Review

A potential golden age to come – current tools, recent use cases, and future avenues for *de novo* sequencing in proteomics

Thilo Muth[1#*], Felix Hartkopf[1#], Marc Vaudel[2,3*], and Bernhard Y. Renard[1]

[1] Bioinformatics Unit (MF 1), Department for Methods Development and Research Infrastructure, Robert Koch Institute, Berlin, Germany.
[2] K.G. Jebsen Center for Diabetes Research, Department of Clinical Science, University of Bergen, Norway
[3] Center for Medical Genetics and Molecular Medicine, Haukeland University Hospital, Bergen, Norway

[#]Shared first-authorship
*Corresponding authors:
Dr. Thilo Muth, Bioinformatics Unit (MF 1), Department for Methods Development and Research Infrastructure, Robert Koch Institute, 13353 Berlin, Germany. E-mail: mutht@rki.de
Dr. Marc Vaudel, Haukeland universitetssykehus, Laboratoriebygget, Postboks 7804, 5020 Bergen, Norway. E-mail: marc.vaudel@uib.no

Abstract

In shotgun proteomics, peptide and protein identification is most commonly conducted using database search engines, the method of choice when reference protein sequences are available. Despite its widespread use the database-driven approach is limited, mainly because of its static search space. In contrast, *de novo* sequencing derives peptide sequence information in an unbiased manner, using only the fragment ion information from the tandem mass spectra. In recent years, with the improvements in MS instrumentation, various new methods have been proposed for *de novo* sequencing.

This review article proposes an overview of existing *de novo* sequencing algorithms and software tools ranging from peptide sequencing to sequence-to-protein mapping. We describe various use cases where *de novo* sequencing was successfully applied. Finally, we highlight limitations of current methods and discuss new directions for a wider acceptance of *de novo* sequencing in the community.

Keywords

De novo sequencing / Bioinformatics / Peptide identification / Software tools / Protein identification

1 Introduction

Nowadays, database searching is the most common approach to identify peptides and proteins in shotgun (bottom-up) proteomics workflows. With this computational method, experimentally acquired tandem mass (MS/MS) spectra are searched against a reference database that contains target protein sequences from the proteomes of interest [1-3]. This protein database is either tailored, containing mainly sample- or (at least) species-specific reference sequences, or more generic, covering a broad variety of potential candidates, as it is often the case for publicly available reference proteomes from UniProtKB [4] or NCBI RefSeq [5].

The great benefit of targeting a specific reference can also be regarded the most critical issue since the ability of database search engines to identify peptides and proteins strongly depends on the availability and quality of appropriate reference sequences. As a consequence, the power of database searching is limited when the proteome reference is unavailable or incomplete, which is the typical case for organisms that have not yet been sequenced (e.g. for samples from non-model systems [6, 7] or microbial communities [8, 9]). Reference-based algorithms also have problems when the proteome reference is unreliable, which often occurs for splice variants [10], single amino acid variations (SAAVs) [11], or proteins with post-translational modifications (PTMs) [12]. In particular, sequence variation presents a major challenge when analyzing clinical cancer [13, 14] or pathogenic samples [15]. In these cases, tailored protein sequence databases are designed to capture biological variation. However, this creates an enlargement of the search space that decreases the discrimination power of search engines and consequently reduces their ability to identify peptides. Consequently, there is a need for complementary approaches that overcome these limitations.

Two alternatives to the above-described method exist for peptide identification: (1) spectral library searching, which matches experimental MS/MS spectra against a collection of pre-recorded spectra using spectrum-to-spectrum comparison [16], and (2) *de novo* sequencing, which infers partial or complete peptide sequences from the spectra. Because of lower processing times [17] and potentially higher identification yields [18] in comparison with database searching, spectral libraries have become a promising alternative for peptide identification. The interested reader is referred to the review article of Griss [19], which provides an extensive overview with detailed descriptions on available algorithms and resources for spectral library searching in proteomics. However, these methods require a solid foundation of previously acquired and well-annotated MS/MS spectra to which experimental data can be compared. Thus, spectral library searching depends on available high-quality references for spectrum data as much as database-driven peptide identification on high-quality sequence information.

In contrast to database and spectral library searching, *de novo* sequencing works in a completely unbiased manner as it does not require any input based on prior knowledge on the sample, but solely uses information available in the experimental MS/MS spectrum to infer the peptide sequence. In general, *de novo* sequencing shows the best performance for high quality data. This is, when the peptide fragmentation is well reflected within the spectrum, with high mass accuracy and sufficient coverage of fragment ions. Therefore, the increase in resolution of modern MS instruments has opened the way to a potential 'golden age' of *de novo* sequencing.

This review provides a detailed overview of state-of-the-art methods and software packages for *de novo* sequencing. We also review sequence-to-protein mapping, which can be combined with the *de novo* technique. In addition, we put particular emphasis on practical use cases, highlighting examples from previous proteomic studies that illustrate the effective application of *de novo* sequencing. Finally, we critically discuss shortcomings of existing methods and speculate on directions of improvement that may yield better performing tools and a wider acceptance of *de novo* sequencing in the proteomics community.

2 Principle of *de novo* sequencing and overview of algorithms

The objective of *de novo* sequencing is to determine the amino acid sequence of a peptide and associated modifications from a given MS/MS spectrum, precursor mass, and charge. As shown in **Figure 1A**, an MS/MS spectrum is essentially a bar plot, in which each fragment ion (acquired from the peptide fragmentation process inside the mass spectrometer) produces a signal peak at a specific mass-to-charge ratio (m/z), indicating its relative abundance (intensity). The key principle of *de novo* sequencing is that mass differences between pairs of fragment ion peaks are compared with the masses of the 20 standard amino acids (with matching mass values for leucine and isoleucine). The amino acids can be modified, either *in vivo* or during sample preparation resulting in additional mass shifts that need to be accounted for. The modifications can target specific amino acids, peptide or protein termini, or specific amino acids at termini. When modifications occur on the vast majority of possible modification sites (typically for chemical modifications with high yield), the modifications are considered as fixed or static and always accounted for. When less prevalent, modifications are considered variable or dynamic, requiring the algorithms to consider possible mass shifts at all modification sites.

To infer sequences along with potential modifications from mass spectra, most modern *de novo* sequencing algorithms employ approaches based on graph theory that construct a so-called *spectrum graph* for each MS/MS spectrum as described originally by Bartels [20]. A spectrum graph consists of nodes and edges. MS/MS peaks are converted into nodes representing masses (i.e. m/z values) of partial peptides. **Figure 1B** shows a simulated spectrum with singly charged fragment ion

peaks for convenience only. Based on this example, b-ion and y-ion spectrum graphs are illustrated exemplarily in **Figures 1C** and **1D**, respectively. A full path in each graph is constructed iteratively by connecting the nodes: an edge is drawn when the mass difference between two peak nodes corresponds to the mass of an amino acid. The spectrum graph in **Figure 1C** shows that the b_2 (at m/z 185.13) and b_3 (at m/z 256.17) fragment ion nodes are connected since their mass difference corresponds to the mass of alanine (71.04 Dalton). From this example, it is clear that the longer the peptide, the more combinations the algorithms will have to take into account. This combinatorial explosion is further amplified when taking into account variable modifications. In addition, a spectrum graph is usually scored, for example, on the basis of m/z peak matching accuracy or peak intensity (not shown here). The best-scoring path through the graph (traversing from N- to the C-terminus) is then used to *de novo* determine a candidate peptide sequence from the spectrum. In the example, b-ion (**Figure 1C**) and y-ion (**Figure 1D**) spectrum graphs are shown independently for the sake of simplicity, however, the information from both graphs is usually combined by *de novo* sequencing algorithms to obtain the sequence.

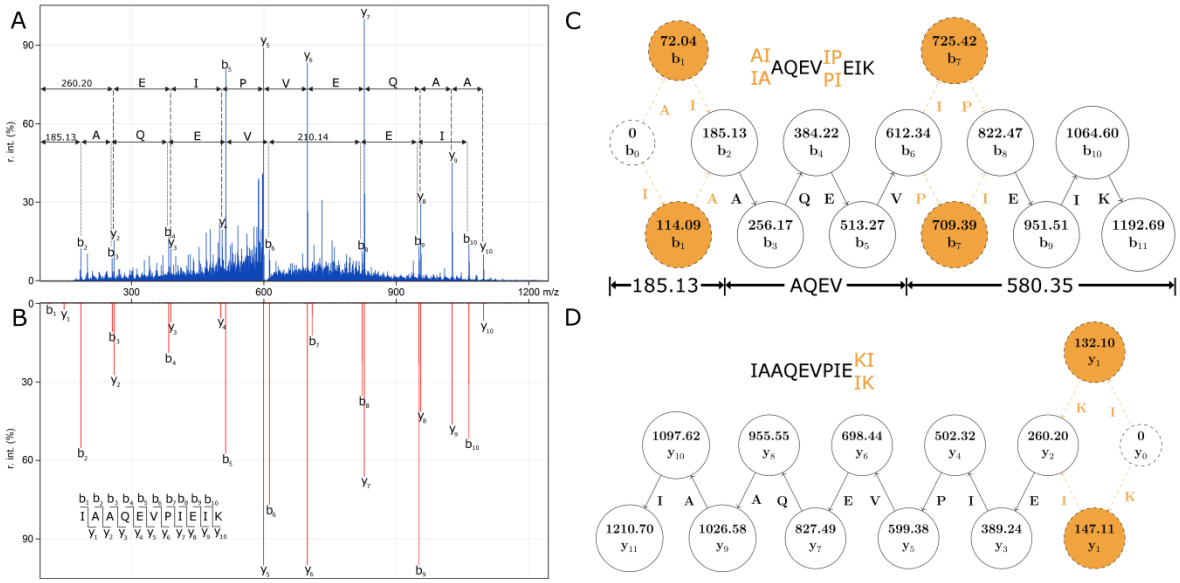

Figure 1. A) The full MS/MS spectrum for the peptide 'IAAQEVPIEIK' is shown in blue. The spectrum has been obtained from the ProteomeTools project [21] and visualized with the mMass software [22]. The singly charged fragment ion peaks for b- and y-ions used to determine the sequence are highlighted in black and the corresponding amino acids between the fragments are presented at the top. **B)** An idealized spectrum for the peptide 'IAAQEVPIEIK', as generated with MS2PIP [23], is shown in red. In the bottom left corner, the full sequence with annotated b- and y-ions is provided. **C)** b-ion spectrum graph for the MS/MS spectrum in (A). Each singly charged b-ion peak is shown as a node, and nodes are connected by edges labeled with the corresponding amino acid symbol. Dashed nodes and edges represent peaks and mass differences that are not found within the spectrum, but suggested only by *de novo* sequencing algorithms. More complex versions of such a spectrum graph, which include other charge states, ion classes and noise, are commonly utilized by these algorithms. The evaluation of all possible paths through the graph can yield different alternative sequences; here, two inversions are highlighted in orange. Note that y_0 and b_0 determine the origin of the graph (and do not exist as peaks in reality). Below the graph the peptide sequence tag 'AQEV' is shown,

which is flanked by unassigned masses of 185.13 Da and 580.35 Da. This reflects the typical output of sequence tagging algorithms. **D)** y-ion spectrum graph for the MS/MS spectrum in (A). This is analogous to the graph in (C), however, based on the singly charged y-ions. As shown in (B), the y-ions are suffixes of the fragmented peptide. Therefore, the calculation and evaluation of the graph is reversed.

