi:6/9/1666 [pii]

Martens L, Vandekerckhove J, Gevaert K (2005) DBToolkit: processing protein databases for peptide-centric proteomics. Bioinformatics 21 (17):3584-3585

Matthews L, Gopinath G, Gillespie M, Caudy M, Croft D, de Bono B, Garapati P, Hemish J, Hermjakob H, Jassal B, Kanapin A, Lewis S, Mahajan S, May B, Schmidt E, Vastrik I, Wu G, Birney E, Stein L, D'Eustachio P (2009) Reactome knowledgebase of human biological pathways and processes. Nucleic Acids Res 37 (Database issue):D619-622. doi:gkn863 [pii] 10.1093/nar/gkn863 [doi]

Mueller M, Vizcaino JA, Jones P, Cote R, Thorneycroft D, Apweiler R, Hermjakob H, Martens L (2008) Analysis of the experimental detection of central nervous system-related genes in human brain and cerebrospinal fluid datasets. Proteomics 8 (6):1138-1148. doi:10.1002/pmic.200700761 [doi]

Na S, Bandeira N, Paek E (2012) Fast multi-blind modification search through tandem mass spectrometry. Mol Cell Proteomics 11 (4):M111 010199. doi:M111.010199 [pii]
10.1074/mcp.M111.010199 [doi]
Nesvizhskii AI, Aebersold R (2005) Interpretation of shotgun proteomic data: the protein inference problem. Mol Cell Proteomics 4 (10):1419-1440
Nilsson T, Mann M, Aebersold R, Yates JR, 3rd, Bairoch A, Bergeron JJ (2010) Mass spectrometry in high-throughput proteomics: ready for the big time. Nat Methods 7 (9):681-685. doi:nmeth0910-681 [pii]
10.1038/nmeth0910-681 [doi]
Oberg AL, Vitek O (2009) Statistical design of quantitative mass spectrometry-based proteomic experiments. J Proteome Res 8 (5):2144-2156. doi:10.1021/pr8010099 [doi]
10.1021/pr8010099 [pii]
Orchard S, Albar JP, Deutsch EW, Eisenacher M, Vizcaino JA, Hermjakob H (2011) Enabling BioSharing - a report on the Annual Spring Workshop of the HUPO-PSI April 11-13, 2011, EMBL-Heidelberg, Germany. Proteomics 11 (22):4284-4290. doi:10.1002/pmic.201190117 [doi]
Picotti P, Rinner O, Stallmach R, Dautel F, Farrah T, Domon B, Wenschuh H, Aebersold R (2010) High-throughput generation of selected reaction-monitoring assays for proteins and proteomes. Nat Methods 7 (1):43-46. doi:nmeth.1408 [pii]
10.1038/nmeth.1408 [doi]
Prlic A, Down TA, Hubbard TJ (2005) Adding some SPICE to DAS. Bioinformatics 21 Suppl 2:ii40-41. doi:21/suppl_2/ii40 [pii]
10.1093/bioinformatics/bti1106 [doi]
Pruitt KD, Tatusova T, Maglott DR (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res 35 (Database issue):D61-65. doi:gkl842 [pii]
10.1093/nar/gkl842 [doi]
Reidegeld KA, Eisenacher M, Kohl M, Chamrad D, Korting G, Bluggel M, Meyer HE, Stephan C (2008) An easy-to-use Decoy Database Builder software tool, implementing different decoy strategies for false discovery rate calculation in automated MS/MS protein identifications. Proteomics 8 (6):1129-1137. doi:10.1002/pmic.200701073 [doi]
Reisinger F, Martens L (2009) Database on Demand - an online tool for the custom generation of FASTA-formatted sequence databases. Proteomics 9 (18):4421-4424. doi:10.1002/pmic.200900254 [doi]
Rodriguez J, Gupta N, Smith RD, Pevzner PA (2008) Does trypsin cut before proline? J Proteome Res 7 (1):300-305. doi:10.1021/pr0705035 [doi]
Sherman J, McKay MJ, Ashman K, Molloy MP (2009) Unique ion signature mass spectrometry, a deterministic method to assign peptide identity. Mol Cell Proteomics 8 (9):2051-2062. doi:M800512-MCP200 [pii]
10.1074/mcp.M800512-MCP200 [doi]
Sherwood CA, Eastham A, Lee LW, Risler J, Vitek O, Martin DB (2009) Correlation between y-type ions observed in ion trap and triple quadrupole mass spectrometers. J Proteome Res 8 (9):4243-4251
Siepen JA, Keevil EJ, Knight D, Hubbard SJ (2007) Prediction of missed cleavage sites in tryptic peptides aids protein identification in proteomics. J Proteome Res 6 (1):399-408. doi:10.1021/pr060507u [doi]

