Title:

Anatomy and evolution of database search engines - a central component of mass spectrometry based proteomic workflows

Running title:

Proteomic database search engines

Kenneth Verheggen[1,2,3], Helge Ræder[4,5], Frode S. Berven[6], Lennart Martens[1,2,3,], Harald Barsnes[4,6,7], Marc Vaudel[6,8,9]*

[1] Department of Medical Protein Research, VIB, B-9000 Ghent, Belgium
[2] Department of Biochemistry, Ghent University, B-9000 Ghent, Belgium
[3] Bioinformatics Institute Ghent, Ghent University, B-9052 Ghent, Belgium
[4] KG Jebsen Center for Diabetes Research, Department of Clinical Science, University of Bergen, Norway
[5] Department of Pediatrics, Haukeland University Hospital, Bergen, Norway
[6] Proteomics Unit, Department of Biomedicine, University of Bergen, Norway
[7] Computational Biology Unit, Department of Informatics, University of Bergen, Norway
[8] KG Jebsen Center for Diabetes Research, Department of Clinical Science, University of Bergen, Bergen, Norway
[9] Center for Medical Genetics and Molecular Medicine, Haukeland University Hospital, Bergen, Norway

[*] Corresponding author: Prof. Dr. Lennart Martens, Department of Medical Protein Research, Ghent University –VIB, A. Baertsoenkaai 3 B-9000 Gent Belgium. Tel: +32 9 264 93 58, Fax: +32 9 264 94 84, E-mail: lennart.martens@vib-ugent.be

Abstract

Sequence database search engines are bioinformatics algorithms that identify peptides from tandem mass spectra using a reference protein sequence database. Two decades of development, notably driven by advances in mass spectrometry, have provided scientists with more than thirty published search engines, each with its own properties. In this review, we present the common paradigm behind the different implementations, and its limitations for modern mass spectrometry datasets. We also detail how the search engines attempt to alleviate these limitations, and provide an overview of the different software frameworks available to the researcher. Finally, we highlight alternative approaches for the identification of proteomic mass spectrometry datasets, either as a replacement for, or as a complement to, sequence database search engines.

Introduction

Mass spectrometry has become the technique of choice for proteomics as part of the large scale analysis of proteins in complex samples (Aebersold & Mann, 2003; Aebersold et al., 2013). The direct measurement of intact proteins, so-called top-down proteomics, remains analytically challenging, impairing its application to complex mixtures (Kelleher, 2004). As illustrated in **Figure 1**, high throughput, shotgun proteomics approaches, also referred to as bottom-up proteomics, therefore digest proteins into peptides after reduction of disulfide bonds. The peptides are subsequently separated, generally by liquid chromatography (LC), and are then brought into the mass spectrometer source for ionization. Two ionization techniques are principally used in proteomics: electrospray ionization (Fenn et al., 1989) and matrix assisted laser desorption ionization (MALDI) (Karas et al., 1985; Tanaka et al., 1988; Soltwisch et al., 2009).

Once ionized, the peptides are subjected to tandem mass spectrometry, in which the mass over charge ratios (m/z) of the peptides entering the mass spectrometer are first scanned, yielding so-called MS or MS1 spectra. Selected peaks (typically the ones with the highest signal) are then isolated and fragmented in the next step, and the m/z of the obtained fragment ions is measured and reported in MS/MS or MS2 spectra. In addition to the fragmentation of a specific precursor peptide m/z, an approach called data dependent acquisition (DDA) (Mann et al., 2001), one can also fragment a larger subset, or even all, of the eluting peptides in an approach called data independent acquisition (DIA) (Doerr, 2015; Kuharev et al., 2015). In the latter, the MS1 scan is not necessarily acquired (Gillet et al., 2012). Three fragmentation methods are widely used in proteomics to generate MS2 spectra,: (1) collision induced dissociation (CID) (Wells & McLuckey, 2005), (2) higher-energy collision induced dissociation (HCD) (Olsen et al., 2007), and (3) electron transfer dissociation (ETD) (Syka et al., 2004), each providing distinct types of fragment ions.

As illustrated in **Figure 1**, the experimental data obtained from shotgun proteomic experiments thus mainly consist of the acquired MS1 and MS2 spectra. With the advent of high acquisition rate instruments, proteomics datasets have reached sizes ranging from thousands to millions of spectra, rendering their manual interpretation impossible. As a result, the interpretation heavily relies on the use of bioinformatics. Three main strategies have been established for the identification of peptide-derived tandem mass spectra (McHugh & Arthur, 2008): (1) sequence database searching, where the spectra are searched against a database of reference protein (or peptide) sequences, (2) spectrum sequencing, where amino acid sequences are directly inferred from the spectra, and (3) spectral library searching, where the spectra are searched against a library of spectra from known compounds.

This review describes the principles of the database searching paradigm. After first presenting the main principles and introducing the various software implementations, the so-called search engines, we detail possible pitfalls of database searching, and provide solutions that alleviate these where available. Finally, we introduce some of the advanced operating modes for improved sample characterization.

I. A brief history of database search engines

Figure 2 illustrates the concept of database searching, the matching of experimentally obtained MS2 spectra against a sequence database. The MS2 spectra to search first undergo preprocessing, which for most search engines consists in filtering out low intensity peaks to retain only the most intense peaks. The sequence database against which the spectra will be matched is also processed, by *in silico* digestion and fragmentation of the sequences. This mimics the experimental enzymatic cleavage and fragmentation of sample proteins, and provides theoretical MS2 spectra that can then be compared to the experimental MS2 spectra. A set of search parameters provided by the user tunes the specifics of this comparison. Finally, the quality of each comparison is evaluated using algorithm-specific scores. The result is a list of

peptide candidates for each spectrum, so-called peptide -spectrum matches (PSMs), along with their respective score.

A standard format, called mzIdentML (Jones et al., 2012), has been developed to encapsulate and exchange peptide and protein identification results from search engines. The mzIdentML format can be directly exported by many of the search engines, and converters are also available (http://www.psidev.info/tools-implementing-mzidentml).

Database searching emerged as a fast and reliable alternative to spectrum sequencing, and was pioneered by the SEQUEST algorithm (Eng et al., 1994). Subsequently, the commercial alternative Mascot (Perkins et al., 1999), was quickly adopted by the community for its server-based infrastructure, and the simplicity of interpretation of its scores. Numerous search engines have since become available to the scientific community, including multiple free and open-source alternatives. **Table 1** lists all search engines (to the best of our knowledge at the time of writing) ordered by date of publication or availability. This table shows how the number of search algorithms has steadily increased since the first publication of SEQUEST, consistent with the growing need for better and faster algorithms that are capable of handling ever larger datasets. In addition, the total number of citations (from 1994 to 2016) according to Thomson Reuters™ Web of Science™ is provided as a rough indicator of community adoption.

The number of citations per year for the most common algorithms, based on the original publication, is also displayed as a timeline in **Figure 3**. Despite the relative inaccuracy of this usage metric, it appears that search engine usage has been dominated by the original search engines, SEQUEST and Mascot, followed closely by the early open-source alternatives OMSSA (Geer et al., 2004) and X! Tandem (Fenyo & Beavis, 2003). Interestingly, the total number of search engine citations seems to have stagnated in the past five years. The dramatic increase in the share of Andromeda (Cox et al., 2011) as part of MaxQuant (Cox & Mann, 2008) shows the importance of search engine integration in global data interpretation pipelines. Finally, the increasing prevalence of other algorithms (a category that includes recent engines such as

MS Amanda (Dorfer et al., 2014) and MS-GF+ (Kim & Pevzner, 2014)) shows the interest of the community for more innovative approaches. For further details on the citations statistics please refer to the supplementary material.

II. Practical use of a search engine

The input for a search consists of peak lists containing the spectra to search, a sequence database, and the search settings used to tailor the search to the experimental setup.

A. Peak lists

The raw mass spectrometer output contains all data acquired by the mass spectrometer in a vendor-proprietary format (Martens et al., 2005). Before these data can be analyzed by external software, the output has to be converted into an open and preferably standardized format. The reference format for mass spectrometry files is mzML (Martens et al., 2011). However, due to the complexity and size of mzML files, many search engines operate on simpler formats that contain only the MS2 peak lists, along with the precursor ion m/z, intensity and charge. The most common formats are dta, pkl, ms2, and mgf as reviewed in (Deutsch, 2012). Note that file format conversion can be conducted easily by ProteoWizard (Chambers et al., 2012), and also that some search engines are able to read the vendor formats directly, either through use of the vendor application programming interfaces (APIs) or by incorporating ProteoWizard as part of their software package.

The recorded data often include peaks that are not derived from peptides, and these can impair the identification efficiency of a search (Du et al., 2008). To address this issue the raw spectra can be submitted to specialized algorithms that improve spectrum quality and that reduce the prevalence of non-peptide derived spectra/peaks (Ning & Leong, 2007; Barbarini & Magni, 2010; Sheng et al., 2015). As detailed in (Renard et al., 2009), this preprocessing can be divided into three categories: (1) spectral quality scoring based on spectrum features and/or

clusters; (2) precursor pre-processing, which can improve precursor charge and isotope inference, as well as mass accuracy (Hsieh et al., 2010); and (3) MS2 spectrum processing which includes the merging of spectra, baseline reduction, noise filtering, and deisotoping. All of these steps can for instance be carried out in the OpenMS open-source proteomic software framework (Sturm et al., 2008). Spectrum processing options were also recently implemented in ProteoWizard as part of the raw files conversion (French et al., 2015). However, advanced preprocessing has become less relevant with the advent of high-resolution mass spectrometers, and of instruments equipped with advanced signal processing units that provide data that is directly interpretable by search engines.

Most search engines expect spectra in the form of peak lists, where a peak is represented as an (m/z, intensity) pair. It is thus important to verify that the spectrum files have been output in *centroid mode* (Deutsch, 2012). If the MS2 peaks take the form of the original, bell-shaped detector trace curve (referred to as *profile mode* data), a peak-picker should be applied, for example *via* OpenMS (Lange et al., 2006) or ProteoWizard.

B. Alternative data sources

In a global effort for scientific transparency, an increasing number of researchers now share the experimental data that support their findings. Vast amounts of proteomics data are thus available to the community, and can, for example, be used to provide preliminary results while setting up an experiment (Barsnes & Martens, 2013). However, in order to fully exploit such data, it can be useful to update the database, or to use different algorithms or settings. In this way, the original data can be reprocessed, and possibly even repurposed (Vaudel et al., 2015). Spectra from online repositories can thus be downloaded and reprocessed as if acquired locally. Numerous public repositories contain data for such reanalysis (Fenyo et al., 2010; Perez-Riverol et al., 2014), including the PRoteome IDEntifications database (PRIDE) (Martens et al., 2005; Vizcaino et al., 2016), the Global Proteome Machine Database (GPMDB)(Craig et al., 2004), MaxQB (Schaab et al., 2012), Massive (http://massive.ucsd.edu), and PeptideAtlas

(Desiere et al., 2006). PRIDE data, for example, can be reprocessed seamlessly using multiple search engines *via* PeptideShaker (Vaudel et al., 2015) in combination with SearchGUI (Vaudel et al., 2011).

C. Protein sequence databases

As illustrated in **Figure 2**, search engines rely on a database of reference sequences for the identification of peptides. Sequence databases can be obtained from public resources such as the UniProt knowledgebase (UniProt, 2015), The National Center for Biotechnology Information (NCBI) Reference Sequence (RefSeq) Database (Pruitt et al., 2005), or the DNA Data Bank of Japan (DDBJ) (Mashima et al., 2016).