Inferring a peptide sequence from an MS/MS spectrum *de novo* can be (and often has been) performed manually, by trained experts. This can still be useful, in particular with samples containing few peptides and when unusual PTMs or structures (e.g. cyclic peptides or disulfide bonds) occur. However, for complex mixtures, the massive amount of high-throughput data produced in current proteomic analysis workflows prohibits the manual approach. As a consequence, different algorithms for automatic *de novo* sequencing have been described, starting in the 1980s. One of the first computer-aided methods to tackle the *de novo* sequencing problem was the use of exhaustive search by Sakurai *et al.* [24]: in their pioneering approach, all potential amino acid sequences and corresponding theoretical spectra are generated for a given precursor mass. In a subsequent step, experimental spectra are matched against the generated theoretical spectra and a scoring function is applied based on the quality of the match. Finally, the best-scoring peptide sequences (based on a so-called "total reliability score") are taken as candidates for identification. Later on, various sub-sequencing approaches [25-27] were described for constraining the exponential growth of peptide sequences generated with increasing precursor mass. Since the 1990s, however, graph-based approaches [28, 29] have been increasingly utilized for solving the *de novo* sequencing problem. These methods are much more efficient as they circumvent the combinatorial explosion of evaluating all possible sequences. A host of further techniques have been employed for *de novo* sequencing to date, including integer linear programming [30, 31], dynamic programming [32-38], divide-and-conquer [39], hidden Markov models [40, 41], machine learning [41-44] and deep learning [45].

Beyond classical *de novo* sequencing algorithms that attempt to infer the complete peptide sequence (i.e. from N- to C-terminus) from the spectrum, sequence tagging methods [46-48] present another interesting algorithmic category: these algorithms derive so-called sequence tags, partial peptide sequences consisting of few amino acids surrounded by mass gaps. **Figure 1C** shows an example of the sequence tag 'AQEV' with the flanking, unassigned mass values of 185.13 Da and 580.35 Da that could be derived from the corresponding MS/MS spectrum using a sequence tagging algorithm. Tag-based *de novo* sequencing was introduced by Mann and Wilm in 1994 as a complementary approach to database searching [49]. The idea of the tag-based approach is that, given the heterogeneous ion coverage in the spectrum, a series of few high-intensity fragment ion peaks can be used to extract well-resolved sequence fragments that can in turn be matched against a reference database. Any peptide containing the sequence tag along with the correct flanking

masses is then considered an identification candidate. This step may also involve error-tolerant database searching: in this manner, peptides can be identified even when PTMs or SAAVs are present – accounting for variants that are not included in the reference database.

Numerous algorithms and software tools have been described in the past 20 years for tackling the *de novo* sequencing problem. In 1997, Lutefisk [29] was proposed as a pioneering software package for *de novo* sequencing. It was meant to be an addition to existing peptide identification tools such as SEQUEST [2] or PeptideSearch [49], e.g. for processing samples from organisms that are underrepresented in protein sequence databases. Lutefisk was the first tool to implement the spectrum graph strategy by Bartels [20]. In 1999, the SHERENGA algorithm [28] introduced a couple of paramount, previously undescribed strategies that considerably improved the performance of *de novo* sequencing. Those included automatic parameter learning, robust spectrum graph application for incomplete peptide fragmentation, and better scoring methods, e.g. for analyzing fragment ions of unknown charge states. The PEAKS software [34], described in 2003, uses a preprocessing step to generate candidate *de novo* sequences and employs dynamic programming in combination with a probabilistic scoring scheme for peptide prediction confidence. PEAKS provides complete peptide sequences in conjunction with confidence scores for individual amino acid assignments. In 2005, in their seminal work on the PepNovo algorithm, Frank *et al.* [35] described a scoring method based on a probabilistic network model reflecting the physical and chemical properties of peptide fragmentation. Besides a default set of models for collision-induced dissociation (CID) fragmentation, new models (e.g. for different fragmentation types) can be created in a separate training step. In the same year, NovoHMM [40] was proposed as the first algorithm using a generative hidden Markov model for solving the *de novo* sequencing problem. This has the benefit of providing an exact estimation of Bayesian posterior probabilities for amino acids rather than arbitrary score values. In 2008, Tabb *et al.* released DirecTag [46], a fast tool to infer sequence tags from MS/MS spectra. Remarkably, it includes three different scoring mechanisms for evaluating the tags based on peak intensity, m/z accuracy, and ion complementarity.

In 2010, with the advent of higher-energy collisional dissociation (HCD), pNovo [36] was proposed by Chi *et al.* and updated as pNovo+ [37] in 2013. The latter is able to process HCD and electron-transfer dissociation (ETD) spectra jointly, with the aim of increasing sequencing accuracy and coverage. According to the authors, the numbers of attained correct full-sequence peptides using *de novo* sequencing in combination with pNovo+ were already comparable to database-driven identification at this point. While such conclusions should generally be drawn with respect to selected test data sets, they have unquestionably given promise for a new era of reliable high-throughput *de novo* sequencing in proteomics. One of the most remarkable advances has been

demonstrated with the recent development of the Novor algorithm, yielding a substantial increase in both processing speed (more than 300 MS/MS spectra per second on a normal laptop computer) and sequencing accuracy when compared with previously published approaches [42]. Open-pNovo [50] overcomes the issue of combinatorial explosion introduced by the consideration of multiple modifications by combining efficient *de novo* sequencing with large precursor tolerance for detecting peptides with arbitrary types of modifications. Finally, state-of-the-art deep learning techniques have shown first, promising results: DeepNovo [45], a recently described algorithm based on a deep neural network and local dynamic programming, shows a significant improvement of up to 64% higher accuracy at the full-peptide level in comparison with its competitors PEAKS, PepNovo, and Novor. It combines the recent improvements in convolutional and recurrent neural networks to train on features from MS/MS spectra, fragment ion information, and peptide sequence patterns. Novel solutions with more specific use cases are being developed, such as Supernovo, a specific solution for *de novo* sequencing of monoclonal antibodies (mAbs) (proteinmetrics.com).

Available algorithms for *de novo* sequencing and sequence tagging are summarized in **Table 1** together with the paradigms, corresponding references, license types, and websites. It should be noted that most of these tools are also listed (among many other software packages for omics analyses) on OMICStools [51] (omictools.com).

Table 1. Currently available tools for MS-based *de novo* peptide sequencing. The list is sorted by publication year.

Algorithm	Paradigm(s)	Reference (author/year)	License	Project website
Lutefisk	spectrum graph	Taylor and Johnson 1997 [29]	free	hairyfatguy.com/lutefisk
SeqMS	spectrum graph	Fernandez-de-Cossio *et al.* 2000 [52]	free	protein.osaka-u.ac.jp/rcsfp/profiling/Seqms/Se
PEAKS	spectrum graph, dynamic programming	Ma 2003 [34]	commercial	bioinfor.com/peaks-studio
PepNovo	spectrum graph, dynamic programming	Frank *et al.* 2005 *[35]*	free	proteomics.ucsd.edu/Software/PepNovo
NovoHMM	hidden Markov model	Fischer *et al.* 2005 [40]	free	www-huber.embl.de/users/befische/software/
DirecTag	spectrum graph, tag generation	Tabb *et al.* 2008 [46]	free	medschool.vanderbilt.edu/msrc-bioinformatics
pNovo+	spectrum graph, dynamic programming	Chi *et al.* 2010 [36, 37]	free	pfind.ict.ac.cn/software/pNovo
ANTILOPE	integer linear programming	Andreotti *et al.* 2012 *[30]*	free	openms.de
UniNovo	spectrum graph, dynamic programming	Jeong *et al.* 2013 [38]	free	proteomics.ucsd.edu/software-tools/uninovo
Novor	spectrum graph, machine learning	Ma 2015 [42]	free	rapidnovor.com/download
LADS	machine learning	Devabhaktuni and Elias 2016 [43]	free	github.com/adevabhaktuni/LADS

Twister	top-down sequencing, tag generation	Vyatkina *et al.* 2016 [53]	free	bioinf.spbau.ru/en/twister
UVNovo	hidden Markov model, machine learning	Robotham *et al.* 2016 [41]	free	github.com/marcottelab/UVnovo
Open-pNovo	spectrum graph, dynamic programming	Yang et al. 2017 [50]	free	pfind.ict.ac.cn/software/pNovo
MRUniNovo	spectrum graph, dynamic programming	Li *et al.* 2017 [54]	free	bioinfo.hupo.org.cn/MRUniNovo
DeepNovo	dynamic programming, deep learning	Tran *et al.* 2017 [45]	free	github.com/nh2tran/DeepNovo
pSite	machine learning	Yang *et al.* 2017 [44]	free	pfind.ict.ac.cn/software/pSite

While many algorithms have been published, we observed in a previous benchmarking study [55] that only few methods are (i) regularly updated to support data from modern MS instruments and different fragmentation modes (e.g. CID and HCD) and (ii) currently available as generic software packages or integrated in workflows. Among those, the most widely-used are PepNovo [35], DirecTag [46], pNovo+ [37], and Novor [42]. To facilitate their use and integration, DeNovoGUI [56] provides a command line and graphical user interface for these tools. With a particular focus on usability, it enables researchers to inspect the *de novo* sequencing results in tabular form and also provides an interactive viewer application that annotates fragment ion spectra with amino acid predictions. The PEAKS tool suite [34] (bioinfor.com) is another powerful yet commercial software package for proteome analysis and includes a dedicated module for *de novo* sequencing. The relative performance of the different software solutions is a recurring matter of debate in the literature [55].

The output of most *de novo* sequencing tools is a list of full or partial candidate peptide sequences along with modifications and a score indicating the sequencing quality. The next step therefore involves mapping the candidate peptides to protein sequences. This is generally achieved using secondary tools, of which we provide an overview in the next section.

3 Survey of tools for mapping peptides and sequence tags to the protein level

As demonstrated in the previous section, the output of *de novo* sequencing tools is usually a list of candidate sequences, potentially including modifications and, frequently, mass gaps or ambiguous combinations of amino acids. Similarly, sequence tagging algorithms commonly output sequence tags of three to six amino acids. This makes it challenging for scientists to interpret the data in a meaningful way. To alleviate this issue, it is possible to aggregate the information at the protein level. This may be achieved by mapping the *de novo* peptides or tags to reference proteomes, for example from public databases such as UniProtKB [4] or NCBI RefSeq [5], or from custom and more tailored genome/transcriptome databases. **Figure 2** depicts a typical MS/MS-based protein identification workflow that includes both *de novo* sequencing and peptide-to-protein mapping.

Beyond *de novo* sequencing, the mapping of peptide sequences to reference proteomes is a crucial step in many proteomics workflows. Importantly, this does not have to happen only once: even when proteins have already been identified, the underlying 'raw' peptide sequences may be reconsidered by mapping them against a different reference proteome, which may contain novel sequences or sequence isoforms (e.g., when protein databases have been updated). It is also worth noting that algorithms for database searching and protein inference may introduce certain biases that one wants to avoid by remapping the peptides against the search database with the original references. The hybrid of *de novo* sequencing and protein mapping combines sequence tagging with database searches. The aim is to reduce overall processing time by filtering for potential candidate sequences using short sequence tags in the first step and performing a classical database matching on these candidates in the second step.