Smedley D, Haider S, Ballester B, Holland R, London D, Thorisson G, Kasprzyk A (2009) BioMart--biological queries made easy. BMC Genomics 10:22. doi:1471-2164-10-22 [pii]
10.1186/1471-2164-10-22 [doi]

Swaney DL, Wenger CD, Coon JJ (2010) Value of using multiple proteases for large-scale mass spectrometry-based proteomics. J Proteome Res 9 (3):1323-1329. doi:10.1021/pr900863u [doi]

Szklarczyk D, Franceschini A, Kuhn M, Simonovic M, Roth A, Minguez P, Doerks T, Stark M, Muller J, Bork P, Jensen LJ, von Mering C (2011) The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. Nucleic Acids Res 39 (Database issue):D561-568. doi:gkq973 [pii]
10.1093/nar/gkq973 [doi]

Tabb DL, Ma ZQ, Martin DB, Ham AJ, Chambers MC (2008) DirecTag: accurate sequence tags from peptide MS/MS through statistical scoring. J Proteome Res 7 (9):3838-3846. doi:10.1021/pr800154p [doi]

Tabb DL, Saraf A, Yates JR, 3rd (2003) GutenTag: high-throughput sequence tagging via an empirically derived fragmentation model. Anal Chem 75 (23):6415-6421. doi:10.1021/ac0347462 [doi]

Taylor CF (2006) Minimum reporting requirements for proteomics: a MIAPE primer. Proteomics 6 Suppl 2:39-44.

Tharakan R, Edwards N, Graham DR (2010) Data maximization by multipass analysis of protein mass spectra. Proteomics 10 (6):1160-1171. doi:10.1002/pmic.200900433 [doi]

Thiede B, Lamer S, Mattow J, Siejak F, Dimmler C, Rudel T, Jungblut PR (2000) Analysis of missed cleavage sites, tryptophan oxidation and N-terminal pyroglutamylation after in-gel tryptic digestion. Rapid Commun Mass Spectrom 14 (6):496-502. doi:10.1002/(SICI)1097-0231(20000331)14:6<496::AID-RCM899>3.0.CO;2-1 [pii]
10.1002/(SICI)1097-0231(20000331)14:6<496::AID-RCM899>3.0.CO;2-1 [doi]

UniProt Consortium (2010) The Universal Protein Resource (UniProt) in 2010. Nucleic Acids Res 38 ((Database issue)):D142-148

Van Damme P, Martens L, Van Damme J, Hugelier K, Staes A, Vandekerckhove J, Gevaert K (2005) Caspase-specific and nonspecific in vivo protein processing during Fas-induced apoptosis. Nat Methods 2 (10):771-777. doi:nmeth792 [pii]
10.1038/nmeth792 [doi]

Vaudel M, Burkhart JM, Sickmann A, Martens L, Zahedi RP (2011) Peptide identification quality control. Proteomics 11 (10):2105-2114. doi:10.1002/pmic.201000704 [doi]

Villaveces JM, Jimenez RC, Garcia LJ, Salazar GA, Gel B, Mulder N, Martin M, Garcia A, Hermjakob H (2011) Dasty3, a WEB framework for DAS. Bioinformatics 27 (18):2616-2617. doi:btr433 [pii]
10.1093/bioinformatics/btr433 [doi]

Vizcaíno JA, Côté R, Reisinger F, Barsnes H, Foster JM, Rameseder J, Hermjakob H, Martens L (2010) The Proteomics Identifications database: 2010 update. Nucleic Acids Res 38 ((Database issue)):D736-742

Vizcaino JA, Martens L, Hermjakob H, Julian RK, Paton NW (2007) The PSI formal document process and its implementation on the PSI website. Proteomics 7 (14):2355-2357. doi:10.1002/pmic.200700064 [doi]