UniProt is the result of an effort to centralize the protein sequences in complete proteomes, along with relevant knowledge about these proteins as extracted from the available literature. As such, it serves a standardized hub for protein sequences and associated information. It is divided into two distinct subsets: (1) UniProt-KB/Swiss-Prot, which contains manually curated and annotated proteins with an evidence ranking based on literature review; and (2) UniProt/TrEMBL, which contains non-reviewed, automatically inferred sequences.

The NCBI reference database provides a non-redundant collection of genomic, transcriptomic, and proteomic sequences, and DDBJ is an online repository that contains both human genotype and phenotype data. Alternatively, specialized databases can be found for specific species, diseases, or sub-proteomes (Hong et al., 2008; Reddy et al., 2009; Lamesch et al., 2012; McQuilton et al., 2012; Howe et al., 2013; Harris et al., 2014; Gaudet et al., 2015; Urban et al., 2015).

The choice of the sequence database to use has a strong impact on the results of the search. Indeed, it is important to note that it is impossible for a search engine to identify peptides from proteins that are not present in the selected database. The search database should thus cover the proteins that are likely to be present in the sample as comprehensively as

possible. If not all proteins in the sample are represented in the database, the spectra obtained from such unexpected proteins can be matched incorrectly to other proteins in the database, which results in false positive identifications (Foster, 2011; Knudsen & Chalkley, 2011). This is notably the case for common contaminants, *e.g.* human keratin proteins in non-human samples, which can lead to incorrect biological conclusions if misinterpreted (Bern et al., 2009; Ghesquiere et al., 2011). Known contaminants should therefore be included in the database, alongside the protein sequences of interest. A useful list of common contaminants can be found in the common Repository of Adventitious Proteins (cRAP) (www.thegpm.org/crap).

It is important to keep in mind that very large sequence databases also affect the search sensitivity, as further discussed below. It is thus recommended to tailor the database to the species under study when possible. However, complete protein sequences are not available for all samples. In such cases, genomic or transcriptomic data can instead be used to infer a suitable search database (Dove, 1999; Nesvizhskii, 2014; Menschaert & Fenyo, 2015). Closely related species can also be used as a substitute database, or one could be derived from elements of sequences that are conserved between multiple species (Penel et al., 2009).

Note that protein sequence databases are regularly updated; for instance, an updated version of UniProtKB is released monthly. It is always recommended to use the latest available database version, and to use the same version throughout a project in order to avoid biases when comparing samples. The database version should be documented in-house and, for the sake of reproducibility, in scientific publications.

Finally, it is important to highlight that shotgun proteomics identifies peptides and *not* proteins. The proteins are only inferred from the peptide evidence, as will be further discussed below.

III. Management of the search space

The search space of a proteomics database search engine is defined as the collection of all possible peptide and fragment ions that need to be taken into account when a spectrum is searched. The number of possible peptides is the number of peptides that can be matched to a precursor m/z in the experimental data. It is influenced by the tolerance used to search the data, and therefore the instrument resolution, but also by all search parameters that can influence the number of possible peptide m/z matching measured precursor m/z. For example, adding a variable modification increases the concentration of possible m/z and therefore the chances to match a precursor m/z.

Additionally, when a modified peptide matches a precursor mass, all possible localization combinations have to be tested and scored. It should also be noted that a substantial fraction of the spectra that end up as unidentified may arise from modified peptides (Chick et al., 2015; Bogdanow et al., 2016).

To illustrate this effect, we used the example dataset of the CompOmics Proteomics Bioinformatics tutorials, a one hour gradient measurement of a HeLa trypsin digest on a Q Exactive, see (Vaudel et al., 2014) for details, and increased the search space in 14 different ways: (1) *Isoforms*, the isoforms of the canonical sequences were included in the database; (2) *Trembl*, UniProt/TrEMBL sequences were included in the database, (3) *Vertebrates*, a non-species-specific database was used including all vertebrates canonical sequences from UniProt; (4) *4 mc*, up to four missed cleavages were allowed; (5) *Semispecific*, semi-specific cleavage was allowed; (6) *Variable Cmm*, cysteine carbamidomethylation was considered as variable; (7) *Phosphorylation*, variable phosphorylation of S, T, and Y was included; (8) *ABY*, a-ions were included in the fragmentation of peptides; (9) *ABCXYZ*, all fragment ions were considered; (10) *MS2 0.5 Da*, the MS2 tolerance was changed to 0.5 Da; (11) *MS1&2 0.5 Da*, both MS1 and MS2 tolerances were changed to 0.5 Da; (12) *-4 to +4 Da*, an isotopic shift of -4 to +4 Da was allowed for the precursors; (13) *1 to 4*, peptides of charge 1 to 4 were included in the search; and (14) *1 to 6*, peptides of charge 1 to 6 were included in the search. Importantly, the minimal peptide

m/z considered was set to 500 m/z, and up to five modification sites were tested per peptide. For more details on the generation of these data, see mvaudel.github.io/onyase/review_figure/review_figure.html.

Figure 4A shows the density of the number of peptides matching per precursor in the different search space enlargement cases sorted from lowest to highest median. As expected, the ms2 settings, fragment ions considered and tolerance do not alter the number of peptides per precursor distribution. Similarly, the inclusion of charge 1 peptides and isoforms, do not substantially alter the number of peptides per precursor. The first increase is observed when including higher charges, higher numbers of missed cleavages, variable cysteine carbamidomethylation, and TrEMBL sequences. Including larger isotope tolerance clearly increases the number of peptides considered.

Similarly, adding phosphorylation as variable modification increases the median number of peptides per precursor by almost one order of magnitude. As visible from the density and inter-quartile distance, the phosphorylation site combinations increase the span of the distribution above all other settings, and without limitation on the number of sites considered the size of the search space can become challenging to manage. The semi-specific, non-species-specific, and relaxed MS1 tolerance all increased the median number of peptides per precursor by over an order of magnitude.

A direct consequence of a large search space is longer search times due to the number of possibilities to evaluate. Figure 4B shows the number of peptides to evaluate in every search space enlargement condition listed above. As expected, the total number of peptides evaluated follows the increase in median number of peptides per precursor. With the notable exception of datasets searched with a low MS2 resolution, most peptides will however be rapidly discarded by the search engines because no fragment ion could be matched, as indicated by a hyperscore of 0 (Fenyo & Beavis, 2003).

Another direct consequence is the increased probability to match a spectrum incorrectly with a high score, and thus either a higher prevalence of false identifications, or a lower identification rate (Colaert et al., 2011; Muth et al., 2015). This is illustrated Figure 4C with the distributions of the scores of decoy matches, that are by design incorrect. One can clearly see an increase of the score attributed to these false hits as the search space increases, the search space enlargements *via* high mass tolerance having the most prominent effect.

The presence of false positive matches with high scores makes it more difficult for search engines to distinguish the true hits from the others. Consequently, after estimation of E-values from the hyperscore distributions, one can observe a drop in the number of identified PSMs at a given false discovery rate (FDR). To maximize the identification rate, the search space is therefore tailored to best represent the proteins present in the sample without bias using the search parameters. **Table 2** lists the standard search parameters encountered in most search engines.

The first search space confinement is achieved by adaptation to the sample under study, mainly by limitation of the sequence database to the sample content and thus only use sequences from a single species when possible (Yen et al., 2006; Borges et al., 2013; Muth et al., 2013). Because the inclusion of all known protein isoforms as separate entries in the sequence database increases the search space, curated consensus sequences that present only one protein sequence per gene are often used. As already mentioned, the presence of potential protein modifications also greatly influences the search space. Protein modifications can be categorized into three groups: (1) *in vivo* modifications that carry biological or functional information, *e.g.* phosphorylation; (2) *in vitro* artefactual modifications that occur spontaneously in sample handling, *e.g.* methionine oxidation or the formation of amino-terminal pyroglutamate; and (3) *in vitro* intended modifications that are part of the sample preparation protocol, *e.g.* isobaric or other isotopically labeled tags.

Proteins that carry specific biological or functional modifications are often present at substoichiometric abundances and are therefore rarely detectable without enrichment (Nielsen et al., 2006; Millioni et al., 2011; Loroch et al., 2013; Solari et al., 2015). In order to avoid unnecessary enlargement of the search space, it is therefore recommended to refrain from the inclusion of protein modifications that fall below the limit of detection of the experimental setup. Artefactual or intended modifications have a much higher prevalence and therefore need to be accounted for in the search. However, many of the intended modifications will occur with a very high efficiency, typically >95%. All possible target residues can then be considered as modified. Such *fixed*, or *static*, modifications do not increase the search space as these simply replace the affected residue by the corresponding modified residue and are thus systematically included at all possible sites. This in contrast to *variable*, or *dynamic*, modifications, where both the modified and unmodified version of each affected residue has to be considered, which increases the search space exponentially.

The search space is also most often adapted to the enzyme used to digest the proteins. This means that only peptides that abide by the enzyme cleavage rules are considered. To adapt to the efficiency of proteolytic cleavage, a certain number of allowed missed cleavages is allowed. Similarly, semi-specificity can be used to account for unanticipated cleavages. Allowed missed cleavages and semi-specificity both lead to a larger search space.

Finally, the search space can be tailored to the performance of the instrument used. This is achieved by adaptation of the m/z tolerances. For most search engines, mass tolerances are set at both the precursor and fragment ion levels, and any theoretical peptide and fragment ions that fall outside of these tolerance ranges are excluded. Hence, less restrictive mass tolerances induce a large search space, while stricter tolerances reduce the search space. Search engines also allow the search space to be tailored to the fragmentation method used. The simplest setting includes the selection of the amino-terminal and carboxy-terminal fragment ions to consider (Roepstorff & Fohlman, 1984; Johnson et al., 1987). More advanced parameters

include the isotope range considered, the selection of neutral losses, or the use of expert fragmentation models (Skilling et al., 2004; Paizs & Suhai, 2005; Klammer et al., 2008; Neuhauser et al., 2012). The search space can also be tailored to the possible peptide charges expected from the applied ionization method. MALDI ionization yields singly charged ions, while higher charges (two to four) are typically considered for electrospray ionization.

Note that many modern search engines support the selection of predefined settings, for example for high or low resolution instruments, or for different fragmentation models. This allows for a simpler setup of the search. Some settings can also be optimized by the search engines themselves, for example through the use of machine learning approaches (Barla et al., 2008; Yang et al., 2012). The optimization of the search space is a complex multi-variable optimization procedure, and one can easily be overwhelmed by the number of settings available for each search engine. It is therefore recommended to start from a set of standard settings and then study the influence of a change in a specific setting to best model the sample, protocol and acquisition (Vaudel et al., 2011; Muth et al., 2015). While optimal values can vary between samples or instruments, the variability for a single setup is not substantial and usually does not require a complete optimization process for every experiment. Some quality control procedures, such as the verification of the efficiency of chemical labelling can however, be mandatory prior to publication, for a detailed example see (Aasebo et al., 2014).

IV. Spectrum matching

Different approaches have been established to match spectra to theoretical peptides. These can be categorized according to the approach used to infer the theoretical spectra (Sadygov et al., 2004): (A) descriptive, (B) interpretative, or (C) stochastic.