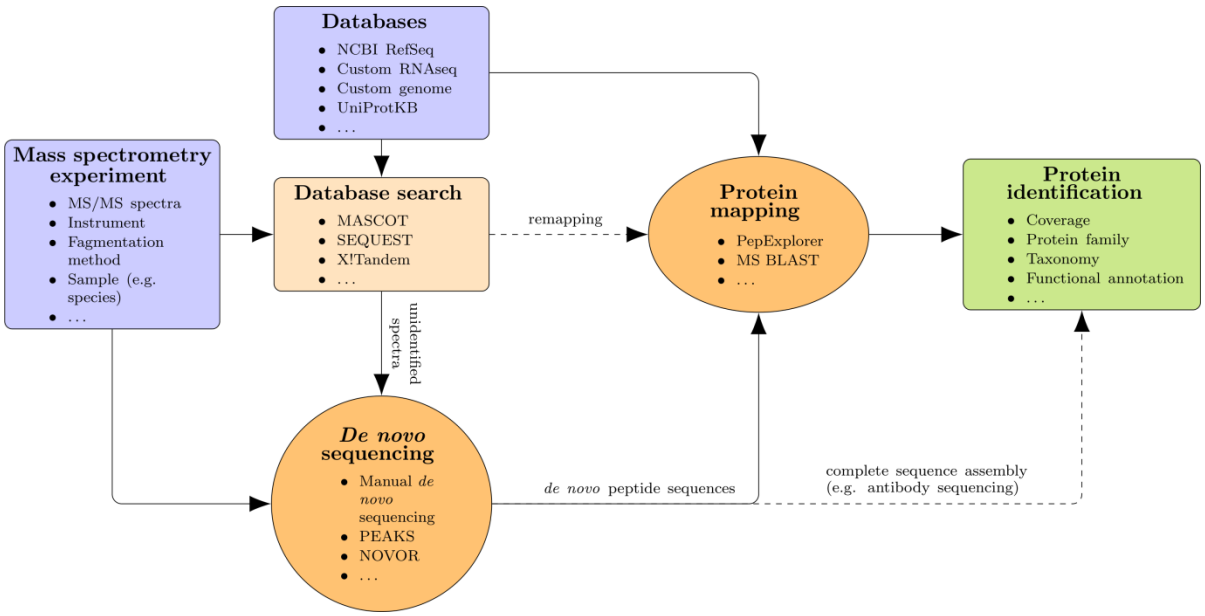

Figure 2. A typical workflow involving *de novo* sequencing. MS/MS spectra are acquired by mass spectrometry, which, together with the databases themselves, serve as input (blue) for database searches. Any unidentified spectra (or even all, e.g. for non-model organisms) can be processed with *de novo* sequencing (orange). Resulting *de novo* peptide sequences are assembled to complete sequences (e.g., for antibody sequencing) or mapped to a protein database. The resulting protein identifications (green) can be aggregated to provide valuable information about a sample, such as protein coverage, protein families, taxonomic and functional distribution.

Existing tools for mapping *de novo* sequencing data (full-length peptide sequences or peptide tags) to reference proteomes are summarized in **Table 2**. The popular BLAST [57] is widely used to compare sequences either at the genome or at the proteome level. However, it is clearly not the best choice for MS-based data, since the information from the spectra (e.g. precursor masses or fragment ion peaks) is not taken into consideration. For example, typical *de novo* sequencing errors such as inversions or mass ambiguities often occur due to missing fragment ion peaks within an MS/MS spectrum. This information is not considered when using BLAST, and erroneous query sequences may therefore lead to false assignments. BLAST has indeed been developed for whole-gene and -protein sequence comparison and is therefore not meant to handle short *de novo* peptide sequences. In 2001, Shevchenko *et al.* introduced the web-based MS BLAST [58] as a tailored solution for processing candidate peptides from *de novo* sequencing. Similar to the original BLAST heuristics, this tool allows users to match sequences in FASTA format against a predefined set of protein databases using different similarity matrices (e.g. PAM or BLOSUM). MS BLAST was specifically designed to align short sequences such as peptides and also makes it possible to include LC-MS/MS-specific parameters. However, it does not consider spectrum information nor does it permit to choose a user-defined reference database. Consequently, with these points in mind, further software packages have been developed. For example, MS-Homology [59] from the

ProteinProspector package allows for sequence tag searches using a user-defined protein database. Similarly, FASTS [60] uses multiple short peptide sequences for identifying proteins from any given protein database. Furthermore, the sequence tags output by DirecTag [46] can be used as input for TagRecon [61] to identify protein sequences containing unanticipated mutations. The TagRecon authors also propose general guidelines for validating such identifications. Corresponding acceptance criteria for PTM validation are found in an earlier study [62]. The basic principle shared by these guidelines is to discard unexpected (i.e. modified or mutated) peptide hits that fail to fulfill certain quality checks.

Further tools have been proposed with the specific aim of detecting mutations or modifications in proteomics data sets. The InsPecT package [63] allows the user to search for unexpected PTMs without explicit parameterization of all variants, via an internal peptide tag generation procedure followed by protein database filtering. The commercial PEAKS software, which has been updated with a specific module for database matching with PEAKS DB [64], also includes SPIDER [65], an algorithm to perform mutation-tolerant protein identification using sequence tags. In a similar fashion, the BICEPS algorithm [66] combines *de novo* sequencing and sequence tag searching, with the latter allowing for mutations in the *de novo* peptides and also at the protein sequence level.

Additional tools focus on data integration and visualization of results. PepExplorer [67] helps interpreting results from various modern *de novo* sequencing algorithms. The similarity-driven tool is that protein inference and false discovery rate (FDR) estimation are established by mapping the peptide sequences to a user-defined target-decoy sequence database. In addition, it provides a user-friendly graphical interface and outputs reports based on the identified proteins. Consequently, the output of this *de novo* sequencing-based workflow can be readily interpreted, similar as the results from established database-driven protein identification search engines. A guided tutorial for using PepExplorer in proteomic studies can be found in [68]. The web-based Unipept [69] focuses on metaproteomics research. Users can enter peptide sequences derived from an MS/MS experiment that are then matched against an indexed peptide database derived from UniProtKB. Unipept provides useful features with respect to taxonomy-based analysis. For example, it displays an interactive circular tree map based on the inferred proteins from different species to give an insight into the biodiversity of a sample. Still, Unipept may not be the best option for processing (particularly low-quality) *de novo* sequencing data, as it performs exact string matching only and therefore does not account for potentially incorrect or incomplete *de novo* sequences. The recently published PeptideMapper [70] was specifically developed for the rapid mapping of full-length or tag peptides to a user-defined protein reference database (in FASTA format). PeptideMapper utilizes a highly efficient full-text substring index data structure [71] based on the Burrows-Wheeler transform

that has already been used successfully for mapping next-generation sequencing reads [72]. This mapping algorithm has already been integrated into the user-friendly DeNovoGUI [56] and PeptideShaker [73] frameworks for post-processing results from *de novo* sequencing and database searching, respectively. While the previously mentioned tools are used to directly map *de novo* peptides to proteome references, Meta-SPS [74] follows a completely different approach: it seeks to obtain complete proteins *de novo* (e.g., with the goal to perform sequencing of whole antibodies, as explained further in the next section). To achieve this difficult goal, proteins need to be digested with multiple enzymes and peptides fragmented using different techniques such as HCD and ETD.

Table 2. Currently available tools for mapping full-length and tag peptides from MS-based *de novo* sequencing to reference proteomes. The list is sorted by publication year.

Algorithm	Reference (author/year)	License	Project website
MS BLAST	Shevchenko *et al.* 2001 [58]	free	genetics.bwh.harvard.edu/msblast
MS-Homology	Huang et al. 2001 [59]	free	prospector.ucsf.edu
FASTS	Mackey *et al.* 2002 [60]	free	fasta.bioch.virginia.edu
InsPect	Tanner *et al.* 2005 [63]	free	proteomics.ucsd.edu/Software/Inspect
SPIDER	Han *et al.* 2005 [65]	commercial	bioinfor.com/peptide-mutations-homology-searching
TagRecon	Dasari *et al.* 2010 [61]	free	medschool.vanderbilt.edu/msrc-bioinformatics/software
PEAKS DB	Zhang *et al.* 2012 [64]	commercial	bioinfor.com/peaksdb
BICEPS	Renard *et al.* 2012 [66]	free	software.steenlab.org
Unipept	Mesuere *et al.* 2012 [69]	free	unipept.ugent.be
Meta-SPS	Guthals *et al.* 2012 [74, 75]	free	proteomics.ucsd.edu/software-tools/metasps
PepExplorer	Leprevost *et al.* 2014 [67]	free	proteomics.fiocruz.br/software/pepexplorer
DeNovoGUI	Muth *et al.* 2014 [56]	free	compomics.github.io/projects/denovogui.html
PeptideMapper	Kopczynski *et al.* 2017 [70]	free	github.com/compomics/compomics-utilities/wiki/PeptideMapper

4 Practical applications of *de novo* sequencing in proteomics

Previous review articles [76-79] have already highlighted how *de novo* sequencing has been effectively used in the past for various use cases. Since then, however, improved MS instrumentation has made it possible to retrieve more data and achieve more complete peptide fragmentation. These improvements are reflected by high-intensity ion signals in the MS/MS scan. For example, in HCD data, prominent b- and y-ion series with high coverage are typically available. Consequently, more accurate methods and novel developments in the recent past enabled researchers to apply *de novo* sequencing for their purposes. At the same time, the lack of suitable reference proteomes has rendered the *de novo* technique the only viable option in many recent studies. This section focuses on applications of *de novo* sequencing in the field with a particular focus on recent studies and developments.

4.1 Antibody sequencing

Antibodies, also commonly known as immunoglobulins, are Y-shaped proteins produced mainly by plasma cells that are used by the immune system to cope with pathogens (e.g. bacteria or viruses) or cancer cells. In the field of immunotherapy, monoclonal antibodies (mAbs) are heavily used as molecules engineered to interact with the immune system and redirect immune responses. Overall, mAbs are very promising drug candidates, as they are more specific and have fewer side effects than conventional small-molecule drugs. In the pharmaceutical industry, determining the amino acid sequence of a monoclonal antibody is an essential step when discovering new drug candidates and innovator products for biosimilar development [80]. The challenge with this is that antibody sequences are (and need to be) highly diverse, as a result of gene recombination and somatic hypermutation events. In the pharmaceutical context, this is added to by further events occurring during the manufacturing process or storage [81]. On the proteome level, the diversity of antibodies manifests itself in sequence mutations, PTMs (e.g. varying glycosylation patterns), as well as terminal amino acid additions [82]. For example, it is known that glycosylation has strong effects on both antigen specificity and function of immunoglobulins [83]. Detailed knowledge about the sequence of a given antibody is critical for understanding the relationship between its structure and function, as well as for evaluating its efficacy and safety when used as a drug. Commonly, to obtain the sequence of an antibody with unknown variable regions, cDNA from the source hybridoma cell line is produced and sequenced. However, the previously mentioned post-DNA level modifications remain invisible in this manner, and sometimes hybridoma cells may be entirely unavailable. For these reasons, appropriate reference templates for antibodies are lacking, which renders the application of database-driven identification methods practically impossible. Since classical Edman degradation, as one possible alternative, is very time-consuming and provides low throughput only, efforts have

been made to directly sequence antibodies using the MS-based *de novo* approach. In 2006, Pham and colleagues at Genentech pioneered in the application of *de novo* sequencing for characterizing a full-length mAb by combining complementary digestion methods, MS-based analysis and Edman degradation [84]. Two years later, Bandeira *et al.* were the first to describe a dedicated workflow making it possible to sequence an antibody within 72 hours [85], much faster than Edman degradation. In their approach, so-called spectral contigs, as an equivalent to sequence contigs in genome sequencing, are first obtained in *de novo* manner. Reference antibody sequences are then used for ordering the contigs and, finally, for sequence mapping. The 95-99% coverage reported in this study translates into a performance superior to the classical Edman technique.