Vizcaino JA, Mueller M, Hermjakob H, Martens L (2009) Charting online OMICS resources: A navigational chart for clinical researchers. Proteomics Clin Appl 3 (1):18-29. doi:10.1002/prca.200800082 [doi]

Wang R, Fabregat A, Rios D, Ovelleiro D, Foster JM, Cote RG, Griss J, Csordas A, Perez-Riverol Y, Reisinger F, Hermjakob H, Martens L, Vizcaino JA (2012) PRIDE Inspector: a tool to visualize and validate MS proteomics data. Nat Biotechnol 30 (2):135-137. doi:nbt.2112 [pii] 10.1038/nbt.2112 [doi]

Woollard PM (2010) Asking complex questions of the genome without programming. Methods Mol Biol 628:39-52. doi:10.1007/978-1-60327-367-1_3 [doi]

Yates JR, III, Eng JK, McCormack AL, Schieltz D (1995) Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database. Anal Chem 67 (8):1426-1436

Yen CY, Houel S, Ahn NG, Old WM (2011) Spectrum-to-spectrum searching using a proteome-wide spectral library. Mol Cell Proteomics 10 (7):M111 007666. doi:M111.007666 [pii] 10.1074/mcp.M111.007666 [doi]

Yen CY, Meyer-Arendt K, Eichelberger B, Sun S, Houel S, Old WM, Knight R, Ahn NG, Hunter LE, Resing KA (2009) A simulated MS/MS library for spectrum-to-spectrum searching in large scale identification of proteins. Mol Cell Proteomics 8 (4):857-869

Yen CY, Russell S, Mendoza AM, Meyer-Arendt K, Sun S, Cios KJ, Ahn NG, Resing KA (2006) Improving sensitivity in shotgun proteomics using a peptide-centric database with reduced complexity: protease cleavage and SCX elution rules from data mining of MS/MS spectra. Anal Chem 78 (4):1071-1084. doi:10.1021/ac051127f [doi]

Zhang X, Li Y, Shao W, Lam H (2011) Understanding the improved sensitivity of spectral library searching over sequence database searching in proteomics data analysis. Proteomics 11 (6):1075-1085. doi:10.1002/pmic.201000492 [doi]

Zhang Z (2004) Prediction of low-energy collision-induced dissociation spectra of peptides. Anal Chem 76 (14):3908-3922

Zhang Z (2005) Prediction of low-energy collision-induced dissociation spectra of peptides with three or more charges. Anal Chem 77 (19):6364-6373

Figure legends

Figure 1: A proteomics experiment can roughly be divided into four parts: (1) From Ideas to Experiments, (2) From Acquired Data to Processed Results, (3) From Proteins to Knowledge, and (4) From Private Local Data to Public Online Data. Each with its distinct set of publicly available proteomics resources. Ideally there should also be a feedback loop from (4) back to (1). See main text for details.

Figure 2: Performing a pre-experiment *in silico* digest can provide valuable insight into the theoretically coverable areas of a given protein sequence. a) In this simple example the goal is to detect the two phosphorylations indicated. b) Performing an *in silico* digest tells us that, assuming the given enzyme properties, some areas of the protein sequence most likely can be detected (green), while others cannot (orange). From this we can infer that performing an experiment targeting the second modification will (most likely) be of little value using the selected protease.

Figure 3: Plot illustrating the database content and database information (points, left axis) and derived information ratio (bars, right axis) for six popular databases in the field of proteomics. Note that IPI has been discontinued on 27 September 2011. See main text for details.

Figure 4: BioMart makes it straightforward to combine data from multiple, unrelated databases. Here an example using the Ensembl BioMart is shown, where UniProt protein accession numbers are provided as input, and via just a few mouse clicks the proteins can be annotated with information from a long list of diverse resources.