A. Descriptive

Descriptive approaches are based on theoretical models of peptide fragmentation. A theoretical spectrum is generated for each peptide based on specified rules, and a similarity score is calculated between the theoretical and experimental spectrum. SEQUEST (Eng et al., 1994) is one example of a search engine that uses this descriptive approach. The number of predicted fragments that are present in the experimental spectrum determines the quality of the match. Peptide fragmentation models can be very simple, with fixed, arbitrary intensity values for all b- and y-ions (Sadygov et al., 2004).

B. Interpretative

Interpretative approaches rely on the assumption that peptides can be identified from a series of fragment ions that are manually or automatically retrieved from the spectra. Each peptide candidate is partitioned into an amino acid sequence flanked by masses of unknown composition, and the algorithm then attempts to match this amino acid sequence and its masses to the search space. This approach was pioneered by PeptideSearch (Mann & Wilm, 1994), which showed that such extracted sequence "islands" could be matched to a database. Note also that through partial matching with sequence tags the search space can be significantly reduced, for example, as seen in open modification searching (Na et al., 2012).

A more recent implementation of this strategy can be found in TagRecon (Dasari et al., 2010) that can be used to identify mutations that occur in the masses that flank the extracted amino acid sequence, and in MS-GF+ (Kim & Pevzner, 2014) that is designed to cope with the emergence of novel mass spectrometry techniques and more accurate data.

C. Stochastic

In stochastic approaches, libraries of already identified spectra are used to model the theoretical spectrum of a given peptide. This model is thus specifically tailored to the instrument used and its performance. An example of such an algorithm is SCOPE (Bafna & Edwards, 2001). Stochastic models require large training datasets to determine the likelihoods

and features of tandem mass spectra. Typically, these models are devised using machine learning and are based on the intrinsic properties of existing data (Kelchtermans et al., 2014). The model is however vulnerable to fluctuations in mass spectrometer performance and experimental setup.

V. Calculation of peptide to spectrum scores

For all PSMs, a scoring algorithm is employed to provide a quality metric for the matches between a spectrum and its proposed progenitor peptides. This score is then used to rank the results and retain only the best peptide-to-spectrum matches (PSMs). This can be achieved by (A) correlation and ion scores, or (B) statistical and probabilistic approaches (Sadygov et al., 2004).

A. Correlation and ion scores

This method was pioneered by SEQUEST (Eng et al., 1994), where scores are calculated in two stages: (1) a preliminary score, Sp, is calculated as the summed intensity of all peaks that match the predicted fragment ion masses; and (2) for the 500 top candidates, a cross-correlation value of the experimental *versus* the theoretical spectra is calculated (the XCorr score), and normalized (the Cn score). Similarly, X! Tandem bases its scoring on the hyperscore, where the cross correlation is further multiplied with the factorial of the number of matched peaks. It is important to note that these correlation-based scores are deterministic, they will always be the same between a given peptide and spectrum.

The main drawbacks of this approach are: (1) its dependence on the quality of the spectra, the peptide length, the considered modifications and charge states, (2) the difficulty to interpret the scores, and (3) the computational load of the cross correlation analysis - note however, that recent implementations of this scheme rely on a faster implementation (Eng et al., 2008).

B. Statistical and probabilistic approaches

Most search engines estimate the significance of deterministic scores in the context of the search. For example, X! Tandem transforms the deterministic hyperscore into an E-value that will depend on the search space. SEQUEST uses a relative score that represents the difference between the Cn of the best and second best peptide candidates for a particular spectrum (the ΔCn score). Statistical and probabilistic approaches hence estimate the probability of a match to occur randomly in the search space. This approach, pioneered by Mascot (Perkins et al., 1999), has become very popular for its simplicity of interpretation. Mascot's algorithm is however, kept a trade secret. Open search engines such as X! Tandem (Fenyo & Beavis, 2003) and OMSSA (Geer et al., 2004) were later released as open-source alternatives that follow a similar approach, with different distributions to model the population of matching scores. More recently, Andromeda (Cox et al., 2011) and MS Amanda (Dorfer et al., 2014) have become available as additional free alternatives, based on related probabilistic models that perform similar to Mascot. Andromeda can either be used as a standalone search engine or as a part of MaxQuant (Cox & Mann, 2008).

In contrast to ion scores, probabilistic scores depend on the experiment and search space. When the search space grows the distribution of scores from random matches spans a wider range, as illustrated Figure 4C. The difference in score between the correct matches and the random identifications therefore decreases, making it harder to distinguish the correct identifications from the others. This loss of discrimination power yields a lower search sensitivity, and eventually a lower identification rate as illustrated Figure 4D.

VI. Advanced search strategies

Advanced search strategies have been developed to circumvent the limitations of enlarged search spaces, and thus increase the identification coverage of proteomic workflows. In multi-stage search strategies, spectra are searched iteratively, where the result of one iteration is used to select seed peptides or proteins presumably present in the sample, and then only these are considered in the next iteration, as illustrated with the feedback loop to the protein list in **Figure 2**.

X! Tandem employs a second search stage called *refinement* where the result of the first search is used to establish a set of high confidence proteins. The spectra are subsequently searched again using relaxed settings (semi-enzymatic cleavage, higher number of missed cleavages, additional modifications, *etc.*), but only against the set of proteins detected in the first iteration. As a result, the search engine quickly identifies more peptides for these proteins, circumventing search space enlargement issues. A similar procedure is called *error-tolerant search* in Mascot (Creasy & Cottrell, 2002) and *iterative search* in OMSSA. However, the arbitrary selection of confident proteins from one stage to the other, can introduce a bias in the scores of matches in multiple stage strategies, which ultimately impairs the reliability of downstream peptide-level false discovery rate estimation strategies (Everett et al., 2010; Bern & Kil, 2011). The results of multiple stage searches should thus be interpreted with care (Jeong et al., 2012).

A similar strategy to reduce the search space is to iteratively filter out spectra and search the remaining non-identified spectra against another database (Noble, 2015). In such cascaded searches, the search space is gradually shifted, starting from background peptides to peptides of interest, which reduces the prevalence of random false positive matches (Kertesz-Farkas et al., 2015). However, the lack of competition between hits induced when the database is tailored towards a given hypothesis is known to generate false positives matches of non-random nature (Colaert et al., 2011). Here again, the possible underestimation of error rates should thus be kept in mind when the results are interpreted.

More than 100,000 peptides elute in a typical proteomic shotgun experiment (Michalski et al., 2011), consequently, two different peptides may be isolated, fragmented, and recorded simultaneously, which results in so-called chimeric spectra (Houel et al., 2010). Almost all search engines are however, designed to identify only a single peptide per spectrum. Consequently, only one peptide is identified, generally the one that corresponds to the dominant peaks in the fragmentation spectrum. (It should be noted however, that if the difference in precursor mass between the peptides included in the chimeric spectrum is larger than the allowed mass tolerance, than the precursor mass will be the primary determinant for identification.) To alleviate this problem, some search engines implement another advanced procedure that consists of a removal of the signal of the identified peptide from the spectrum, followed by a re-search of the remaining unidentified peaks for possible co-fragmented peptides. This method is notably implemented in Andromeda where it was reported to provide up to 10 % additional hits (Cox et al., 2011).

Finally, error-tolerant searches can be used as a strategy to identify peptides outside of the search space, as additional degrees of freedom are allowed, *e.g.* mass differences induced by unexpected modifications, sequence variants, or non-enzymatic cleavage sites. This strategy is notably available in Mascot (Creasy & Cottrell, 2002) (www.matrixscience.com/help/error_tolerant_help.html). A similar approach is to search with very high tolerances and filter matches *a posteriori* (Beausoleil et al., 2006). Mass tolerant searches can be combined with clustering of PSMs to match peptides to a wide range of modifications and sequence variants (Chick et al., 2015).

VII. Conclusion and overview

Database searching has become the identification method of choice in proteomics. In a global attempt at enhancing the performance of searches, multiple implementations have been made available to researchers. Each of these offers different variations on the basic principles

presented in **Figure 2**. The increasing number of available search engines (**Figure 3**)**,** highlights the highly dynamic development of this field of research. Additional resources have also become available for upstream and downstream processing of the data, and search engines have been integrated in software environments that allow the design of complex workflows, as for example in the Trans Proteomic Pipeline (TPP) (Deutsch et al., 2010), OpenMS (Sturm et al., 2008), MaxQuant (Cox & Mann, 2008), and Pladipus (Verheggen et al., 2016). Among the downstream procedures, several are notable for an intricate link to the search engines: (1) error rate estimation, (2) multiple algorithm integration, and (3) protein inference.

Error rate estimation is generally achieved by a false discovery rate (FDR) estimation, which provides an estimate of the share of incorrect matches retained. This is important because, as detailed above, search engines provide only a list of candidate peptides along with a score for each peptide. The researcher must somehow establish a certain score cut-off from these results to control the quality of the retained set, as reviewed in (Vaudel et al., 2012). Two main methods are available for this control of error rates: search engine score modeling (Keller et al., 2002) and the use of target/decoy databases (Elias & Gygi, 2007). A third approach, which relies on lower ranked hits has also been proposed recently (Gonnelli et al., 2015). For a detailed review on error rate estimation procedures in proteomic identification results, see (Nesvizhskii, 2010).

In order to benefit from the complementarity of available search engines, methods for the integration of multiple search engines have been established, and these can provide a substantial gain in identification coverage (Yen et al., 2006; Searle et al., 2008; Yu et al., 2010; Shteynberg et al., 2013). In addition, search engines can also be combined with alternative identification approaches to overcome the drawbacks of using a sequence database. The first useful alternative is the use of spectral libraries, where newly acquired experimental spectra are matched to previously identified spectra (Craig et al., 2006; Bandeira et al., 2007; Lam et al., 2007; Frank et al., 2011). The second alternative approach that can be combined with search

engines is *de novo* sequencing (Seidler et al., 2010; Allmer, 2011; Medzihradszky & Chalkley, 2015). In *de novo* sequencing, the peptide amino acid sequence is partially or completely inferred from the spectrum. While computationally intensive, this method presents the advantage of being virtually unbiased toward sequence databases. The combination of these approaches can, for example, be achieved using IDPicker (Ma et al., 2009).

As detailed in the introduction, shotgun proteomics only allows the identification of peptides. Before drawing conclusions at the protein level, the presence of a protein must therefore first be inferred from the identified peptides. This task is made complex by the presence of peptides shared between proteins (Nesvizhskii & Aebersold, 2005). Moreover, this problem propagates to the downstream tasks in a proteomic bioinformatics workflow, for instance protein identification error rate estimation, protein quantification, and post-translational modification studies. The inference of proteins from peptides is particularly complicated in the case of multiple search engine workflows, due to the inconsistencies of peptide-to-protein association between algorithms, and for multiple samples or fractionated samples, where the protein inference step must take into account the complexity of the experimental design (Vaudel et al., 2013).

Such newer and more complex search strategies, along with the growth in data set size, make it increasingly difficult to conduct a search within a reasonable time frame. This is particularly problematic in the case of proteogenomics studies (Nesvizhskii, 2014) and metaproteomics (Muth et al., 2013). Distributed computing can help overcoming some of these limitations, through the use of grid or cloud computing (Verheggen et al., 2014). This way, extensive processing power can be made available to the community at large, notably through the establishment of dedicated environments, like the Galaxy project (Giardine et al., 2005; Boekel et al., 2015). It is also possible to distribute tasks on a local cluster of computers, making it possible for most labs to carry out demanding searches even with limited informatics resources (Verheggen et al., 2016). Moreover, cloud-based systems that run on third-party

hardware over the internet, have also been devised (Halligan et al., 2009; Trudgian & Mirzaei, 2012; Muth et al., 2013; Slagel et al., 2015).