In recent years, different studies have described direct antibody sequencing based on the *de novo* technique with or without assisting databases. In 2016, Tran *et al.* proposed an integrated system for assembling antibody sequences [86] that combines *de novo* sequencing peptides (derived from PEAKS DB [64]), quality scores, and information from databases, using a weighted De Bruijn graph. They report unprecedented performance, with 100% coverage and 96-100% accuracy for three complete monoclonal antibody sequences. In a different study, the same authors also applied their DeepNovo algorithm for sequencing the heavy and light chains of a mouse antibody and reported coverage and accuracy values in the same range [45]. Further, Bogdanoff *et al.* combined mass spectrometry and crystallography to determine the protein sequence, structure and glycosylation pattern of the Fab fragment of a human astrovirus-neutralizing mAb [87]. In their study, they used the commercial Byonic software [88] and employed the so-called wild card search to determine the missing sequence gaps and unanticipated modifications. In 2017, Guthals and colleagues extended the application of *de novo* sequencing to characterize polyclonal antibodies directly from donor blood plasma [89], without genome-sequencing any peripheral B cells from the same donor. Savidor *et al.* have recently proposed a database-independent workflow for full-length protein and antibody sequencing [90] in which non-enzymatic microwave-assisted acid hydrolysis is used for semi-random cleavage, followed by solid-phase extraction, peptide *de novo* sequencing and contig assembly. Two commercial software developments for antibody sequencing, namely PEAKS AB (bioinfor.com/peaks-ab-software) and Supernovo (proteinmetrics.com/products/supernovo), underline current market needs for professional platforms. Although *de novo* sequencing of full-length proteins is still challenging, the rapidly increasing interest in therapeutic human antibodies has undoubtedly already led to the development of better and faster algorithms.

4.2 Application to non-model organisms and cross-species identification

Despite many ongoing sequencing efforts worldwide, most organisms have not been sequenced to date. Particularly non-model organisms are still hard to analyze on the proteome level because appropriate reference sequences for database searching are lacking [6]. A good example is the recent study by Saha *et al.* [91] on the coconut palm, whose genome sequence is still largely undetermined. Therefore, a combination of conventional database searching with MASCOT [92] and manual *de novo* sequencing was used to eventually determine twelve proteins responsible for coconut pollen allergies with MS BLAST [58]. A similar study by Bordas-Le Floch *et al.* [93] evaluated the allergens of the house dust mites *Dermatophagoides farinae* and *D*. *pteronyssinus* and included transcriptome sequencing of the two species. Subsequently, database searches and *de novo* sequencing were performed with PEAKS [34] to determine peptide and protein sequences. Out of a test cohort, 42% of patients reacted positively to the newly discovered allergens. Both of these studies support a similar workflow to detect allergens in non-model organisms. A related research topic is the study of microbial consortia. Such are found almost anywhere on Earth and are also essential to human health. The relatively novel field of metaproteomics deals with the analysis of these microbial communities at the proteome level. The main difference to conventional proteomic studies is that large numbers of different organisms are contained within the samples of interest [94, 95]. This leads to several challenges on the experimental side, and arguably even more severe ones on the computational level. In particular, this refers to missing proteome references and an inflated search space [96]. Since *de novo* sequencing makes it possible to obtain full or partial peptide sequences on the basis of high-quality MS/MS spectra from organisms without sequenced genomes, it has already been used in different metaproteomics studies. For example, a study by Cantarel *et al.* [97] demonstrated that information could be gained when using *de novo* sequencing additionally to conventional reference-based methods. The authors combined the *de novo* peptide predictions of PepNovo and PEAKS, thereby identifying more than 8,000 additional peptides. In a benchmarking study on intestinal metaproteomes [98], *de novo* sequencing was also evaluated for being complementary to database searching. PepNovo could on average recover 23% of the peptides that were obtained using database searching. Recently, Speda *et al.* made use of *de novo* sequencing for a metaproteomics-guided selection of targeted enzymes from mixed microbial communities [99]. Interestingly, they found that the mutation-tolerant SPIDER algorithm could identify more proteins with a different function than the ones identified by PEAKS, suggesting that allowing for sequence mutations might be very useful in this context. However, the authors also encountered the common difficulty of unambiguously linking identified peptides to the correct complete sequence entry, as *de novo* sequence tags rarely cover an entire protein sequence. While this also applies to conventional

database searches, the weak spot of insufficient *de novo* sequence coverage is worsened by the protein inference problem [100]. The proper validation of *de novo* sequencing hits in metaproteomics settings is still an active field of research.

Single amino acid variations between evolutionarily related organisms strongly affect the success of protein identification. In fact, peptides without an exactly matching reference sequence will remain unidentified using classic database search methods. In contrast, if a reference from a (closely or even distantly) related organism is available, homology searching on the basis of *de novo* sequencing results can be employed to alleviate this issue [77]. In 2004, Habermann *et al.* first evaluated this strategy using the MS BLAST protocol [58] for cross-species identification [101]. In another study [102] on a non-human primate species for which database information was limited, *de novo* sequencing and homology searching was successfully used with PEAKS [34]. The BICEPS software [66] makes use of *de novo* sequencing internally, while being tailored to overcome species boundaries in peptide identification. In benchmarking, it showed a similar performance on reference data from remotely related organisms when compared with database search algorithms running on the respective sample-specific database. Very recently, Welker conducted a computational paleoproteomics experiment [103]. Human bone protein samples are searched against three different databases containing sequences with increasing evolutionary distances, from human, chimpanzee and orangutan. Albeit using PEAKS and the mutation-tolerant SPIDER [65], the results of this study confirm that the identification rate decreases with increasing evolutionary distance, and that there is a bias towards conserved sequences. Importantly, a considerable loss in protein hits was observed despite using error-tolerant methods. Overall, this issue can strongly affect the outcome of proteomics studies and, if not anticipated, may lead to incorrect divergence dating and invalid comparisons between samples.

4.3 Venom-based studies

As venoms are causing a noteworthy amount of deaths and injuries worldwide, their characterization from tissues of various toxic animals is an important field of research. Strikingly, the World Health Organization (WHO) has started to consider snake bites a form of Neglected Tropical Disease (NTD) in 2017 [104]. Snake venoms are mixtures of mainly polypeptides and carbohydrates, with proteins being the main component in terms of venom dry weight. To understand the pathogenic processes triggered by venoms and develop efficient treatments, it is essential to study these proteins. Since almost no reference databases exist, *de novo* sequencing has proven useful to this end. In a recent study, de Oliveira *et al.* were able to discover multiple isoforms of crotapotin in the venom of the South American rattlesnake (*Crotalus durissus terrificus*) using *de novo* sequencing

[105]. Further recent studies have applied the technique other organisms, including snails [106, 107], ants [108], scorpions [109-111] and spiders [112]. Mainly PEAKS [34] and manual *de novo* sequencing were used to evaluate the MS data in these studies. The manual approach was shown to enhance the detection of peptide and protein modifications. Trevisan-Silva *et al.* used multiple proteases and dissociation techniques in conjunction with the Meta-SPS pipeline [74] to characterize the venom of the brown spider (*Loxosceles intermedia)* [112]. Conotoxins are oligopeptides found in cone snails and have been the subject of multiple studies. The work of Figueroa-Montiel *et al.* [107] surveys the utilization of conotoxins with antimycobacterial activity as a potential *M. tuberculosis* treatment. Specifically, the results of the sequencing of the venom gland transcriptome of *Conasprella ximenes* were used as a database for proteomic identification with MASCOT [92], ProteinPilot (SCIEX) and PEAKS. In combination with manual *de novo* sequencing of the produced MS/MS spectra, fragmented in two complementary modes, reliable characterization of the conopeptides could be achieved. Abdel-Wahab *et al.* used *de novo* sequencing for the structural and biological characterization of pn3a and pn4c [106], which are conopeptides originating in the venom of *Conus pennaceus*. The analyses of ant venom (*Pachycondyla striata*) by Santos *et al.* [104] revealed a complex mixture of venom proteins, allergenic and bioactive peptides. In this study, spectra produced by MALDI-TOF/TOF and ESI-Q/TOF experiments were sequenced with PEAKS, after searching against the UniProtKB [4] and NCBI [5] databases.

Beside snake bites, scorpion stings present another serious health threat. A study of the venom of *Thorellius atrox* by Romero-Gutierrez *et al.* [110] consisted of an RNA-seq analysis of the venom gland transcriptome followed by bottom-up LC-MS/MS analysis. The latter study included database searching with SEQUEST [2] and *de novo* sequencing with PEAKS. This combination yielded a detailed description of the venom composition. Miyashita *et al.* were able to define new antimicrobial peptides in the venom of *Isometrus maculatus* [109] using *de novo* sequencing based on two types of MS with different peptide fragmentation modes. Finally, Amorim *et al.* used *de novo* sequencing to investigate hyaluronidase rTsHyal-1 from the *Tityus serrulatus* venom [111]. In conclusion, *de novo* sequencing is as a powerful tool in the field of venom research.

4.4 Glycomics and miscellaneous studies

The above-listed studies only cover a fraction of the use cases that have been examined with MS-based *de novo* sequencing so far. Further topics reach beyond classical proteomics applications, for example, into the related field of glycomics. This is concerned with the systematic study of all oligosaccharide structures, so-called glycans, of a given cell type or organism [113]. Glycans constitute a large part of the observed protein modifications and are of high biological relevance.

Also, in the context of personalized medicine and diagnostics, glycoproteins are considered interesting candidates for biomarker discovery, as changes in protein glycosylation are associated with disease states [114, 115]. Since dedicated experimental setups are thus far required to analyze these carbohydrate structures, the determination of glycan sequences directly from MS/MS spectra has become an important matter of research. Since glycans form complex, branched molecules, dedicated algorithms have to cope with a combinatorial explosion of possible structures. Therefore, *de novo* glycan sequencing is considered a challenging and computationally hard problem that has only been addressed by few research groups to date [116-118]. Although different sophisticated algorithms and tools were proposed more recently [119-123], more development concerning efficient glycan structure determination is required. Overall, glycomics will strongly benefit from improvements in both MS instrumentation and *de novo* sequencing algorithms.

A last use case of *de novo* sequencing worth mentioning is the study of bioactive neuropeptides and cyclic peptides. Since such endogenous peptides have various (both beneficial and detrimental) functions in physiological systems and act as transmitters for cells, they can be used as markers for disease detection. *De novo* sequencing is an important tool here, too, as accurate sequence references may be lacking and the peptides are often chemically modified. In their study [124] Knickelbine *et al*. were able to obtain twelve bioactive neuropeptides that are expressed in the nematode *Ascaris suum* using *de novo* sequencing. Ogrinc Potočnik *et al.* [125] successfully applied the technique in determining endogenous neuropeptides via matrix enhanced secondary ion mass spectrometry. Untypically shaped peptide molecules are another interesting use case, such as cyclic peptides: Narayani *et al.* [126] used manual *de novo* sequencing for analysis of so-called cyclotides from the plant *Viola odorata*. Cyclotides are cyclic peptides with cysteine bonds, for which conventional database searches cannot be used. In this case, d*e novo* sequencing enabled the authors to discover three new cyclotides (vodo I1, vodo I2 and vodo I3). The abovementioned studies only constitute few examples for a wide area of application for *de novo* sequencing in various omics-driven fields. It can be expected that, with ongoing and future improvements, in particular with respect to better automated algorithmic solutions, more researchers will be able to use the technique in their studies.

5 Challenges of current methods and novel promising avenues

So far, we have given an overview of the available methods, software tools, and typical use cases for *de novo* sequencing that have been described in the literature. Although many contributions and efforts have been made by various research groups in the past, *de novo* sequencing is still not being widely used within the proteomics community. In the following, we discuss intrinsic limitations of the approach and highlight potential solutions that have been developed to overcome these barriers. Finally, we indicate ideas for future directions that may help the technique to step out of its hitherto exotic niche.