Figure 1

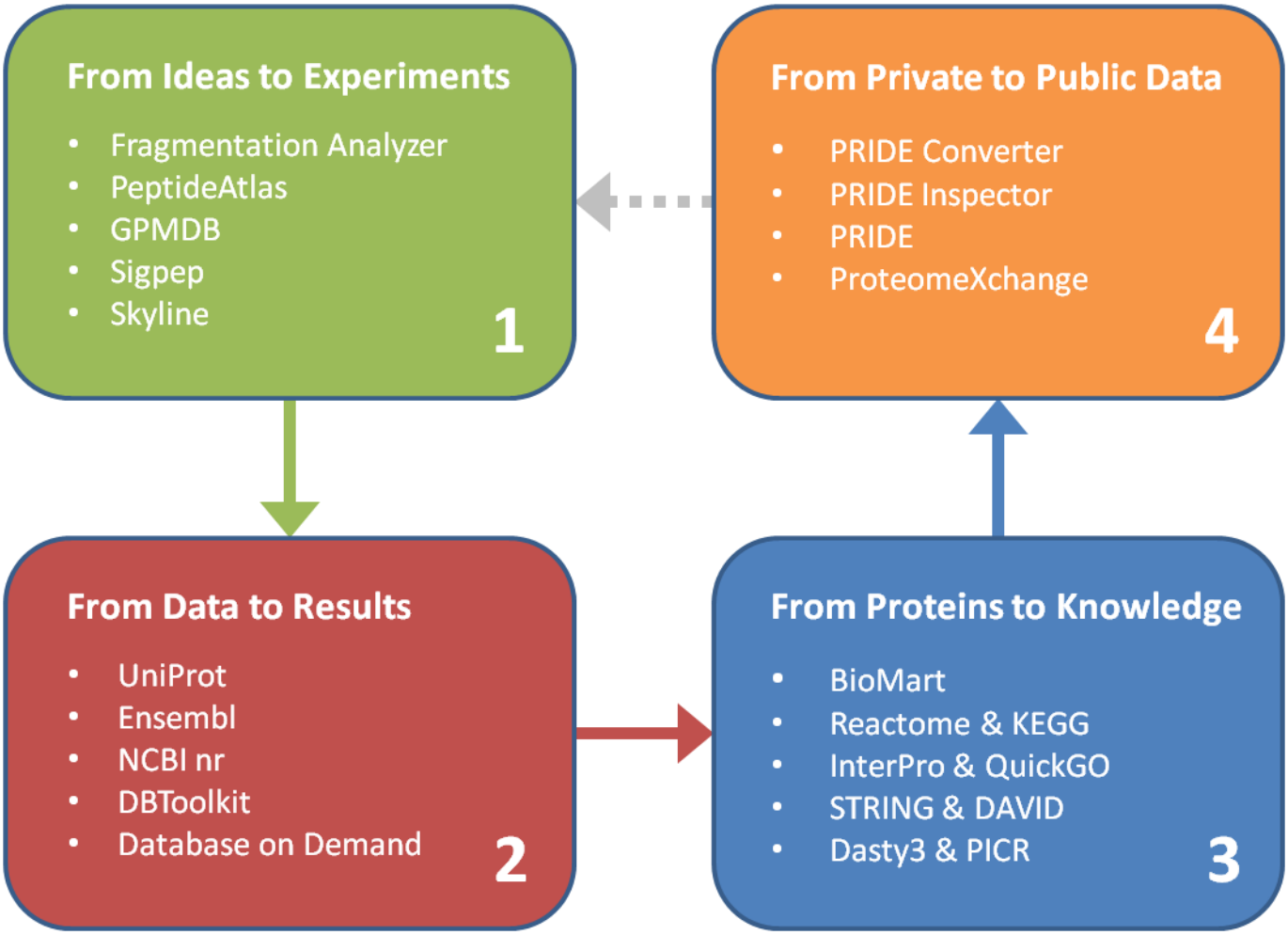

From Ideas to Experiments
Fragmentation Analyzer
PeptideAtlas
GPMDB
Sigpep
Skyline
1
From Private to Public Data
PRIDE Converter
PRIDE Inspector
PRIDE
ProteomeXchange
4
From Data to Results
UniProt
Ensembl
NCBI nr
DBToolkit
Database on Demand
2
From Proteins to Knowledge
BioMart
Reactome & KEGG
InterPro & QuickGO
STRING & DAVID
Dasty3 & PICR
3