The increase in performance offered by these new database search setups has made it possible to conduct global proteome analyses and provide the first maps of the human proteome (Kim et al., 2014; Wilhelm et al., 2014). Such large scale investigations can in turn be combined with other omics results that together provide an unprecedented characterization of a biological system (Cabezas-Wallscheid et al., 2014; Robles et al., 2014; Hein et al., 2015). A promising multi-omics application is the growing field of proteogenomics, where genomics, transcriptomics, proteomics, and epigenomics are combined to provide a fine-grained analysis of the gene translational and transcriptional processes (Jaffe et al., 2004).

Proteomic search engines play a key role in these approaches (Menschaert & Fenyo, 2015), and the control of the search space size and prevalence of false positives, especially of a non-random nature, is vital for their success (Nesvizhskii, 2014). The availability of increasing amounts of data in ever-improving quality from public repositories is a great advantage when searching for low abundant compounds , and also enables big data mining of all the globally acquired data in order to achieve unprecedented insights into biological systems (Vaudel et al., 2015; Volders et al., 2015; Olexiouk et al., 2016).

Acknowledgements

K.V. is thankful to Ghent University and VIB. L.M. acknowledges the support of Ghent University (Multidisciplinary Research Partnership “Bioinformatics: from nucleotides to networks”), and the SBO grant “InSPECtor” (120025) of the Flemish agency for Innovation by Science and Technology (IWT). H.B. and H.R are supported by Bergen Forskningsstiftelse, and H.R. is further supported by Novo Nordisk Fonden, Diabetesforbundet and Western Norway Regional Health Authority. H,B. is also supported by the Research Council of Norway. F.B. is supported by the Kristian Gerhard Jebsen foundation.

References

Aasebo E, Vaudel M, Mjaavatten O, Gausdal G, Van der Burgh A, Gjertsen BT, Doskeland SO, Bruserud O, Berven FS, Selheim F. 2014. Performance of super-SILAC based quantitative proteomics for comparison of different acute myeloid leukemia (AML) cell lines. Proteomics 14:1971-1976.

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

Aebersold R, Burlingame AL, Bradshaw RA. 2013. Western blots versus selected reaction monitoring assays: time to turn the tables? Mol Cell Proteomics 12:2381-2382.

Allmer J. 2011. Algorithms for the de novo sequencing of peptides from tandem mass spectra. Expert Rev Proteomics 8:645-657.

Alves G, Ogurtsov AY, Yu YK. 2008. RAId_DbS: mass-spectrometry based peptide identification web server with knowledge integration. BMC Genomics 9:505.

Bafna V, Edwards N. 2001. SCOPE: a probabilistic model for scoring tandem mass spectra against a peptide database. Bioinformatics 17 Suppl 1:S13-21.

Bandeira N, Tsur D, Frank A, Pevzner PA. 2007. Protein identification by spectral networks analysis. Proc Natl Acad Sci U S A 104:6140-6145.

Barbarini N, Magni P. 2010. Accurate peak list extraction from proteomic mass spectra for identification and profiling studies. BMC Bioinformatics 11:518.

Barla A, Jurman G, Riccadonna S, Merler S, Chierici M, Furlanello C. 2008. Machine learning methods for predictive proteomics. Brief Bioinform 9:119-128.

Barsnes H, Martens L. 2013. Crowdsourcing in proteomics: public resources lead to better experiments. Amino Acids 44:1129-1137.

Beausoleil SA, Villen J, Gerber SA, Rush J, Gygi SP. 2006. A probability-based approach for high-throughput protein phosphorylation analysis and site localization. Nat Biotechnol 24:1285-1292.

Bern M, Phinney BS, Goldberg D. 2009. Reanalysis of Tyrannosaurus rex Mass Spectra. J Proteome Res 8:4328-4332.

Bern M, Kil YJ. 2011. Comment on "Unbiased statistical analysis for multi-stage proteomic search strategies". J Proteome Res 10:2123-2127.

Bern M, Kil YJ, Becker C. 2012. Byonic: advanced peptide and protein identification software. Curr Protoc Bioinformatics Chapter 13:Unit13 20.

Boekel J, Chilton JM, Cooke IR, Horvatovich PL, Jagtap PD, Kall L, Lehtio J, Lukasse P, Moerland PD, Griffin TJ. 2015. Multi-omic data analysis using Galaxy. Nat Biotech 33:137-139.

Bogdanow B, Zauber H, Selbach M. 2016. Systematic Errors in Peptide and Protein Identification and Quantification by Modified Peptides. Mol Cell Proteomics 15:2791-2801.

Borges D, Perez-Riverol Y, Nogueira FC, Domont GB, Noda J, da Veiga Leprevost F, Besada V, Franca FM, Barbosa VC, Sanchez A, Carvalho PC. 2013. Effectively addressing complex proteomic search spaces with peptide spectrum matching. Bioinformatics 29:1343-1344.

Cabezas-Wallscheid N, Klimmeck D, Hansson J, Lipka DB, Reyes A, Wang Q, Weichenhan D, Lier A, von Paleske L, Renders S, Wunsche P, Zeisberger P, Brocks D, Gu L, Herrmann C, Haas S, Essers MA, Brors B, Eils R, Huber W, Milsom MD, Plass C, Krijgsveld J, Trumpp A. 2014.

Identification of regulatory networks in HSCs and their immediate progeny via integrated proteome, transcriptome, and DNA methylome analysis. Cell Stem Cell 15:507-522.

Chambers MC, Maclean B, Burke R, Amodei D, Ruderman DL, Neumann S, Gatto L, Fischer B, Pratt B, Egertson J, Hoff K, Kessner D, Tasman N, Shulman N, Frewen B, Baker TA, Brusniak MY, Paulse C, Creasy D, Flashner L, Kani K, Moulding C, Seymour SL, Nuwaysir LM, Lefebvre B, Kuhlmann F, Roark J, Rainer P, Detlev S, Hemenway T, Huhmer A, Langridge J, Connolly B, Chadick T, Holly K, Eckels J, Deutsch EW, Moritz RL, Katz JE, Agus DB, MacCoss M, Tabb DL, Mallick P. 2012. A cross-platform toolkit for mass spectrometry and proteomics. Nat Biotechnol 30:918-920.

Chick JM, Kolippakkam D, Nusinow DP, Zhai B, Rad R, Huttlin EL, Gygi SP. 2015. A mass-tolerant database search identifies a large proportion of unassigned spectra in shotgun proteomics as modified peptides. Nat Biotechnol 33:743-749.

Chu F, Baker PR, Burlingame AL, Chalkley RJ. 2010. Finding chimeras: a bioinformatics strategy for identification of cross-linked peptides. Mol Cell Proteomics 9:25-31.

Colaert N, Degroeve S, Helsens K, Martens L. 2011. Analysis of the resolution limitations of peptide identification algorithms. J Proteome Res 10:5555-5561.

Colinge J, Masselot A, Giron M, Dessingy T, Magnin J. 2003. OLAV: towards high-throughput tandem mass spectrometry data identification. Proteomics 3:1454-1463.

Cox J, Mann M. 2008. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat Biotechnol 26:1367-1372.

Cox J, Neuhauser N, Michalski A, Scheltema RA, Olsen JV, Mann M. 2011. Andromeda: a peptide search engine integrated into the MaxQuant environment. J Proteome Res 10:1794-1805.

Craig R, Beavis RC. 2004. TANDEM: matching proteins with tandem mass spectra. Bioinformatics 20:1466-1467.

Craig R, Cortens JP, Beavis RC. 2004. Open source system for analyzing, validating, and storing protein identification data. J Proteome Res 3:1234-1242.

Craig R, Cortens JC, Fenyo D, Beavis RC. 2006. Using annotated peptide mass spectrum libraries for protein identification. J Proteome Res 5:1843-1849.

Creasy DM, Cottrell JS. 2002. Error tolerant searching of uninterpreted tandem mass spectrometry data. Proteomics 2:1426-1434.

Dasari S, Chambers MC, Slebos RJ, Zimmerman LJ, Ham AJ, Tabb DL. 2010. TagRecon: high-throughput mutation identification through sequence tagging. J Proteome Res 9:1716-1726.

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:D655-658.

Deutsch EW, Mendoza L, Shteynberg D, Farrah T, Lam H, Tasman N, Sun Z, Nilsson E, Pratt B, Prazen B, Eng JK, Martin DB, Nesvizhskii AI, Aebersold R. 2010. A guided tour of the Trans-Proteomic Pipeline. Proteomics 10:1150-1159.

Deutsch EW. 2012. File formats commonly used in mass spectrometry proteomics. Mol Cell Proteomics 11:1612-1621.

Diament BJ, Noble WS. 2011. Faster SEQUEST searching for peptide identification from tandem mass spectra. J Proteome Res 10:3871-3879.

Doerr A. 2015. DIA mass spectrometry. Nature Methods 12.

Dorfer V, Pichler P, Stranzl T, Stadlmann J, Taus T, Winkler S, Mechtler K. 2014. MS Amanda, a Universal Identification Algorithm Optimized for High Accuracy Tandem Mass Spectra. J Proteome Res.

Dove A. 1999. Proteomics: translating genomics into products? Nat Biotechnol 17:233-236.

Du P, Stolovitzky G, Horvatovich P, Bischoff R, Lim J, Suits F. 2008. A noise model for mass spectrometry based proteomics. Bioinformatics 24:1070-1077.

Elias JE, Gygi SP. 2007. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. Nat Methods 4:207-214.

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 5:976–989.

Eng JK, Fischer B, Grossmann J, Maccoss MJ. 2008. A fast SEQUEST cross correlation algorithm. J Proteome Res 7:4598-4602.

Eng JK, Jahan TA, Hoopmann MR. 2013. Comet: an open-source MS/MS sequence database search tool. Proteomics 13:22-24.

Everett LJ, Bierl C, Master SR. 2010. Unbiased statistical analysis for multi-stage proteomic search strategies. J Proteome Res 9:700-707.

Faherty BK, Gerber SA. 2010. MacroSEQUEST: efficient candidate-centric searching and high-resolution correlation analysis for large-scale proteomics data sets. Anal Chem 82:6821-6829.

Fenn JB, Mann M, Meng CK, Wong SF, Whitehouse CM. 1989. Electrospray ionization for mass spectrometry of large biomolecules. Science 246:64-71.

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

Fenyo D, Eriksson J, Beavis R. 2010. Mass spectrometric protein identification using the global proteome machine. Methods Mol Biol 673:189-202.

Field HI, Fenyo D, Beavis RC. 2002. RADARS, a bioinformatics solution that automates proteome mass spectral analysis, optimises protein identification, and archives data in a relational database. Proteomics 2:36-47.

Foster LJ. 2011. Interpretation of data underlying the link between colony collapse disorder (CCD) and an invertebrate iridescent virus. Mol Cell Proteomics 10:M110 006387.

Frank AM, Monroe ME, Shah AR, Carver JJ, Bandeira N, Moore RJ, Anderson GA, Smith RD, Pevzner PA. 2011. Spectral archives: extending spectral libraries to analyze both identified and unidentified spectra. Nat Methods 8:587-591.

French WR, Zimmerman LJ, Schilling B, Gibson BW, Miller CA, Townsend RR, Sherrod SD, Goodwin CR, McLean JA, Tabb DL. 2015. Wavelet-based peak detection and a new charge inference procedure for MS/MS implemented in ProteoWizard's msConvert. J Proteome Res 14:1299-1307.