To some extent, the low level of adoption of *de novo* sequencing throughout the proteomics community can be attributed to limitations of the algorithms themselves, leading to insufficient peptide-level accuracy and low coverage when mapping *de novo* peptides to protein references. There are several classical challenges for *de novo* sequencing arising from mass ambiguities: for instance, in low mass accuracy data, algorithms cannot distinguish between glutamine or lysine nor between oxidized methionine and phenylalanine [77]. While better instrumentation may readily solve some of these typical issues, algorithms also need to be extended by parameters fully accounting for higher mass accuracy. Medzihradszky and Chalkley [127] further encourage developers to better understand the biological, chemical, and physical experimental constraints behind the data. Many well-known fragmentation rules are not incorporated into algorithm scoring methods. For example, a neutral loss of 64 Da from the precursor and/or product ions typically occurs in CID for peptides containing oxidized methionine [128], but various algorithms ignore such distinct fragmentation pathways and predict phenylalanine instead. It is also important to acknowledge that more information can be obtained when specific fragment ion types are available as parameters for the algorithms. Prominent examples are typical satellite fragment ions (due to the loss of NH3 or H20) or immonium ions (as markers for specific amino acid modifications) [129], but more fragment ion types and rules can be found across different fragmentation techniques [127]. Further, even when considering all these points, MS/MS spectra still frequently contain significant amounts of peaks that are difficult to interpret or non-interpretable. These signals may originate, for example, from chemical noise or side chain cleavages [130]. The complex mechanisms of peptide fragmentation have been explained in various publications in the past. For instance, peptide fragmentation was described using the ‘mobile proton model’ that provides a general framework for understanding and predicting peptide dissociation in the gas phase [131]. When mobile protons are not available, poor fragmentation may occur that causes uneven fragmentation patterns, which may lead to ambiguous sequence predictions. It was also shown that the position of residues within peptides can also have a significant influence on the peak intensities of fragment ions [132]. The interested reader is also referred to a review article by Paizs and Suhai [133] that summarizes

dissociation chemistry and fragmentation pathways of protonated peptides. Chemical noise, unexplainable peaks and missing signals in MS/MS spectra are critical issues for *de novo* sequencing, as they cause ambiguities and make it very difficult to obtain a resolved peptide sequence in many cases. In this context, Zhang published several works describing a mathematical model that extends the 'mobile proton' framework and considers fragmentation as a series of chemical reactions [134-136]. Importantly, algorithms for automated *de novo* sequencing benefit from such established models: for example, certain features in Novor [42] were inspired by previously published spectrum and fragment ion intensity prediction methods [134, 137].

Another inherent difficulty of *de novo* sequencing is amino acid permutation complexity: the number of residues that potentially match increases with peptide mass, leading to decreased accuracy values for longer peptides in general [55]. In this context, it is further problematic that prediction algorithms often generate different peptide candidates that vary in few residues only for the same spectrum. For example, inversions of two subsequent amino acids frequently occur when the determining fragment ion peak is not available. Consequently, these predictions carry similar or equal confidence scores and can hardly be distinguished or ranked properly due to marginal differences. For many MS/MS spectra, it is very difficult (or even impossible) to establish a common score threshold above which predictions are accepted [98]. This is an example for the more general problem that there is no search space restriction when generating *de novo* sequence candidates, leading to high amounts of false positives. It should be noted here that increased resolution (and higher mass accuracy) of modern MS instruments can counteract these detrimental effects. Nevertheless, mass tolerance windows in the low parts-per-million range are still used more often at the MS rather than at the MS/MS level in most studies. Therefore, the benefit of modern instrumentation might not have been fully exploited so far, although many tools provide the parameter settings to do so. In this context, the choice of fragmentation mode also plays a role: for example, it could be found that the accuracy of *de novo* sequencing is significantly increased for HCD in comparison with CID spectra [36], but is still limited when compared with database searching [55]. In addition, the overall coverage of *de novo* sequencing is not sufficient so far [79]. Despite past and ongoing improvements, issues concerning accuracy and sensitivity in combination with missing control of the false discovery rate (similar to the target-decoy approach for database searching [138]) can still be regarded as the main caveats concerning the validity of the results of *de novo* sequencing. As a consequence, proper quality control mechanisms are required for evaluating the results of *de novo* sequencing prior to downstream analysis. Still, establishing such strategies remains challenging and will require more development work.

Due to the growing number of possibilities, the impact of these challenges on sequencing results is even higher when including variable modifications. *De novo* sequencing with multiple modifications therefore results in both longer processing times and reduced discrimination power between candidate peptides. In addition, for machine learning based algorithms, modified peptides present the additional issue that the amount of data available for training can be limited, especially when considering combinations of different modifications. For database-driven identification, new promising algorithms [147, 148] were recently proposed that perform so-called 'open' searches with a large precursor mass tolerance window to capture peptides with unexpected modifications [12]. In contrast, sequencing algorithms are rarely benchmarked for their ability to identify modified peptides – in some tools modifications are not even supported. Recently, the open-search paradigm was adapted for *de novo* sequencing in an efficient manner presenting an interesting option for discovering unanticipated modifications [50]. Multiple modifications can however easily be used to fill mass gaps, and it is not uncommon to see multiple modifications stacked at the termini of sequences inferred by sequencing or database matching algorithms in order to match the precursor mass. Modified peptides therefore require very careful quality control, ideally including the verification that modified residues are unambiguously flanked by fragment ions.

At this point, different strategies have been proposed either to increase the accuracy or to validate the results of *de novo* sequencing. Approaches for accuracy improvement include combining complementary fragmentation techniques by applying different dissociation strategies, such as CID, HCD, and ETD to the same precursor [37, 38, 75, 139-141], or using overlapping complementary protein digestion methods [84, 112, 142]. Another recently proposed method employs differential chemical labeling of peptides [143] and, on the basis of experimentally disambiguated fragmentation spectra, features a dedicated algorithm for *de novo* sequencing with improved sensitivity and accuracy, particularly for longer peptides, in comparison with previously published algorithms [43]. These are useful strategies for improving the performance of *de novo* sequencing, however, they come with more complex workflows and lowered acquisition rates when spectral acquisition cannot be parallelized. To decrease the number of false positives and increase overall confidence in the results obtained, the combination of different *de novo* sequencing methods was suggested recently [144]. The proposed workflow combines three different algorithms and led to a three-fold increase of peptide identifications at 5% FDR compared to the single best performing algorithm. This combination strategy however requires a robust integration of data, as previously established for database search algorithms. Another downside is that executing multiple algorithms naturally leads to higher overall execution runtimes. Concerning the use of different algorithms, Gorshkov *et al.* [145] conducted a study for evaluating the impact of mixture fragmentation spectra on *de novo*

sequencing performance. Since co-isolated peptides often increase spectrum complexity, a proper deconvolution strategy is required. The authors propose a mixture spectra deconvolution method, tested on four different *de novo* sequencing algorithms, and found more correct sequence predictions when using deconvolution processing. Notably, however, some peptides were only correctly identified using unprocessed data, suggesting that the deconvolution strategy still needs to be improved. Tschager *et al.* [146] recently suggested a new scoring model for *de novo* sequencing, in which the algorithm minimizes the symmetric difference between explained and measured masses. In proof-of-concept experiments, they used synthesized peptides to demonstrate that the approach has a better performance than methods that maximize the shared peaks count. While results on synthetic data are promising, further evaluation on real-world proteomics data sets is required. Very recently, a so-called false amino-acid rate was proposed, defined as the number of incorrectly predicted residues divided by the number of all reported amino acids [44]. This generic confidence measure can be applied to validate the results from different algorithms.

Recently, the large amounts of available experimental proteomics data in public repositories and the significantly increased computational power available (e.g. *via* CPU- or GPU-cluster computing) have led to the application of machine and deep learning algorithms to enhance both accuracy and speed of *de novo* sequencing. In this context, DeepNovo [45] and Novor [42] stand out as flagship examples for application of state-of-the-art machine learning. In this context, the ProteomeTools project [21] (proteometools.org) presents a highly valuable resource for both researchers and developers, with data for more than 330,000 tryptic peptides from the human proteome that have been synthesized and analyzed.

While *de novo* sequencing has often been considered highly time-consuming, this is clearly not the case anymore when using the most recent tools, as we have recently shown in a dedicated benchmarking study [55]. Further, various efforts have been made to make the technique more popular among researchers by developing software tools with graphical user interfaces. For example, such integrating command line-based algorithms (e.g. DeNovoGUI [56]) with dedicated visualization features, or such offering advanced post-processing features for peptide-to-protein mapping and FDR estimation (e.g. PepExplorer [67]). In a similar fashion, the useful mapping of peptide identifications into genome browser visualizations has become increasingly important, and various proteogenomics tools have been proposed [149-151]. It is the responsibility of the developer community to foster such user-friendly developments that go beyond the execution of algorithms in command line to become accepted and applied by those researchers that lack expertise and infrastructure in bioinformatics.

6 Concluding remarks

Computational methods for automated *de novo* sequencing have been constantly improved over the last decades, and its application rate and breadth has increased. For an overview, we have summarized the most significant improvements of the technique in a timeline diagram in **Figure 3**.

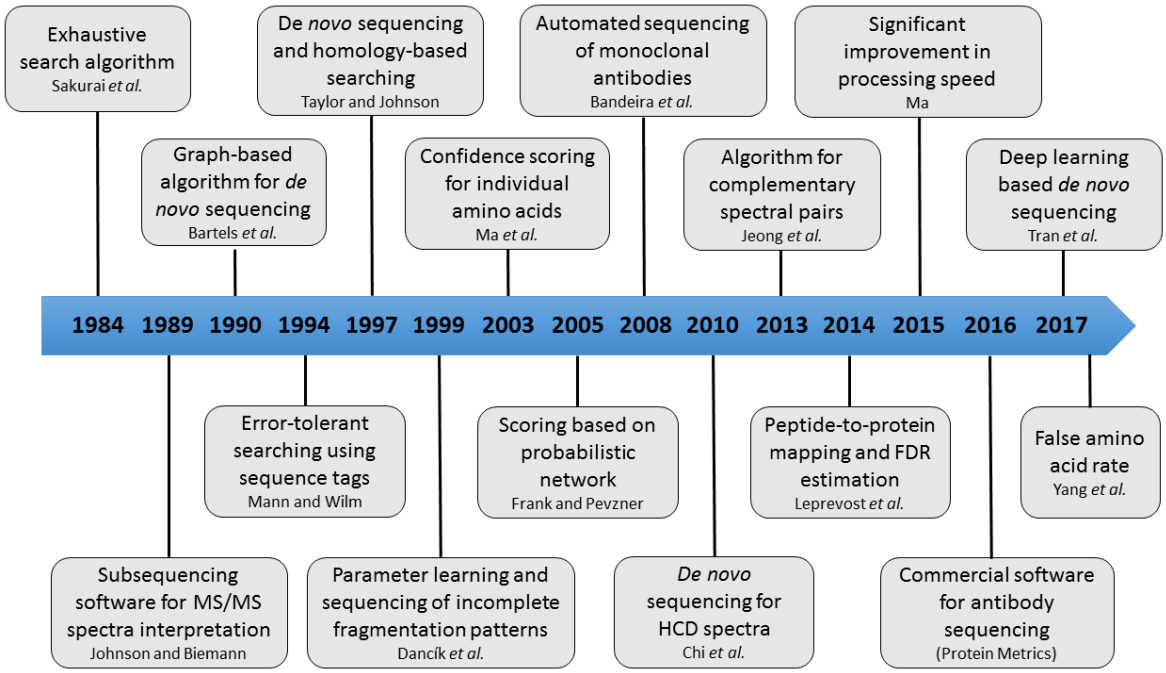

Figure 3. The most significant milestones in the development of *de novo* sequencing. The timeline (blue) indicates the publication year of each milestone publication (grey textboxes).