Figure 2

Figure 3

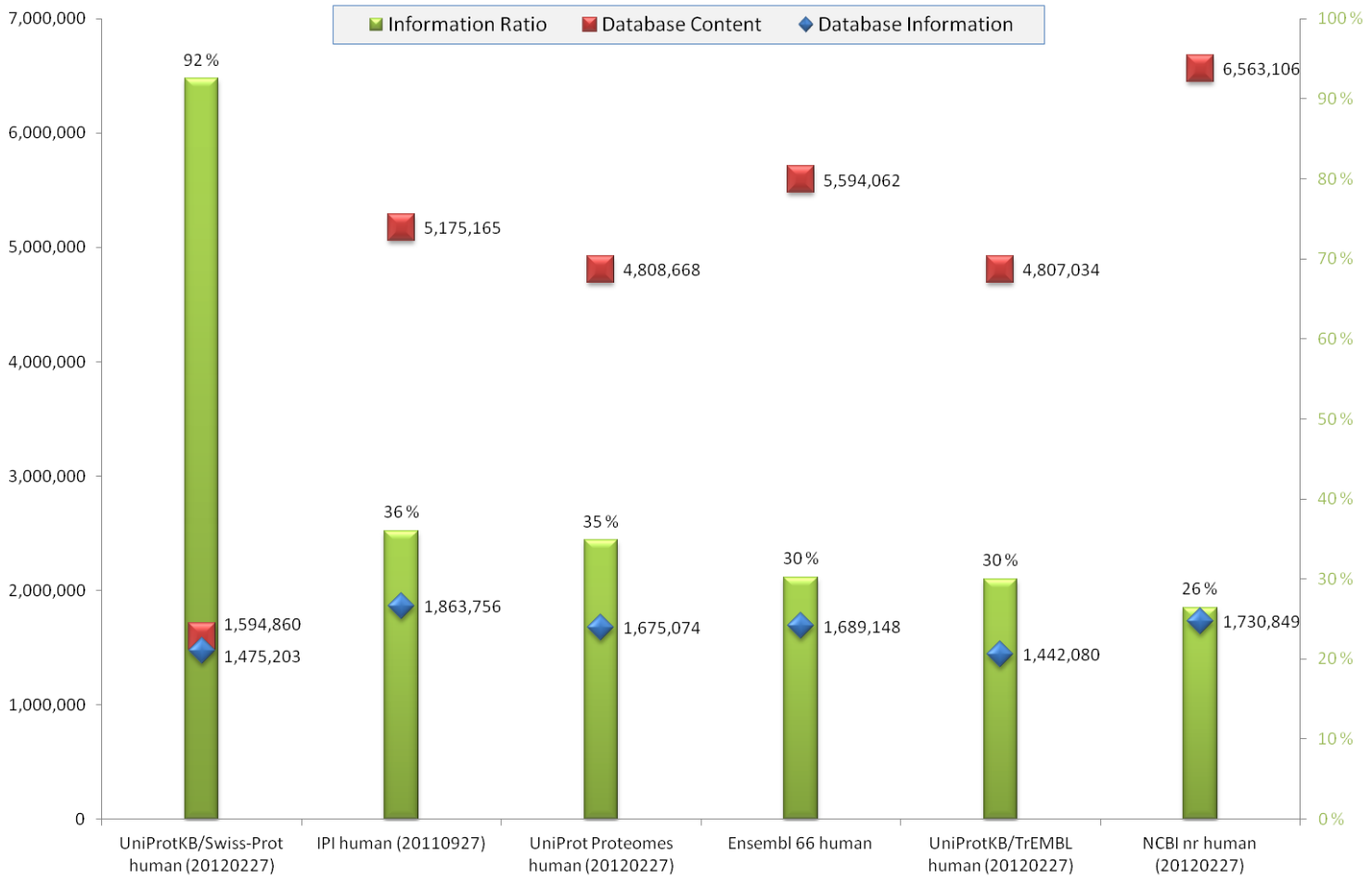
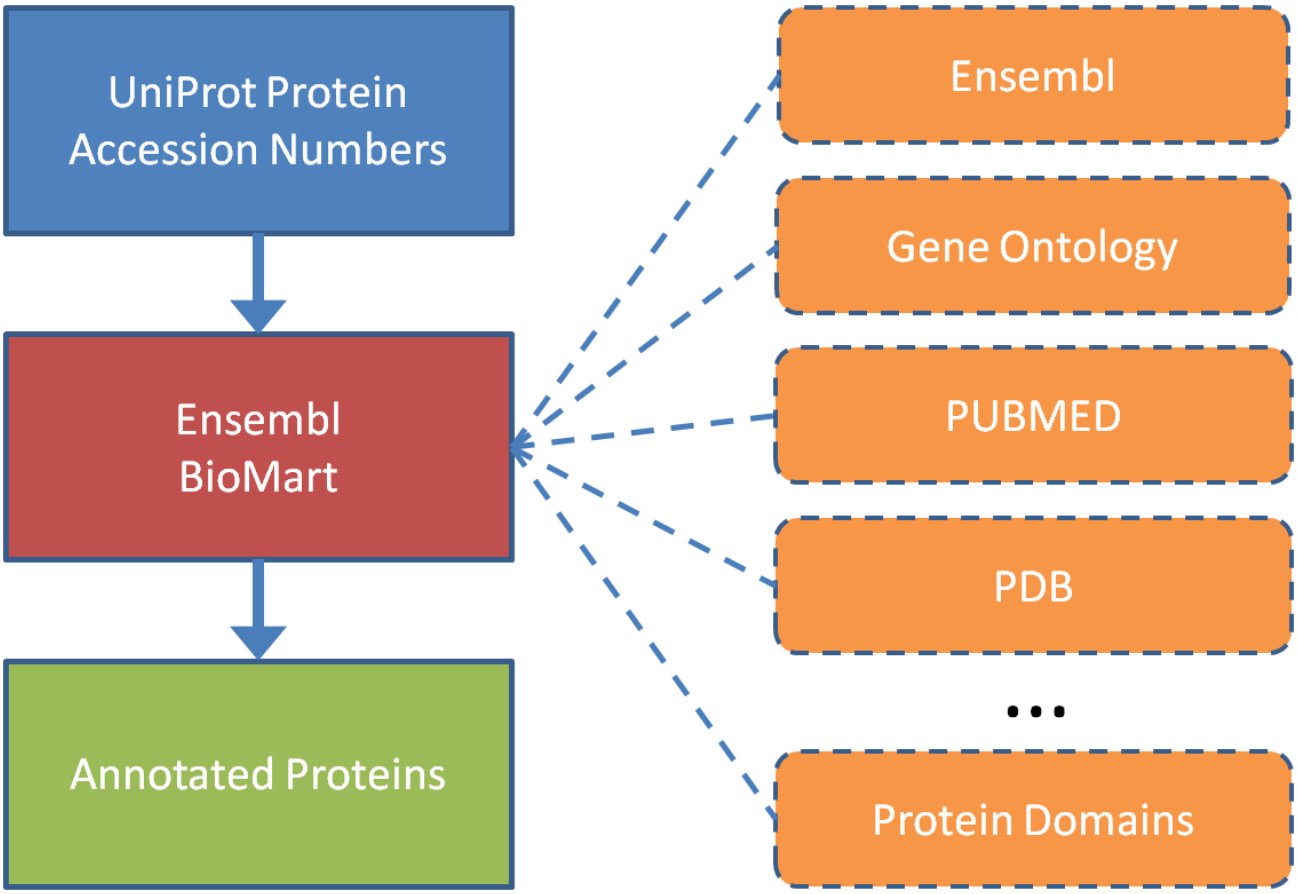

Information Ratio
Database Content
Database Information
7,000,000
6,000,000
5,000,000
4,000,000
3,000,000
2,000,000
1,000,000
0
100 %
90 %
80 %
70 %
60 %
50 %
40 %
30 %
20 %
10 %
0 %
92 %
1,594,860
1,475,203
36 %
5,175,165
1,863,756
35 %
4,808,668
1,675,074
30 %
5,594,062
1,689,148
30 %
4,807,034
1,442,080
26 %
6,563,106
1,730,849
UniProtKB/Swiss-Prot human (20120227)
IPI human (20110927)
UniProt Proteomes human (20120227)
Ensembl 66 human
UniProtKB/TrEMBL human (20120227)
NCBI nr human (20120227)

Figure 4


UniProt Protein Accession Numbers
Ensembl BioMart
Annotated Proteins
Ensembl
Gene Ontology
PUBMED
PDB
...
Protein Domains