Gaudet P, Michel PA, Zahn-Zabal M, Cusin I, Duek PD, Evalet O, Gateau A, Gleizes A, Pereira M, Teixeira D, Zhang Y, Lane L, Bairoch A. 2015. The neXtProt knowledgebase on human proteins: current status. Nucleic Acids Res 43:D764-770.

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 3:958-964.

Ghesquiere B, Helsens K, Vandekerckhove J, Gevaert K. 2011. A stringent approach to improve the quality of nitrotyrosine peptide identifications. Proteomics 11:1094-1098.

Giardine B, Riemer C, Hardison RC, Burhans R, Elnitski L, Shah P, Zhang Y, Blankenberg D, Albert I, Taylor J, Miller W, Kent WJ, Nekrutenko A. 2005. Galaxy: a platform for interactive large-scale genome analysis. Genome Res 15:1451-1455.

Gillet LC, Navarro P, Tate S, Rost H, Selevsek N, Reiter L, Bonner R, Aebersold R. 2012. Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis. Mol Cell Proteomics 11:O111 016717.

Gonnelli G, Stock M, Verwaeren J, Maddelein D, De Baets B, Martens L, Degroeve S. 2015. A decoy-free approach to the identification of peptides. J Proteome Res 14:1792-1798.

Halligan BD, Geiger JF, Vallejos AK, Greene AS, Twigger SN. 2009. Low cost, scalable proteomics data analysis using Amazon's cloud computing services and open source search algorithms. J Proteome Res 8:3148-3153.

Harris TW, Baran J, Bieri T, Cabunoc A, Chan J, Chen WJ, Davis P, Done J, Grove C, Howe K, Kishore R, Lee R, Li Y, Muller HM, Nakamura C, Ozersky P, Paulini M, Raciti D, Schindelman G, Tuli MA, Van Auken K, Wang D, Wang X, Williams G, Wong JD, Yook K, Schedl T, Hodgkin J, Berriman M, Kersey P, Spieth J, Stein L, Sternberg PW. 2014. WormBase 2014: new views of curated biology. Nucleic Acids Res 42:D789-793.

Hein MY, Hubner NC, Poser I, Cox J, Nagaraj N, Toyoda Y, Gak IA, Weisswange I, Mansfeld J, Buchholz F, Hyman AA, Mann M. 2015. A human interactome in three quantitative dimensions organized by stoichiometries and abundances. Cell 163:712-723.

Hong EL, Balakrishnan R, Dong Q, Christie KR, Park J, Binkley G, Costanzo MC, Dwight SS, Engel SR, Fisk DG, Hirschman JE, Hitz BC, Krieger CJ, Livstone MS, Miyasato SR, Nash RS, Oughtred R, Skrzypek MS, Weng S, Wong ED, Zhu KK, Dolinski K, Botstein D, Cherry JM. 2008. Gene Ontology annotations at SGD: new data sources and annotation methods. Nucleic Acids Res 36:D577-581.

Houel S, Abernathy R, Renganathan K, Meyer-Arendt K, Ahn NG, Old WM. 2010. Quantifying the impact of chimera MS/MS spectra on peptide identification in large-scale proteomics studies. J Proteome Res 9:4152-4160.

Howe DG, Bradford YM, Conlin T, Eagle AE, Fashena D, Frazer K, Knight J, Mani P, Martin R, Moxon SA, Paddock H, Pich C, Ramachandran S, Ruef BJ, Ruzicka L, Schaper K, Shao X, Singer A, Sprunger B, Van Slyke CE, Westerfield M. 2013. ZFIN, the Zebrafish Model Organism Database: increased support for mutants and transgenics. Nucleic Acids Res 41:D854-860.

Hsieh EJ, Hoopmann MR, MacLean B, MacCoss MJ. 2010. Comparison of database search strategies for high precursor mass accuracy MS/MS data. J Proteome Res 9:1138-1143.

Jaffe JD, Berg HC, Church GM. 2004. Proteogenomic mapping as a complementary method to perform genome annotation. Proteomics 4:59-77.

Jeong K, Kim S, Bandeira N. 2012. False discovery rates in spectral identification. BMC Bioinformatics 13 Suppl 16:S2.

Johnson R, Martin S, Biemann K, Stults J, Watson J. 1987. Novel fragmentation process of peptides by collision-induced decomposition in a tandem mass spectrometer: differentiation of leucine and isoleucine. Anal Chem 59:2621-2625.

Jones AR, Eisenacher M, Mayer G, Kohlbacher O, Siepen J, Hubbard SJ, Selley JN, Searle BC, Shofstahl J, Seymour SL, Julian R, Binz PA, Deutsch EW, Hermjakob H, Reisinger F, Griss J, Vizcaino JA,

Chambers M, Pizarro A, Creasy D. 2012. The mzIdentML data standard for mass spectrometry-based proteomics results. Mol Cell Proteomics 11:M111 014381.

Kalyanaraman A, Cannon WR, Latt B, Baxter DJ. 2011. MapReduce implementation of a hybrid spectral library-database search method for large-scale peptide identification. Bioinformatics 27:3072-3073.

Karas M, D. B, Hillenkamp F. 1985. Influence of the wavelength in high-irradiance ultraviolet laser desorption mass spectrometry of organic molecules. Anal. Chem. 57:2935–2939.

Kelchtermans P, Bittremieux W, De Grave K, Degroeve S, Ramon J, Laukens K, Valkenborg D, Barsnes H, Martens L. 2014. Machine learning applications in proteomics research: how the past can boost the future. Proteomics 14:353-366.

Kelleher NL. 2004. Top-down proteomics. Anal Chem 76:197A-203A.

Keller A, Nesvizhskii AI, Kolker E, 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.

Kertesz-Farkas A, Keich U, Noble WS. 2015. Tandem Mass Spectrum Identification via Cascaded Search. J Proteome Res 14:3027-3038.

Kim MS, Pinto SM, Getnet D, Nirujogi RS, Manda SS, Chaerkady R, Madugundu AK, Kelkar DS, Isserlin R, Jain S, Thomas JK, Muthusamy B, Leal-Rojas P, Kumar P, Sahasrabuddhe NA, Balakrishnan L, Advani J, George B, Renuse S, Selvan LD, Patil AH, Nanjappa V, Radhakrishnan A, Prasad S, Subbannayya T, Raju R, Kumar M, Sreenivasamurthy SK, Marimuthu A, Sathe GJ, Chavan S, Datta KK, Subbannayya Y, Sahu A, Yelamanchi SD, Jayaram S, Rajagopalan P, Sharma J, Murthy KR, Syed N, Goel R, Khan AA, Ahmad S, Dey G, Mudgal K, Chatterjee A, Huang TC, Zhong J, Wu X, Shaw PG, Freed D, Zahari MS, Mukherjee KK, Shankar S, Mahadevan A, Lam H, Mitchell CJ, Shankar SK, Satishchandra P, Schroeder JT, Sirdeshmukh R, Maitra A, Leach SD, Drake CG, Halushka MK, Prasad TS, Hruban RH, Kerr CL, Bader GD, Iacobuzio-Donahue CA, Gowda H, Pandey A. 2014. A draft map of the human proteome. Nature 509:575-581.

Kim S, Pevzner PA. 2014. MS-GF+ makes progress towards a universal database search tool for proteomics. Nat Commun 5:5277.

Klammer AA, Reynolds SM, Bilmes JA, MacCoss MJ, Noble WS. 2008. Modeling peptide fragmentation with dynamic Bayesian networks for peptide identification. Bioinformatics 24:i348-356.

Knudsen GM, Chalkley RJ. 2011. The effect of using an inappropriate protein database for proteomic data analysis. PLoS One 6:e20873.

Kuharev J, Navarro P, Distler U, Jahn O, Tenzer S. 2015. In-depth evaluation of software tools for data-independent acquisition based label-free quantification. Proteomics 15:3140-3151.

Lam H, Deutsch EW, Eddes JS, Eng JK, King N, Stein SE, Aebersold R. 2007. Development and validation of a spectral library searching method for peptide identification from MS/MS. Proteomics 7:655-667.

Lamesch P, Berardini TZ, Li D, Swarbreck D, Wilks C, Sasidharan R, Muller R, Dreher K, Alexander DL, Garcia-Hernandez M, Karthikeyan AS, Lee CH, Nelson WD, Ploetz L, Singh S, Wensel A, Huala E. 2012. The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. Nucleic Acids Res 40:D1202-1210.

Lange E, Gropl C, Reinert K, Kohlbacher O, Hildebrandt A. 2006. High-accuracy peak picking of proteomics data using wavelet techniques. Pac Symp Biocomput:243-254.

Li D, Fu Y, Sun R, Ling CX, Wei Y, Zhou H, Zeng R, Yang Q, He S, Gao W. 2005. pFind: a novel database-searching software system for automated peptide and protein identification via tandem mass spectrometry. Bioinformatics 21:3049-3050.

Li W, Ji L, Goya J, Tan G, Wysocki VH. 2011. SQID: an intensity-incorporated protein identification algorithm for tandem mass spectrometry. J Proteome Res 10:1593-1602.

Loroch S, Dickhut C, Zahedi RP, Sickmann A. 2013. Phosphoproteomics--more than meets the eye. Electrophoresis 34:1483-1492.

Ma ZQ, Dasari S, Chambers MC, Litton MD, Sobecki SM, Zimmerman LJ, Halvey PJ, Schilling B, Drake PM, Gibson BW, Tabb DL. 2009. IDPicker 2.0: Improved protein assembly with high discrimination peptide identification filtering. J Proteome Res 8:3872-3881.

Mann M, Wilm M. 1994. Error-tolerant identification of peptides in sequence databases by peptide sequence tags. Anal Chem 66:4390-4399.

Mann M, Hendrickson RC, Pandey A. 2001. Analysis of proteins and proteomes by mass spectrometry. Annu Rev Biochem 70:437-473.

Martens L, Hermjakob H, Jones P, Adamski M, Taylor C, States D, Gevaert K, Vandekerckhove J, Apweiler R. 2005. PRIDE: the proteomics identifications database. Proteomics 5:3537-3545.

Martens L, Nesvizhskii AI, Hermjakob H, Adamski M, Omenn GS, Vandekerckhove J, Gevaert K. 2005. Do we want our data raw? Including binary mass spectrometry data in public proteomics data repositories. Proteomics 5:3501-3505.

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:R110 000133.

Mashima J, Kodama Y, Kosuge T, Fujisawa T, Katayama T, Nagasaki H, Okuda Y, Kaminuma E, Ogasawara O, Okubo K, Nakamura Y, Takagi T. 2016. DNA data bank of Japan (DDBJ) progress report. Nucleic Acids Res 44:D51-57.

Matthiesen R, Lundsgaard M, Welinder KG, Bauw G. 2003. Interpreting peptide mass spectra by VEMS. Bioinformatics 19:792-793.

McHugh L, Arthur JW. 2008. Computational methods for protein identification from mass spectrometry data. PLoS Comput Biol 4:e12.

McQuilton P, St Pierre SE, Thurmond J, FlyBase C. 2012. FlyBase 101--the basics of navigating FlyBase. Nucleic Acids Res 40:D706-714.

Medzihradszky KF, Chalkley RJ. 2015. Lessons in de novo peptide sequencing by tandem mass spectrometry. Mass Spectrom Rev 34:43-63.

Menschaert G, Fenyo D. 2015. Proteogenomics from a bioinformatics angle: A growing field. Mass Spectrom Rev.