Despite these advances, *de novo* sequencing is still commonly regarded slow and inaccurate, and therefore not used in most proteomics projects. As discussed above, on the one hand, this can be attributed to the use of overly simplistic scoring methods and algorithms. On the other hand, low accuracy and coverage may also arise from insufficient quality of MS/MS spectra, which has its origins in the error-prone process of peptide fragmentation associated with experimental procedures and instrumentation. In particular, spectral noise and missing fragment ion peaks still render the *de novo* sequencing problem highly difficult. Despite all this, we expect the performance of the approach to further increase in the near future, for the following reasons:

1. Improved mass spectrometers with higher mass accuracy and resolution should lead to further increased overall performance, given that the corresponding algorithms are continuously adapted.
2. Multi-protease and multi-dissociation strategies will evolve further and, once they reach the critical point of being time and cost efficient enough to enable analyses in high throughput, may be routinely applied to enhance *de novo* sequencing results.

3. Algorithmic developments, particularly sophisticated dynamic programming and machine learning approaches, should also lead to a further performance increase. The most important factors driving these developments are:
    a. the increased availability of proteomics data in public repositories
    b. benchmarking studies performing independent comparisons between algorithms
    c. data standards that provide for a better integration with other peptide identification approaches, such as database and spectral library searching

Joint efforts of the bioinformatics and analytical proteomics communities will be necessary to overcome limitations, for instance, by means of improving the understanding of peptide fragmentation in order to develop more appropriate scoring models. The primary effect of better performing methods should be an increased application of *de novo* sequencing in all kinds of proteomics studies, instead of only those cases where database searching cannot be readily applied. At the same time, the integration with reference-based methods will become more and more important. A practical scenario would be that *de novo* sequencing is executed as an obligatory second step to detect unexpected protein sequence variations, which would remain undetected in database-only workflows. Overall, with further algorithmic and technical improvements, the 'golden age' of *de novo* sequencing may be just around the corner. Eventually, the technique should be clearly superior to database search engines, not only in niche applications, thanks to its speed and unbiased way of obtaining sequence information.

Acknowledgments

Bernhard Y. Renard acknowledges financial support by Deutsche Forschungsgemeinschaft (DFG), grant number RE 3474/2-2.

The authors would like to thank Tobias Loka and Dr. Robert Rentzsch for critical reading of the manuscript and their valuable suggestions.

The authors have declared no conflict of interest.

7 References

[1] Henzel, W. J., Billeci, T. M., Stults, J. T., Wong, S. C.*, et al.*, Identifying proteins from two-dimensional gels by molecular mass searching of peptide fragments in protein sequence databases. *Proc Natl Acad Sci U S A* 1993, *90*, 5011-5015.

[2] Eng, J. K., McCormack, A. L., Yates, J. R., An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J Am Soc Mass Spectrom* 1994, *5*, 976-989.

[3] Verheggen, K., Raeder, H., Berven, F. S., Martens, L.*, et al.*, Anatomy and evolution of database search engines-a central component of mass spectrometry based proteomic workflows. *Mass Spectrom Rev* 2017.

[4] Apweiler, R., Bairoch, A., Wu, C. H., Barker, W. C.*, et al.*, UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res* 2004, *32*, D115-119.

[5] Pruitt, K. D., Tatusova, T., Maglott, D. R., NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* 2005, *33*, D501-504.

[6] Armengaud, J., Trapp, J., Pible, O., Geffard, O.*, et al.*, Non-model organisms, a species endangered by proteogenomics. *J Proteomics* 2014, *105*, 5-18.

[7] Kuhring, M., Renard, B. Y., Estimating the computational limits of detection of microbial non-model organisms. *Proteomics* 2015, *15*, 3580-3584.

[8] Seifert, J., Herbst, F. A., Halkjaer Nielsen, P., Planes, F. J.*, et al.*, Bioinformatic progress and applications in metaproteogenomics for bridging the gap between genomic sequences and metabolic functions in microbial communities. *Proteomics* 2013, *13*, 2786-2804.

[9] Wilmes, P., Heintz-Buschart, A., Bond, P. L., A decade of metaproteomics: where we stand and what the future holds. *Proteomics* 2015, *15*, 3409-3417.

[10] Zhu, Y., Hultin-Rosenberg, L., Forshed, J., Branca, R. M.*, et al.*, SpliceVista, a tool for splice variant identification and visualization in shotgun proteomics data. *Mol Cell Proteomics* 2014, *13*, 1552-1562.

[11] Giese, S. H., Zickmann, F., Renard, B. Y., Detection of Unknown Amino Acid Substitutions Using Error-Tolerant Database Search. *Methods Mol Biol* 2016, *1362*, 247-264.

[12] Chick, J. M., Kolippakkam, D., Nusinow, D. P., Zhai, B.*, et al.*, A mass-tolerant database search identifies a large proportion of unassigned spectra in shotgun proteomics as modified peptides. *Nat Biotechnol* 2015, *33*, 743-749.

[13] Ellis, M. J., Gillette, M., Carr, S. A., Paulovich, A. G.*, et al.*, Connecting genomic alterations to cancer biology with proteomics: the NCI Clinical Proteomic Tumor Analysis Consortium. *Cancer Discov* 2013, *3*, 1108-1112.

[14] Mertins, P., Mani, D. R., Ruggles, K. V., Gillette, M. A.*, et al.*, Proteogenomics connects somatic mutations to signalling in breast cancer. *Nature* 2016, *534*, 55-62.

[15] Zickmann, F., Renard, B. Y., MSProGene: integrative proteogenomics beyond six-frames and single nucleotide polymorphisms. *Bioinformatics* 2015, *31*, 106-115.

[16] Lam, H., Deutsch, E. W., Eddes, J. S., Eng, J. K.*, et al.*, Development and validation of a spectral library searching method for peptide identification from MS/MS. *Proteomics* 2007, *7*, 655-667.

[17] Lam, H., Building and searching tandem mass spectral libraries for peptide identification. *Mol Cell Proteomics* 2011, *10*, R111 008565.

[18] Zhang, X., Li, Y., Shao, W., Lam, H., Understanding the improved sensitivity of spectral library searching over sequence database searching in proteomics data analysis. *Proteomics* 2011, *11*, 1075-1085.

[19] Griss, J., Spectral library searching in proteomics. *Proteomics* 2016, *16*, 729-740.

[20] Bartels, C., Fast algorithm for peptide sequencing by mass spectroscopy. *Biomed Environ Mass Spectrom* 1990, *19*, 363-368.

[21] Zolg, D. P., Wilhelm, M., Schnatbaum, K., Zerweck, J.*, et al.*, Building ProteomeTools based on a complete synthetic human proteome. *Nat Methods* 2017, *14*, 259-262.

[22] Strohalm, M., Kavan, D., Novak, P., Volny, M., Havlicek, V., mMass 3: a cross-platform software environment for precise analysis of mass spectrometric data. *Anal Chem* 2010, *82*, 4648-4651.
[23] Degroeve, S., Martens, L., MS2PIP: a tool for MS/MS peak intensity prediction. *Bioinformatics* 2013, *29*, 3199-3203.
[24] Sakurai, T., Matsuo, T., Matsuda, H., Katakuse, I., PAAS 3: A computer program to determine probable sequence of peptides from mass spectrometric data. *Biological Mass Spectrometry* 1984, *11*, 396-399.
[25] Hamm, C. W., Wilson, W. E., Harvan, D. J., Peptide sequencing program. *Comput Appl Biosci* 1986, *2*, 115-118.
[26] Siegel, M. M., Bauman, N., An efficient algorithm for sequencing peptides using fast atom bombardment mass spectral data. *Biological Mass Spectrometry* 1988, *15*, 333-343.
[27] Johnson, R. S., Biemann, K., Computer program (SEQPEP) to aid in the interpretation of high-energy collision tandem mass spectra of peptides. *Biomed Environ Mass Spectrom* 1989, *18*, 945-957.
[28] Dancik, V., Addona, T. A., Clauser, K. R., Vath, J. E., Pevzner, P. A., De novo peptide sequencing via tandem mass spectrometry. *J Comput Biol* 1999, *6*, 327-342.
[29] Taylor, J. A., Johnson, R. S., Sequence database searches via de novo peptide sequencing by tandem mass spectrometry. *Rapid Commun Mass Spectrom* 1997, *11*, 1067-1075.
[30] Andreotti, S., Klau, G. W., Reinert, K., Antilope--a Lagrangian relaxation approach to the de novo peptide sequencing problem. *IEEE/ACM Trans Comput Biol Bioinform* 2012, *9*, 385-394.
[31] DiMaggio, P. A., Jr., Floudas, C. A., De novo peptide identification via tandem mass spectrometry and integer linear optimization. *Anal Chem* 2007, *79*, 1433-1446.
[32] Mo, L., Dutta, D., Wan, Y., Chen, T., MSNovo: a dynamic programming algorithm for de novo peptide sequencing via tandem mass spectrometry. *Anal Chem* 2007, *79*, 4870-4878.
[33] Chen, T., Kao, M. Y., Tepel, M., Rush, J., Church, G. M., A dynamic programming approach to de novo peptide sequencing via tandem mass spectrometry. *J Comput Biol* 2001, *8*, 325-337.
[34] Ma, B., Zhang, K., Hendrie, C., Liang, C.*, et al.*, PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. *Rapid Commun Mass Spectrom* 2003, *17*, 2337-2342.
[35] Frank, A., Pevzner, P., PepNovo: de novo peptide sequencing via probabilistic network modeling. *Anal Chem* 2005, *77*, 964-973.
[36] Chi, H., Sun, R. X., Yang, B., Song, C. Q.*, et al.*, pNovo: de novo peptide sequencing and identification using HCD spectra. *J Proteome Res* 2010, *9*, 2713-2724.
[37] Chi, H., Chen, H., He, K., Wu, L.*, et al.*, pNovo+: de novo peptide sequencing using complementary HCD and ETD tandem mass spectra. *J Proteome Res* 2013, *12*, 615-625.
[38] Jeong, K., Kim, S., Pevzner, P. A., UniNovo: a universal tool for de novo peptide sequencing. *Bioinformatics* 2013, *29*, 1953-1962.
[39] Zhang, Z., De novo peptide sequencing based on a divide-and-conquer algorithm and peptide tandem spectrum simulation. *Anal Chem* 2004, *76*, 6374-6383.
[40] Fischer, B., Roth, V., Roos, F., Grossmann, J.*, et al.*, NovoHMM: a hidden Markov model for de novo peptide sequencing. *Anal Chem* 2005, *77*, 7265-7273.
[41] Robotham, S. A., Horton, A. P., Cannon, J. R., Cotham, V. C.*, et al.*, UVnovo: A de Novo Sequencing Algorithm Using Single Series of Fragment Ions via Chromophore Tagging and 351 nm Ultraviolet Photodissociation Mass Spectrometry. *Analytical Chemistry* 2016, *88*, 3990-3997.
[42] Ma, B., Novor: real-time peptide de novo sequencing software. *J Am Soc Mass Spectrom* 2015, *26*, 1885-1894.
[43] Devabhaktuni, A., Elias, J. E., Application of de Novo Sequencing to Large-Scale Complex Proteomics Data Sets. *Journal of Proteome Research* 2016, *15*, 732-742.
[44] Yang, H., Chi, H., Zhou, W. J., Zeng, W. F.*, et al.*, pSite: Amino Acid Confidence Evaluation for Quality Control of De Novo Peptide Sequencing and Modification Site Localization. *J Proteome Res* 2018, *17*, 119-128.