Michalski A, Cox J, Mann M. 2011. More than 100,000 detectable peptide species elute in single shotgun proteomics runs but the majority is inaccessible to data-dependent LC-MS/MS. J Proteome Res 10:1785-1793.

Millioni R, Tolin S, Puricelli L, Sbrignadello S, Fadini GP, Tessari P, Arrigoni G. 2011. High abundance proteins depletion vs low abundance proteins enrichment: comparison of methods to reduce the plasma proteome complexity. PLoS One 6:e19603.

Milloy JA, Faherty BK, Gerber SA. 2012. Tempest: GPU-CPU computing for high-throughput database spectral matching. J Proteome Res 11:3581-3591.

Muth T, Benndorf D, Reichl U, Rapp E, Martens L. 2013. Searching for a needle in a stack of needles: challenges in metaproteomics data analysis. Mol Biosyst 9:578-585.

Muth T, Peters J, Blackburn J, Rapp E, Martens L. 2013. ProteoCloud: a full-featured open source proteomics cloud computing pipeline. J Proteomics 88:104-108.

Muth T, Kolmeder CA, Salojarvi J, Keskitalo S, Varjosalo M, Verdam FJ, Rensen SS, Reichl U, de Vos WM, Rapp E, Martens L. 2015. Navigating through metaproteomics data: a logbook of database searching. Proteomics 15:3439-3453.

Na S, Bandeira N, Paek E. 2012. Fast multi-blind modification search through tandem mass spectrometry. Mol Cell Proteomics 11:M111 010199.

Nesvizhskii AI, Aebersold R. 2005. Interpretation of shotgun proteomic data: the protein inference problem. Mol Cell Proteomics 4:1419-1440.

Nesvizhskii AI. 2010. A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. J Proteomics 73:2092-2123.

Nesvizhskii AI. 2014. Proteogenomics: concepts, applications and computational strategies. Nat Methods 11:1114-1125.

Neuhauser N, Michalski A, Cox J, Mann M. 2012. Expert system for computer-assisted annotation of MS/MS spectra. Mol Cell Proteomics 11:1500-1509.

Nielsen ML, Savitski MM, Zubarev RA. 2006. Extent of modifications in human proteome samples and their effect on dynamic range of analysis in shotgun proteomics. Mol Cell Proteomics 5:2384-2391.

Ning K, Leong HW. 2007. Algorithm for peptide sequencing by tandem mass spectrometry based on better preprocessing and anti-symmetric computational model. Comput Syst Bioinformatics Conf 6:19-30.

Noble WS. 2015. Mass spectrometrists should search only for peptides they care about. Nat Methods 12:605-608.

Olexiouk V, Crappe J, Verbruggen S, Verhegen K, Martens L, Menschaert G. 2016. sORFs.org: a repository of small ORFs identified by ribosome profiling. Nucleic Acids Res 44:D324-329.

Olsen JV, Macek B, Lange O, Makarov A, Horning S, Mann M. 2007. Higher-energy C-trap dissociation for peptide modification analysis. Nat Methods 4:709-712.

Paizs B, Suhai S. 2005. Fragmentation pathways of protonated peptides. Mass Spectrom Rev 24:508-548.

Park CY, Klammer AA, Kall L, MacCoss MJ, Noble WS. 2008. Rapid and accurate peptide identification from tandem mass spectra. J Proteome Res 7:3022-3027.

Penel S, Arigon AM, Dufayard JF, Sertier AS, Daubin V, Duret L, Gouy M, Perriere G. 2009. Databases of homologous gene families for comparative genomics. BMC Bioinformatics 10 Suppl 6:S3.

Perez-Riverol Y, Alpi E, Wang R, Hermjakob H, Vizcaino JA. 2014. Making proteomics data accessible and reusable: Current state of proteomics databases and repositories. Proteomics.

Perkins DN, Pappin DJ, Creasy DM, Cottrell JS. 1999. Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis 20:3551-3567.

Pruitt KD, Tatusova T, Maglott DR. 2005. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res 33:D501-504.

Reddy TB, Riley R, Wymore F, Montgomery P, DeCaprio D, Engels R, Gellesch M, Hubble J, Jen D, Jin H, Koehrsen M, Larson L, Mao M, Nitzberg M, Sisk P, Stolte C, Weiner B, White J, Zachariah ZK, Sherlock G, Galagan JE, Ball CA, Schoolnik GK. 2009. TB database: an integrated platform for tuberculosis research. Nucleic Acids Res 37:D499-508.

Renard BY, Kirchner M, Monigatti F, Ivanov AR, Rappsilber J, Winter D, Steen JA, Hamprecht FA, Steen H. 2009. When less can yield more - Computational preprocessing of MS/MS spectra for peptide identification. Proteomics 9:4978-4984.

Robles MS, Cox J, Mann M. 2014. In-vivo quantitative proteomics reveals a key contribution of post-transcriptional mechanisms to the circadian regulation of liver metabolism. PLoS Genet 10:e1004047.

Roepstorff P, Fohlman J. 1984. Proposal for a common nomenclature for sequence ions in mass spectra of peptides. Biomed Mass Spectrom 11:601.

Sadygov RG, Yates JR, 3rd. 2003. A hypergeometric probability model for protein identification and validation using tandem mass spectral data and protein sequence databases. Anal Chem 75:3792-3798.

Sadygov RG, Cociorva D, Yates JR, 3rd. 2004. Large-scale database searching using tandem mass spectra: looking up the answer in the back of the book. Nat Methods 1:195-202.

Sadygov RG, Good DM, Swaney DL, Coon JJ. 2009. A new probabilistic database search algorithm for ETD spectra. J Proteome Res 8:3198-3205.

Schaab C, Geiger T, Stoehr G, Cox J, Mann M. 2012. Analysis of high accuracy, quantitative proteomics data in the MaxQB database. Mol Cell Proteomics 11:M111 014068.

Searle BC, Turner M, Nesvizhskii AI. 2008. Improving sensitivity by probabilistically combining results from multiple MS/MS search methodologies. J Proteome Res 7:245-253.

Seidler J, Zinn N, Boehm ME, Lehmann WD. 2010. De novo sequencing of peptides by MS/MS. Proteomics 10:634-649.

Sheng Q, Li R, Dai J, Li Q, Su Z, Guo Y, Li C, Shyr Y, Zeng R. 2015. Preprocessing significantly improves the peptide/protein identification sensitivity of high-resolution isobarically labeled tandem mass spectrometry data. Mol Cell Proteomics 14:405-417.

Shilov IV, Seymour SL, Patel AA, Loboda A, Tang WH, Keating SP, Hunter CL, Nuwaysir LM, Schaeffer DA. 2007. The Paragon Algorithm, a next generation search engine that uses sequence temperature values and feature probabilities to identify peptides from tandem mass spectra. Mol Cell Proteomics 6:1638-1655.

Shteynberg D, Nesvizhskii AI, Moritz RL, Deutsch EW. 2013. Combining results of multiple search engines in proteomics. Mol Cell Proteomics 12:2383-2393.

Skilling J, Denny R, Richardson K, Young P, McKenna T, Campuzano I, Ritchie M. 2004. ProbSeq--a fragmentation model for interpretation of electrospray tandem mass spectrometry data. Comp Funct Genomics 5:61-68.

Slagel J, Mendoza L, Shteynberg D, Deutsch EW, Moritz RL. 2015. Processing shotgun proteomics data on the Amazon cloud with the trans-proteomic pipeline. Mol Cell Proteomics 14:399-404.

Solari FA, Dell'Aica M, Sickmann A, Zahedi RP. 2015. Why phosphoproteomics is still a challenge. Mol Biosyst 11:1487-1493.

Soltwisch J, Souady J, Berkenkamp S, Dreisewerd K. 2009. Effect of gas pressure and gas type on the fragmentation of peptide and oligosaccharide ions generated in an elevated pressure UV/IR-

MALDI ion source coupled to an orthogonal time-of-flight mass spectrometer. Anal Chem 81:2921-2934.

Sturm M, Bertsch A, Gröpl C, Hildebrandt A, Hussong R, Lange E, Pfeifer N, Schulz-Trieglaff O, Zerck A, Reinert K, O. K. 2008. OpenMS - an open-source software framework for mass spectrometry. BMC Bioinformatics 9.

Syka JE, Coon JJ, Schroeder MJ, Shabanowitz J, Hunt DF. 2004. Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. Proc Natl Acad Sci U S A 101:9528-9533.

Tabb DL, Narasimhan C, Strader MB, Hettich RL. 2005. DBDigger: reorganized proteomic database identification that improves flexibility and speed. Anal Chem 77:2464-2474.

Tabb DL, Fernando CG, Chambers MC. 2007. MyriMatch: highly accurate tandem mass spectral peptide identification by multivariate hypergeometric analysis. J Proteome Res 6:654-661.

Tanaka K, Waki H, Ido Y, Akita S, Yoshida Y, Yoshida T, T. M. 1988. Protein and polymer analyses up to m/z 100 000 by laser ionization time-of-flight mass spectrometry. Rapid Commun Mass Spectrom 2:151-153.

Tanner S, Shu H, Frank A, Wang LC, Zandi E, Mumby M, Pevzner PA, Bafna V. 2005. InsPecT: identification of posttranslationally modified peptides from tandem mass spectra. Anal Chem 77:4626-4639.

Trudgian DC, Mirzaei H. 2012. Cloud CPFP: a shotgun proteomics data analysis pipeline using cloud and high performance computing. J Proteome Res 11:6282-6290.

UniProt C. 2015. UniProt: a hub for protein information. Nucleic Acids Res 43:D204-212.

Urban M, Pant R, Raghunath A, Irvine AG, Pedro H, Hammond-Kosack KE. 2015. The Pathogen-Host Interactions database (PHI-base): additions and future developments. Nucleic Acids Res 43:D645-655.

Vaudel M, Barsnes H, Berven FS, Sickmann A, Martens L. 2011. SearchGUI: An open-source graphical user interface for simultaneous OMSSA and X!Tandem searches. Proteomics 11:996-999.

Vaudel M, Burkhart JM, Sickmann A, Martens L, Zahedi RP. 2011. Peptide identification quality control. Proteomics 11:2105-2114.

Vaudel M, Sickmann A, Martens L. 2012. Current methods for global proteome identification. Expert Rev Proteomics 9:519-532.

Vaudel M, Sickmann A, Martens L. 2013. Introduction to opportunities and pitfalls in functional mass spectrometry based proteomics. Biochim Biophys Acta.

Vaudel M, Venne AS, Berven FS, Zahedi RP, Martens L, Barsnes H. 2014. Shedding light on black boxes in protein identification. Proteomics 14:1001-1005.

Vaudel M, Burkhart JM, Zahedi RP, Oveland E, Berven FS, Sickmann A, Martens L, Barsnes H. 2015. PeptideShaker enables reanalysis of MS-derived proteomics data sets. Nat Biotech 33:22-24.

Vaudel M, Verheggen K, Csordas A, Raeder H, Berven FS, Martens L, Vizcaino JA, Barsnes H. 2015. Exploring the potential of public proteomics data. Proteomics.

Verheggen K, Barsnes H, Martens L. 2014. Distributed computing and data storage in proteomics: many hands make light work, and a stronger memory. Proteomics 14:367-377.

Verheggen K, Maddelein D, Hulstaert N, Martens L, Barsnes H, Vaudel M. 2016. Pladipus Enables Universal Distributed Computing in Proteomics Bioinformatics. J Proteome Res 15:707-712.