[45] Tran, N. H., Zhang, X., Xin, L., Shan, B., Li, M., De novo peptide sequencing by deep learning. *Proc Natl Acad Sci U S A* 2017.
[46] Tabb, D. L., Ma, Z. Q., Martin, D. B., Ham, A. J., Chambers, M. C., DirecTag: accurate sequence tags from peptide MS/MS through statistical scoring. *J Proteome Res* 2008, *7*, 3838-3846.
[47] Sunyaev, S., Liska, A. J., Golod, A., Shevchenko, A., Shevchenko, A., MultiTag: multiple error-tolerant sequence tag search for the sequence-similarity identification of proteins by mass spectrometry. *Anal Chem* 2003, *75*, 1307-1315.
[48] Tabb, D. L., Saraf, A., Yates, J. R., 3rd, GutenTag: high-throughput sequence tagging via an empirically derived fragmentation model. *Anal Chem* 2003, *75*, 6415-6421.
[49] Mann, M., Wilm, M., Error-tolerant identification of peptides in sequence databases by peptide sequence tags. *Anal Chem* 1994, *66*, 4390-4399.
[50] Yang, H., Chi, H., Zhou, W. J., Zeng, W. F.*, et al.*, Open-pNovo: De Novo Peptide Sequencing with Thousands of Protein Modifications. *J Proteome Res* 2017, *16*, 645-654.
[51] Henry, V. J., Bandrowski, A. E., Pepin, A. S., Gonzalez, B. J., Desfeux, A., OMICtools: an informative directory for multi-omic data analysis. *Database (Oxford)* 2014, *2014*.
[52] Fernandez-de-Cossio, J., Gonzalez, J., Satomi, Y., Shima, T.*, et al.*, Automated interpretation of low-energy collision-induced dissociation spectra by SeqMS, a software aid for de novo sequencing by tandem mass spectrometry. *Electrophoresis* 2000, *21*, 1694-1699.
[53] Vyatkina, K., Wu, S., Dekker, L. J., VanDuijn, M. M.*, et al.*, De Novo Sequencing of Peptides from Top-Down Tandem Mass Spectra. *J Proteome Res* 2015, *14*, 4450-4462.
[54] Li, C., Chen, T., He, Q., Zhu, Y., Li, K., MRUniNovo: an efficient tool for de novo peptide sequencing utilizing the hadoop distributed computing framework. *Bioinformatics* 2017, *33*, 944-946.
[55] Muth, T., Renard, B. Y., Evaluating de novo sequencing in proteomics: already an accurate alternative to database-driven peptide identification? *Brief Bioinform* 2017.
[56] Muth, T., Weilnbock, L., Rapp, E., Huber, C. G.*, et al.*, DeNovoGUI: an open source graphical user interface for de novo sequencing of tandem mass spectra. *J Proteome Res* 2014, *13*, 1143-1146.
[57] Altschul, S. F., Gish, W., Miller, W., Myers, E. W., Lipman, D. J., Basic local alignment search tool. *J Mol Biol* 1990, *215*, 403-410.
[58] Shevchenko, A., Sunyaev, S., Loboda, A., Bork, P.*, et al.*, Charting the proteomes of organisms with unsequenced genomes by MALDI-quadrupole time-of-flight mass spectrometry and BLAST homology searching. *Anal Chem* 2001, *73*, 1917-1926.
[59] Huang, L., Jacob, R. J., Pegg, S. C., Baldwin, M. A.*, et al.*, Functional assignment of the 20 S proteasome from Trypanosoma brucei using mass spectrometry and new bioinformatics approaches. *J Biol Chem* 2001, *276*, 28327-28339.
[60] Mackey, A. J., Haystead, T. A., Pearson, W. R., Getting more from less: algorithms for rapid protein identification with multiple short peptide sequences. *Mol Cell Proteomics* 2002, *1*, 139-147.
[61] Dasari, S., Chambers, M. C., Slebos, R. J., Zimmerman, L. J.*, et al.*, TagRecon: high-throughput mutation identification through sequence tagging. *J Proteome Res* 2010, *9*, 1716-1726.
[62] Wilmarth, P. A., Tanner, S., Dasari, S., Nagalla, S. R.*, et al.*, Age-related changes in human crystallins determined from comparative analysis of post-translational modifications in young and aged lens: does deamidation contribute to crystallin insolubility? *J Proteome Res* 2006, *5*, 2554-2566.
[63] Tanner, S., Shu, H., Frank, A., Wang, L. C.*, et al.*, InsPecT: identification of posttranslationally modified peptides from tandem mass spectra. *Anal Chem* 2005, *77*, 4626-4639.
[64] Zhang, J., Xin, L., Shan, B., Chen, W.*, et al.*, PEAKS DB: de novo sequencing assisted database search for sensitive and accurate peptide identification. *Mol Cell Proteomics* 2012, *11*, M111 010587.
[65] Han, Y., Ma, B., Zhang, K., SPIDER: software for protein identification from sequence tags with de novo sequencing error. *J Bioinform Comput Biol* 2005, *3*, 697-716.
[66] Renard, B. Y., Xu, B., Kirchner, M., Zickmann, F.*, et al.*, Overcoming species boundaries in peptide identification with Bayesian information criterion-driven error-tolerant peptide search (BICEPS). *Mol Cell Proteomics* 2012, *11*, M111 014167.

[67] Leprevost, F. V., Valente, R. H., Borges, D. L., Perales, J.*, et al.*, PepExplorer: a similarity-driven tool for analyzing de novo sequencing results. *Mol Cell Proteomics* 2014.
[68] da Veiga Leprevost, F., Barbosa, V. C., Carvalho, P. C., Using PepExplorer to Filter and Organize De Novo Peptide Sequencing Results. *Curr Protoc Bioinformatics* 2015, *51*, 13 27 11-19.
[69] Mesuere, B., Devreese, B., Debyser, G., Aerts, M.*, et al.*, Unipept: tryptic peptide-based biodiversity analysis of metaproteome samples. *J Proteome Res* 2012, *11*, 5773-5780.
[70] Kopczynski, D., Barsnes, H., Njolstad, P. R., Sickmann, A.*, et al.*, PeptideMapper: efficient and versatile amino acid sequence and tag mapping. *Bioinformatics* 2017, *33*, 2042-2044.
[71] Ferragina, P., Manzini, G., *Foundations of Computer Science, 2000. Proceedings. 41st Annual Symposium on*, IEEE 2000, pp. 390-398.
[72] Reinert, K., Langmead, B., Weese, D., Evers, D. J., Alignment of Next-Generation Sequencing Reads. *Annu Rev Genomics Hum Genet* 2015, *16*, 133-151.
[73] Vaudel, M., Burkhart, J. M., Zahedi, R. P., Oveland, E.*, et al.*, PeptideShaker enables reanalysis of MS-derived proteomics data sets. *Nat Biotechnol* 2015, *33*, 22-24.
[74] Guthals, A., Clauser, K. R., Bandeira, N., Shotgun protein sequencing with meta-contig assembly. *Mol Cell Proteomics* 2012, *11*, 1084-1096.
[75] Guthals, A., Clauser, K. R., Frank, A. M., Bandeira, N., Sequencing-grade de novo analysis of MS/MS triplets (CID/HCD/ETD) from overlapping peptides. *J Proteome Res* 2013, *12*, 2846-2857.
[76] Lu, B., Chen, T., Algorithms for de novo peptide sequencing using tandem mass spectrometry. *Drug Discovery Today: BIOSILICO* 2004, *2*, 85-90.
[77] Ma, B., Johnson, R., De novo sequencing and homology searching. *Mol Cell Proteomics* 2012, *11*, O111 014902.
[78] Allmer, J., Algorithms for the de novo sequencing of peptides from tandem mass spectra. *Expert Rev Proteomics* 2011, *8*, 645-657.
[79] Seidler, J., Zinn, N., Boehm, M. E., Lehmann, W. D., De novo sequencing of peptides by MS/MS. *Proteomics* 2010, *10*, 634-649.
[80] Sen, K. I., Tang, W. H., Nayak, S., Kil, Y. J.*, et al.*, Automated Antibody De Novo Sequencing and Its Utility in Biopharmaceutical Discovery. *J Am Soc Mass Spectrom* 2017, *28*, 803-810.
[81] Beck, A., Wurch, T., Bailly, C., Corvaia, N., Strategies and challenges for the next generation of therapeutic antibodies. *Nature Reviews Immunology* 2010, *10*, 345.
[82] Zhang, Z., Pan, H., Chen, X., Mass spectrometry for structural characterization of therapeutic antibodies. *Mass Spectrom Rev* 2009, *28*, 147-176.
[83] Plomp, R., Bondt, A., de Haan, N., Rombouts, Y., Wuhrer, M., Recent Advances in Clinical Glycoproteomics of Immunoglobulins (Igs). *Mol Cell Proteomics* 2016, *15*, 2217-2228.
[84] Pham, V., Henzel, W. J., Arnott, D., Hymowitz, S.*, et al.*, De novo proteomic sequencing of a monoclonal antibody raised against OX40 ligand. *Anal Biochem* 2006, *352*, 77-86.
[85] Bandeira, N., Pham, V., Pevzner, P., Arnott, D., Lill, J. R., Automated de novo protein sequencing of monoclonal antibodies. *Nature Biotechnology* 2008, *26*, 1336-1338.
[86] Tran, N. H., Rahman, M. Z., He, L., Xin, L.*, et al.*, Complete De Novo Assembly of Monoclonal Antibody Sequences. *Sci Rep* 2016, *6*, 31730.
[87] Bogdanoff, W. A., Morgenstern, D., Bern, M., Ueberheide, B. M.*, et al.*, De Novo Sequencing and Resurrection of a Human Astrovirus-Neutralizing Antibody. *Acs Infect Dis* 2016, *2*, 313-321.
[88] Bern, M., Kil, Y. J., Becker, C., Byonic: advanced peptide and protein identification software. *Curr Protoc Bioinformatics* 2012, *Chapter 13*, Unit13 20.
[89] Guthals, A., Gan, Y., Murray, L., Chen, Y.*, et al.*, De Novo MS/MS Sequencing of Native Human Antibodies. *J Proteome Res* 2017, *16*, 45-54.
[90] Savidor, A., Barzilay, R., Elinger, D., Yarden, Y.*, et al.*, Database-independent Protein Sequencing (DiPS) Enables Full-length de Novo Protein and Antibody Sequence Determination. *Mol Cell Proteomics* 2017, *16*, 1151-1161.