Vizcaino JA, Csordas A, del-Toro N, Dianes JA, Griss J, Lavidas I, Mayer G, Perez-Riverol Y, Reisinger F, Ternent T, Xu QW, Wang R, Hermjakob H. 2016. 2016 update of the PRIDE database and its related tools. Nucleic Acids Res 44:D447-456.

Volders PJ, Verheggen K, Menschaert G, Vandepoele K, Martens L, Vandesompele J, Mestdagh P. 2015. An update on LNCipedia: a database for annotated human lncRNA sequences. Nucleic Acids Res 43:4363-4364.

Wang LH, Li DQ, Fu Y, Wang HP, Zhang JF, Yuan ZF, Sun RX, Zeng R, He SM, Gao W. 2007. pFind 2.0: a software package for peptide and protein identification via tandem mass spectrometry. Rapid Commun Mass Spectrom 21:2985-2991.

Wells JM, McLuckey SA. 2005. Collision-induced dissociation (CID) of peptides and proteins. Methods Enzymol 402:148-185.

Wenger CD, Coon JJ. 2013. A proteomics search algorithm specifically designed for high-resolution tandem mass spectra. J Proteome Res 12:1377-1386.

Wilhelm M, Schlegl J, Hahne H, Moghaddas Gholami A, Lieberenz M, Savitski MM, Ziegler E, Butzmann L, Gessulat S, Marx H, Mathieson T, Lemeer S, Schnatbaum K, Reimer U, Wenschuh H, Mollenhauer M, Slotta-Huspenina J, Boese JH, Bantscheff M, Gerstmair A, Faerber F, Kuster B. 2014. Mass-spectrometry-based draft of the human proteome. Nature 509:582-587.

Xu H, Freitas MA. 2009. MassMatrix: a database search program for rapid characterization of proteins and peptides from tandem mass spectrometry data. Proteomics 9:1548-1555.

Yadav AK, Kumar D, Dash D. 2011. MassWiz: a novel scoring algorithm with target-decoy based analysis pipeline for tandem mass spectrometry. J Proteome Res 10:2154-2160.

Yang P, Humphrey SJ, Fazakerley DJ, Prior MJ, Yang G, James DE, Yang JY. 2012. Re-fraction: a machine learning approach for deterministic identification of protein homologues and splice variants in large-scale MS-based proteomics. J Proteome Res 11:3035-3045.

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 78:1071-1084.

Yu W, Taylor JA, Davis MT, Bonilla LE, Lee KA, Auger PL, Farnsworth CC, Welcher AA, Patterson SD. 2010. Maximizing the sensitivity and reliability of peptide identification in large-scale proteomic experiments by harnessing multiple search engines. Proteomics 10:1172-1189.

Zhang J, Xin L, Shan B, Chen W, Xie M, Yuen D, Zhang W, Zhang Z, Lajoie GA, Ma B. 2012. PEAKS DB: de novo sequencing assisted database search for sensitive and accurate peptide identification. Mol Cell Proteomics 11:M111 010587.

Zhang N, Aebersold R, Schwikowski B. 2002. ProbID: a probabilistic algorithm to identify peptides through sequence database searching using tandem mass spectral data. Proteomics 2:1406-1412.

Zhang N, Li XJ, Ye M, Pan S, Schwikowski B, Aebersold R. 2005. ProbIDtree: an automated software program capable of identifying multiple peptides from a single collision-induced dissociation spectrum collected by a tandem mass spectrometer. Proteomics 5:4096-4106.

Figure legends

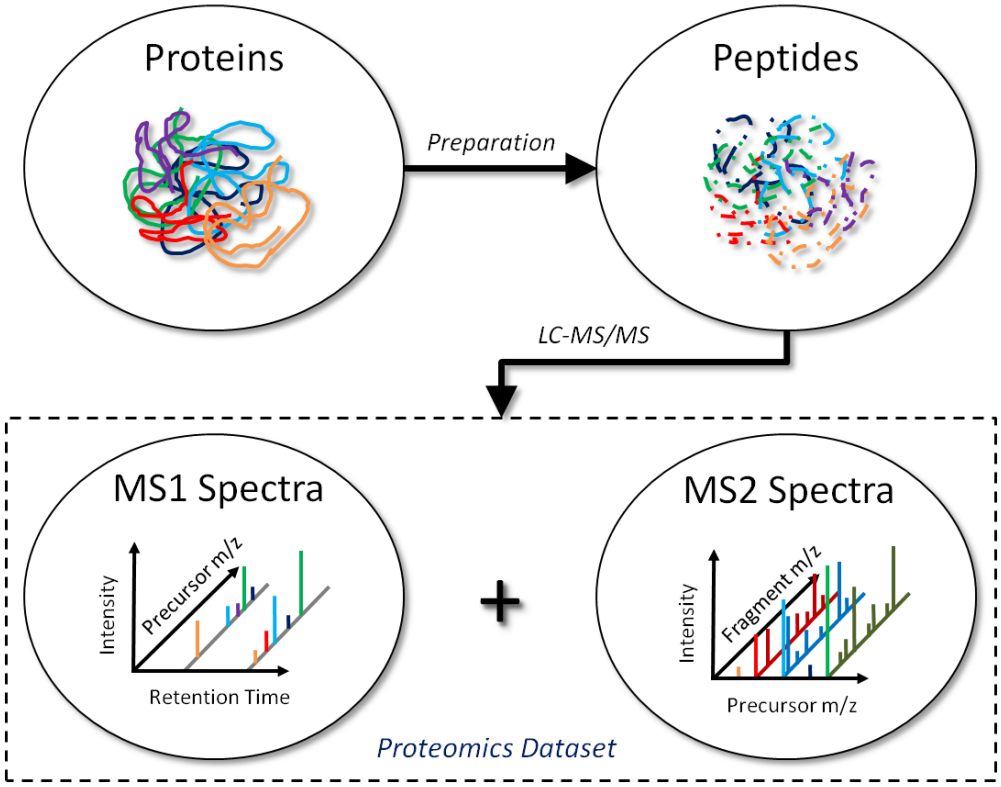

Figure 1: In a typical shotgun proteomics workflow, proteins are extracted from biological samples, their tertiary and secondary structures are reduced to a linear form, and undergo proteolytic digestion. The obtained peptide mixture is usually fractionated to reduce its complexity and analyzed by liquid chromatography coupled to tandem mass spectrometry (LC-MS/MS). The acquired data consist of two types of mass spectra: MS1 spectra that show the intensity *versus* m/z of the different ionized analytes at a given LC retention time, and MS2 spectra that show the intensity versus m/z of the fragmentation products of analytes called precursors that are isolated at a given retention time and mass range prior to dissociation, e.g. induced by collision with an inert gas.

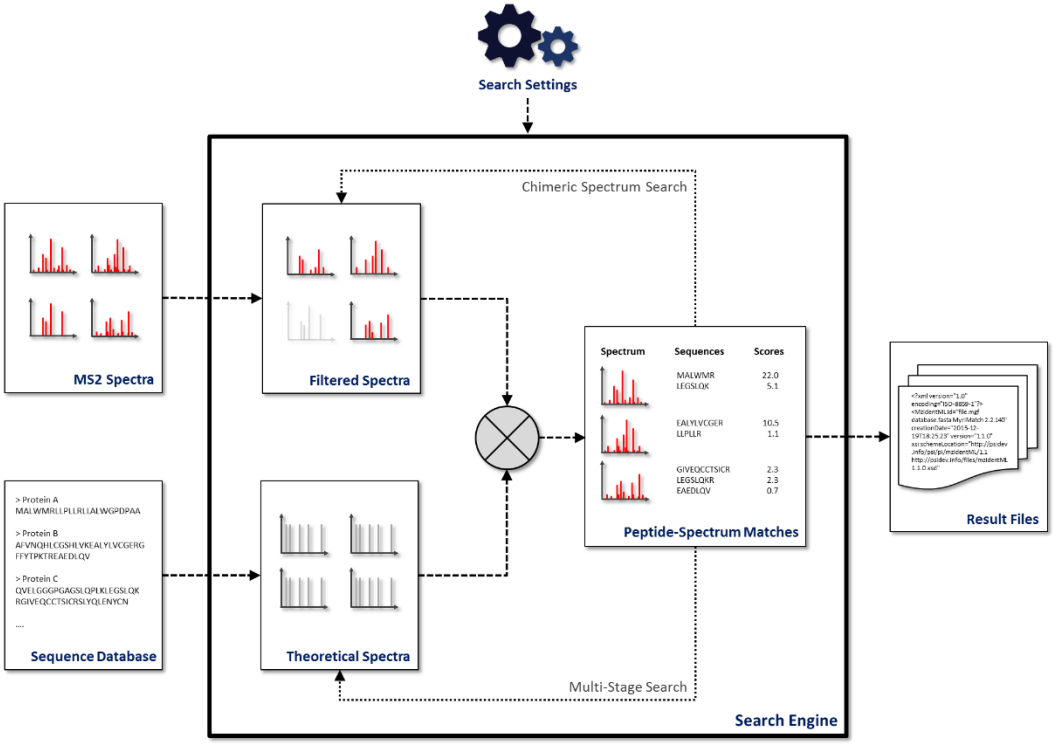

Figure 2: Database search engines attempt to match experimentally obtained MS2 spectra to peptides derived from sequence databases. Their function can be summarized in four steps: (1) spectra are filtered to reduce the number of peaks to process, (2) theoretical spectra are derived from the database sequence, (3) theoretical and experimental spectra are compared and their match is scored, and (4) peptide-spectrum matches (PSMs) are exported for post-processing. The different steps are controlled by search settings, meant to tune the search engine to the experimental conditions. Additionally, advanced search configurations allow the identification of chimeric spectra, and multi-stage strategies (see main text for details).

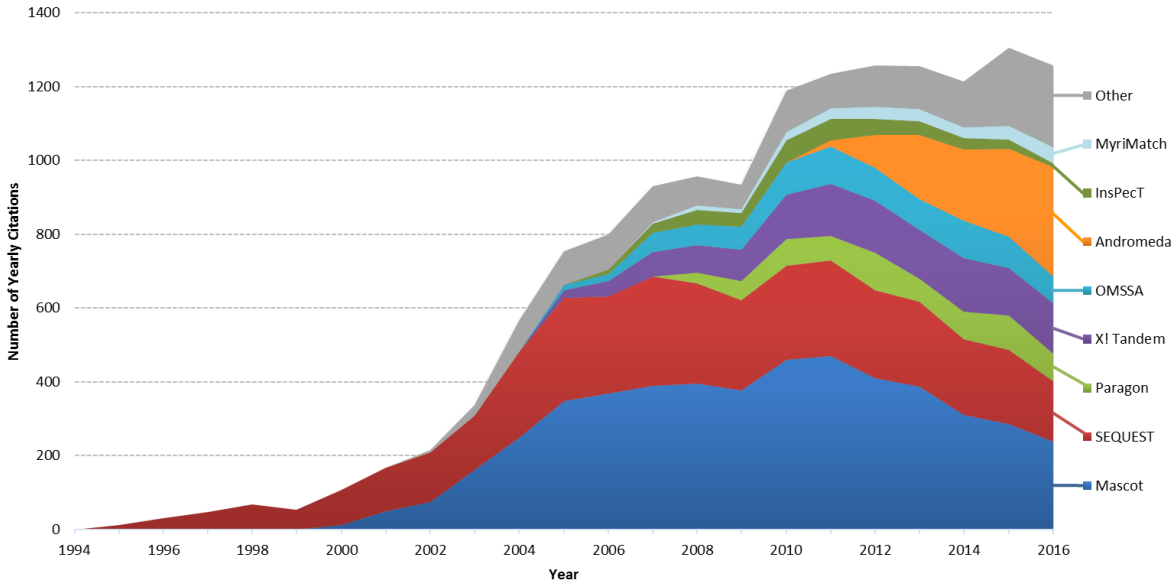

Figure 3: The yearly number of citations (from 1994 - 2016) for the search engines listed in Table 1 according to Thomson Reuters™ Web of Science™, counting the original publication only. The eight most cited search engines are listed individually, while the rest are grouped in the *Other* category. The number of citations can be used as an indicator of the prevalence of a given search engine in the literature, albeit with caution (see main text and the supplementary material for details).

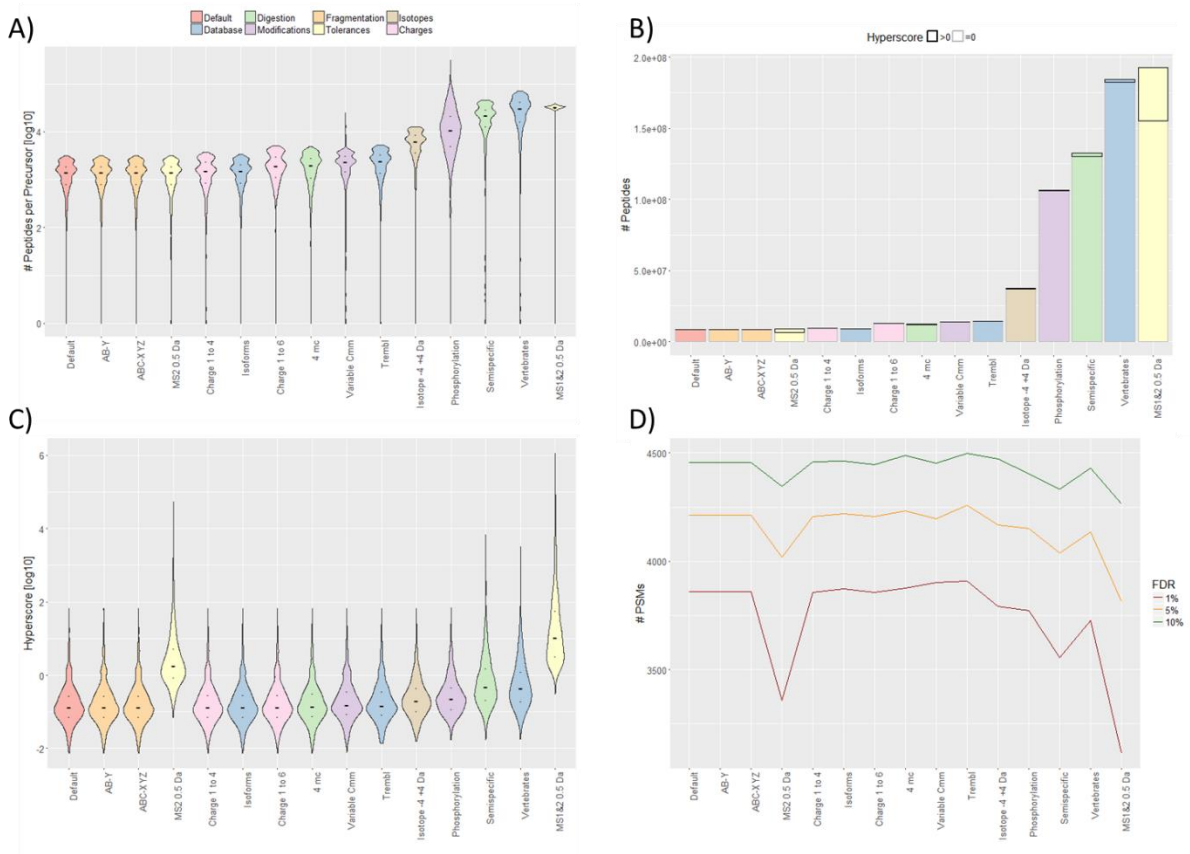

Figure 4: **A)** The density of the number of possible peptides per precursor is plotted as violin plot for every search space enlargement parameter (see main text) after logarithm base 10 transformation. In each case, a large dash represents the median and two smaller dashes represent the upper and lower quartiles. The densities are colored according to the type of parameter changed, and ordered by increasing median. **B)** The number of peptides considered during the search is plotted for every setting of Table 1 in the same order and coloring as in 4A. The peptides with a hyperscore > 0 are outlined in black. **C)** The density of the scores of decoy hits is plotted as in 4A using the same order. **D)** The number of target PSMs in every condition is plotted at 1%, 5%, and 10% False Discovery Rate (FDR) in green, orange, and red, respectively. The FDR is estimated using the share of decoy hits retained at a given score (Elias & Gygi, 2007).

Tables

Name	Year	Website	Publication	#Citations
SEQUEST	1994	fields.scripps.edu/sequest	(Eng et al., 1994)	3844
Mascot	1999	matrixscience.com	(Perkins et al., 1999)	4976
ProbID	2002	tools.proteomecenter.org/wiki/index.php?title=Software:ProbID	(Zhang et al., 2002)	159
Sonar	2002	-	(Field et al., 2002)	164
PEP_Probe	2003	bart.scripps.edu/public/search/pep_probe/search.jsp	(Sadygov & Yates, 2003)	147
OLAV	2003	-	(Colinge et al., 2003)	220
VEMS	2003	portugene.com/vems.html	(Matthiesen et al., 2003)	17
Phenyx	2004	genebio.com/products/phenyx/index.html	-	-
OMSSA	2004	ftp.ncbi.nlm.nih.gov/pub/lewisg/omssa	(Geer et al., 2004)	821
X! Tandem	2004	thegpm.org/TANDEM	(Craig & Beavis, 2004)	1228
ProbIDTree	2005	-	(Zhang et al., 2005)	53
DBDigger	2005	-	(Tabb et al., 2005)	43
pFind	2005	pfind.ict.ac.cn	(Li et al., 2005)	57
InSpect	2005	proteomics.ucsd.edu/Software/Inspect	(Tanner et al., 2005)	383
IdentityE	2007	-	-	-
pFind2.0	2007	pfind.ict.ac.cn	(Wang et al., 2007)	67
Paragon	2007	sciex.com/products/software/proteinpilot-software	(Shilov et al., 2007)	630
MyriMatch	2007	medschool.vanderbilt.edu/msrc-bioinformatics/myrimatch-source	(Tabb et al., 2007)	253
Crux	2008	cruxtoolkit.sourceforge.net	(Park et al., 2008)	73
RAId_Dbs	2008	ncbi.nlm.nih.gov/CBBResearch/qmbp/RAId_DbS/index.html	(Alves et al., 2008)	10
Zcore	2009	-	(Sadygov et al., 2009)	25
MassMatrix	2009	massmatrix.net	(Xu & Freitas, 2009)	95
MacroSequest	2010	proteomics.dartmouth.edu/k/software/macrosequest.kldk	(Faherty & Gerber, 2010)	16
MS-Tag and Batch-Tag	2010	prospector2.ucsf.edu/prospector	(Chu et al., 2010)	35
Tide	2011	noble.gs.washington.edu/proj/tide	(Diament & Noble, 2011)	34
Andromeda	2011	maxquant.org	(Cox et al., 2011)	1009
SpectrumMill	2011	proteomics.broadinstitute.org	-	-
MassWiz	2011	masswiz.igib.res.in	(Yadav et al., 2011)	19
SQID	2011	-	(Li et al., 2011)	27
PeaksDB	2011	bioinfor.com/peaks/features/peaksdb.html	(Zhang et al., 2012)	3
MSPolygraph	2011	compbio.eecs.wsu.edu	(Kalyanaraman et al., 2011)	9
Tempest	2012	-	(Milloy et al., 2012)	9
Byonic	2012	proteinmetrics.com	(Bern et al., 2012)	0

Morpheus	2013	sourceforge.net/projects/morpheus-ms	(Wenger & Coon, 2013)	33
Comet	2013	comet-ms.sourceforge.net	(Eng et al., 2013)	104
ProLuCID	2013	fields.scripps.edu/prolucid	-	-
MS-GF+	2014	proteomics.ucsd.edu/software-tools/ms-gf	(Kim & Pevzner, 2014)	71
MS Amanda	2014	ms.imp.ac.at/?goto=msamanda	(Dorfer et al., 2014)	40
Greylag	2015	greylag.org	-	-

Table 1: Search engines listed by year of publication or availability. When available, the search engine website is provided, along with the related *original* publication and its total number of citations according to Thomson Reuters™ Web of Science™ from 1994 to 2016.

Name	Description	Value
Database	A text file containing a list of amino acid sequences to search in the FASTA format.	The sequences included should best cover the sequences present in the sample, but not contain a large proportion of additional sequences. Contaminants must be included.
Modifications	Mass modifications to be applied to the amino acids in the database. These can be fixed or variable, target single amino acids, amino acid patters, and specific locations on peptide or protein sequences.	The modifications searched need to account for modifications introduced during sample processing, artefactual modifications, and biological modifications.
Digestion	The sequences in the database can be searched in their entirety, cleaved unspecifically, or cleaved using an enzyme.	The digestion setting needs to best represent the method used to obtained peptides, if any.
Enzyme	The cleavage rule of the enzyme, e.g. “After K or R when not followed by P”. Some search engines support multiple enzyme digestion.	If an enzyme was used to digest the proteins, the cleavage rule should best model what peptides can be expected from proteins.
Specificity	It is possible to search for fully specific peptides, with both termini abiding by the cleavage rule, as well as semi-specific peptides, where only one terminus abides by the cleavage rule. The semi-specificity can be two sided, or limited to the C- or N-terminus.	This setting provides a degree of freedom in case the digestion was not complete, or if the proteins are not expected to be in full length in the sample.
Missed Cleavages	A certain number of missed cleavages can be allowed to account for partial digestion of peptides or the inaccuracy of cleavage rules.	This setting is usually set to 2 for trypsin, but should be optimized to account for digestion efficiency.
Fragment Ions	The fragment ions to annotate in spectra need to be set to account for the fragmentation method used.	Generally, b and y ions are used for CID and HCD fragmentation, c and z ions are used for ETD fragmentation.
Precursor Tolerance	The tolerance used to match a theoretic peptide m/z to a measured precursor m/z. This tolerance can be absolute or relative.	The tolerance needs to be adapted to the resolution at which the MS1 spectra were measured. A relative tolerance in ppm is used for high resolution MS1 data.
Fragment Tolerance	The tolerance used to match a theoretic fragment m/z to a measured fragment ion m/z. As for the precursor, this tolerance can be absolute or relative.	The tolerance needs to be adapted to the resolution at which the MS2 spectra were measured. A relative tolerance in ppm is used for high resolution MS2 data.
Precursor Charge	The charge of the peptide to search for can be set.	The charge needs to be adapted to the charge targeted by the mass spectrometer for fragmentation, typically 1 for MALDI ionization, 2 to 4 for electrospray ionization.
Isotopes	The isotopes to account for relative to the monoisotopic peak.	Peptides with one 13C are usually included in the search to account for incorrect monoisotopic peak assignment by the mass spectrometer. This value needs to be optimized based on the monoisotopic peak selection settings.

Table 2: The standard parameters enountered in most search engines. A description of each parameter is provided along with guidance on how to set and optimize the value for a given search.