[91] Saha, B., Sircar, G., Pandey, N., Gupta Bhattacharya, S., Mining Novel Allergens from Coconut Pollen Employing Manual De Novo Sequencing and Homology-Driven Proteomics. *J Proteome Res* 2015, *14*, 4823-4833.
[92] Perkins, D. N., Pappin, D. J., Creasy, D. M., Cottrell, J. S., Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* 1999, *20*, 3551-3567.
[93] Bordas-Le Floch, V., Le Mignon, M., Bussieres, L., Jain, K*., et al.*, A combined transcriptome and proteome analysis extends the allergome of house dust mite Dermatophagoides species. *PLoS One* 2017, *12*, e0185830.
[94] Wilmes, P., Bond, P. L., Metaproteomics: studying functional gene expression in microbial ecosystems. *Trends Microbiol* 2006, *14*, 92-97.
[95] Hettich, R. L., Pan, C., Chourey, K., Giannone, R. J., Metaproteomics: harnessing the power of high performance mass spectrometry to identify the suite of proteins that control metabolic activities in microbial communities. *Anal Chem* 2013, *85*, 4203-4214.
[96] Muth, T., Renard, B. Y., Martens, L., Metaproteomic data analysis at a glance: advances in computational microbial community proteomics. *Expert Rev Proteomics* 2016, *13*, 757-769.
[97] Cantarel, B. L., Erickson, A. R., VerBerkmoes, N. C., Erickson, B. K*., et al.*, Strategies for metagenomic-guided whole-community proteomics of complex microbial environments. *PLoS One* 2011, *6*, e27173.
[98] Muth, T., Kolmeder, C. A., Salojarvi, J., Keskitalo, S*., et al.*, Navigating through metaproteomics data: a logbook of database searching. *Proteomics* 2015, *15*, 3439-3453.
[99] Speda, J., Jonsson, B. H., Carlsson, U., Karlsson, M., Metaproteomics-guided selection of targeted enzymes for bioprospecting of mixed microbial communities. *Biotechnol Biofuels* 2017, *10*, 128.
[100] Nesvizhskii, A. I., Aebersold, R., Interpretation of shotgun proteomic data: the protein inference problem. *Mol Cell Proteomics* 2005, *4*, 1419-1440.
[101] Habermann, B., Oegema, J., Sunyaev, S., Shevchenko, A., The power and the limitations of cross-species protein identification by mass spectrometry-driven sequence similarity searches. *Mol Cell Proteomics* 2004, *3*, 238-249.
[102] Tannu, N. S., Hemby, S. E., De novo protein sequence analysis of Macaca mulatta. *BMC Genomics* 2007, *8*, 270.
[103] Welker, F., Elucidation of cross-species proteomic effects in human and hominin bone proteome identification through a bioinformatics experiment. *BMC Evol Biol* 2018, *18*, 23.
[104] Chippaux, J. P., Snakebite envenomation turns again into a neglected tropical disease! *J Venom Anim Toxins Incl Trop Dis* 2017, *23*, 38.
[105] de Oliveira, L. A., Ferreira, R. S., Jr., Barraviera, B., de Carvalho, F. C. T*., et al.*, Crotalus durissus terrificus crotapotin naturally displays preferred positions for amino acid substitutions. *J Venom Anim Toxins Incl Trop Dis* 2017, *23*, 46.
[106] Abdel-Wahab, M., Miyashita, M., Ota, Y., Juichi, H*., et al.*, Isolation, structural identification and biological characterization of two conopeptides from the Conus pennaceus venom. *Biosci Biotechnol Biochem* 2017, *81*, 2086-2089.
[107] Figueroa-Montiel, A., Bernaldez, J., Jimenez, S., Ueberhide, B*., et al.*, Antimycobacterial Activity: A New Pharmacological Target for Conotoxins Found in the First Reported Conotoxin from Conasprella ximenes. *Toxins (Basel)* 2018, *10*.
[108] Santos, P. P., Games, P. D., Azevedo, D. O., Barros, E*., et al.*, Proteomic analysis of the venom of the predatory ant Pachycondyla striata (Hymenoptera: Formicidae). *Arch Insect Biochem Physiol* 2017, *96*.
[109] Miyashita, M., Kitanaka, A., Yakio, M., Yamazaki, Y*., et al.*, Complete de novo sequencing of antimicrobial peptides in the venom of the scorpion Isometrus maculatus. *Toxicon* 2017, *139*, 1-12.

[110] Romero-Gutierrez, T., Peguero-Sanchez, E., Cevallos, M. A., Batista, C. V. F.*, et al.*, A Deeper Examination of Thorellius atrox Scorpion Venom Components with Omic Techonologies. *Toxins (Basel)* 2017, *9*.
[111] Amorim, F. G., Boldrini-Franca, J., de Castro Figueiredo Bordon, K., Cardoso, I. A.*, et al.*, Heterologous expression of rTsHyal-1: the first recombinant hyaluronidase of scorpion venom produced in Pichia pastoris system. *Appl Microbiol Biotechnol* 2018.
[112] Trevisan-Silva, D., Bednaski, A. V., Fischer, J. S. G., Veiga, S. S.*, et al.*, A multi-protease, multi-dissociation, bottom-up-to-top-down proteomic view of the Loxosceles intermedia venom. *Sci Data* 2017, *4*, 170090.
[113] Raman, R., Raguram, S., Venkataraman, G., Paulson, J. C., Sasisekharan, R., Glycomics: an integrated systems approach to structure-function relationships of glycans. *Nat Methods* 2005, *2*, 817-824.
[114] Hennig, R., Cajic, S., Borowiak, M., Hoffmann, M.*, et al.*, Towards personalized diagnostics via longitudinal study of the human plasma N-glycome. *Biochim Biophys Acta* 2016, *1860*, 1728-1738.
[115] Almeida, A., Kolarich, D., The promise of protein glycosylation for personalised medicine. *Biochim Biophys Acta* 2016, *1860*, 1583-1595.
[116] Bocker, S., Kehr, B., Rasche, F., Determination of glycan structure from tandem mass spectra. *IEEE/ACM Trans Comput Biol Bioinform* 2011, *8*, 976-986.
[117] Shan, B., Ma, B., Zhang, K., Lajoie, G., Complexities and algorithms for glycan sequencing using tandem mass spectrometry. *J Bioinform Comput Biol* 2008, *6*, 77-91.
[118] Tang, H., Mechref, Y., Novotny, M. V., Automated interpretation of MS/MS spectra of oligosaccharides. *Bioinformatics* 2005, *21 Suppl 1*, i431-439.
[119] Tang, Y., Pu, Y., Gao, J., Hong, P.*, et al.*, De Novo Glycan Sequencing by Electronic Excitation Dissociation and Fixed-Charge Derivatization. *Anal Chem* 2018.
[120] Horlacher, O., Jin, C., Alocci, D., Mariethoz, J.*, et al.*, Glycoforest 1.0. *Anal Chem* 2017, *89*, 10932-10940.
[121] Hong, P., Sun, H., Sha, L., Pu, Y.*, et al.*, GlycoDeNovo - an Efficient Algorithm for Accurate de novo Glycan Topology Reconstruction from Tandem Mass Spectra. *J Am Soc Mass Spectrom* 2017, *28*, 2288-2301.
[122] Kumozaki, S., Sato, K., Sakakibara, Y., A Machine Learning Based Approach to de novo Sequencing of Glycans from Tandem Mass Spectrometry Spectrum. *IEEE/ACM Trans Comput Biol Bioinform* 2015, *12*, 1267-1274.
[123] Dong, L., Shi, B., Tian, G., Li, Y.*, et al.*, An Accurate de novo Algorithm for Glycan Topology Determination from Mass Spectra. *IEEE/ACM Trans Comput Biol Bioinform* 2015, *12*, 568-578.
[124] Knickelbine, J. J., Konop, C. J., Viola, I. R., Rogers, C. B.*, et al.*, Different Bioactive Neuropeptides are Expressed in Two Sub-Classes of GABAergic RME Nerve Ring Motorneurons in Ascaris suum. *ACS Chem Neurosci* 2018.
[125] Ogrinc Potocnik, N., Fisher, G. L., Prop, A., Heeren, R. M. A., Sequencing and Identification of Endogenous Neuropeptides with Matrix-Enhanced Secondary Ion Mass Spectrometry Tandem Mass Spectrometry. *Anal Chem* 2017, *89*, 8223-8227.
[126] Narayani, M., Chadha, A., Srivastava, S., Cyclotides from the Indian Medicinal Plant Viola odorata (Banafsha): Identification and Characterization. *J Nat Prod* 2017, *80*, 1972-1980.
[127] Medzihradszky, K. F., Chalkley, R. J., Lessons in de novo peptide sequencing by tandem mass spectrometry. *Mass Spectrom Rev* 2015, *34*, 43-63.
[128] Guan, Z., Yates, N. A., Bakhtiar, R., Detection and characterization of methionine oxidation in peptides by collision-induced dissociation and electron capture dissociation. *J Am Soc Mass Spectrom* 2003, *14*, 605-613.
[129] Steen, H., Mann, M., The ABC's (and XYZ's) of peptide sequencing. *Nat Rev Mol Cell Biol* 2004, *5*, 699-711.
[130] Cottrell, J. S., Protein identification using MS/MS data. *J Proteomics* 2011, *74*, 1842-1851.

[131] Wysocki, V. H., Tsaprailis, G., Smith, L. L., Breci, L. A., Mobile and localized protons: a framework for understanding peptide dissociation. *J Mass Spectrom* 2000, *35*, 1399-1406.
[132] Tabb, D. L., Huang, Y., Wysocki, V. H., Yates, J. R., 3rd, Influence of basic residue content on fragment ion peak intensities in low-energy collision-induced dissociation spectra of peptides. *Anal Chem* 2004, *76*, 1243-1248.
[133] Paizs, B., Suhai, S., Fragmentation pathways of protonated peptides. *Mass Spectrom Rev* 2005, *24*, 508-548.
[134] Zhang, Z., Prediction of low-energy collision-induced dissociation spectra of peptides. *Anal Chem* 2004, *76*, 3908-3922.
[135] Zhang, Z., Prediction of low-energy collision-induced dissociation spectra of peptides with three or more charges. *Anal Chem* 2005, *77*, 6364-6373.
[136] Zhang, Z., Prediction of collision-induced-dissociation spectra of peptides with post-translational or process-induced modifications. *Anal Chem* 2011, *83*, 8642-8651.
[137] Sun, S., Yang, F., Yang, Q., Zhang, H.*, et al.*, MS-Simulator: predicting y-ion intensities for peptides with two charges based on the intensity ratio of neighboring ions. *J Proteome Res* 2012, *11*, 4509-4516.
[138] Elias, J. E., Gygi, S. P., Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat Methods* 2007, *4*, 207-214.
[139] Savitski, M. M., Nielsen, M. L., Kjeldsen, F., Zubarev, R. A., Proteomics-grade de novo sequencing approach. *J Proteome Res* 2005, *4*, 2348-2354.
[140] Savitski, M. M., Nielsen, M. L., Zubarev, R. A., New data base-independent, sequence tag-based scoring of peptide MS/MS data validates Mowse scores, recovers below threshold data, singles out modified peptides, and assesses the quality of MS/MS techniques. *Mol Cell Proteomics* 2005, *4*, 1180-1188.
[141] Datta, K. K., Madugundu, A. K., Gowda, H., Proteogenomic Methods to Improve Genome Annotation. *Methods Mol Biol* 2016, *1410*, 77-89.
[142] Bandeira, N., Clauser, K. R., Pevzner, P. A., Shotgun protein sequencing: assembly of peptide tandem mass spectra from mixtures of modified proteins. *Mol Cell Proteomics* 2007, *6*, 1123-1134.
[143] Ji, C., Guo, N., Li, L., Differential dimethyl labeling of N-termini of peptides after guanidination for proteome analysis. *J Proteome Res* 2005, *4*, 2099-2108.
[144] Blank-Landeshammer, B., Kollipara, L., Biss, K., Pfenninger, M.*, et al.*, Combining De Novo Peptide Sequencing Algorithms, A Synergistic Approach to Boost Both Identifications and Confidence in Bottom-up Proteomics. *J Proteome Res* 2017, *16*, 3209-3218.
[145] Gorshkov, V., Hotta, S. Y., Verano-Braga, T., Kjeldsen, F., Peptide de novo sequencing of mixture tandem mass spectra. *Proteomics* 2016, *16*, 2470-2479.
[146] Tschager, T., Rosch, S., Gillet, L., Widmayer, P., A better scoring model for de novo peptide sequencing: the symmetric difference between explained and measured masses. *Algorithms Mol Biol* 2017, *12*, 12.
[147] Kong, A. T., Leprevost, F. V., Avtonomov, D. M., Mellacheruvu, D., Nesvizhskii, A. I., MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics. *Nat Methods* 2017, *14*, 513-520.
[148] Chi, H., He, K., Yang, B., Chen, Z.*, et al.*, pFind-Alioth: A novel unrestricted database search algorithm to improve the interpretation of high-resolution MS/MS data. *J Proteomics* 2015, *125*, 89-97.
[149] Kuhring, M., Renard, B. Y., iPiG: integrating peptide spectrum matches into genome browser visualizations. *PLoS One* 2012, *7*, e50246.
[150] Pang, C. N., Tay, A. P., Aya, C., Twine, N. A.*, et al.*, Tools to covisualize and coanalyze proteomic data with genomes and transcriptomes: validation of genes and alternative mRNA splicing. *J Proteome Res* 2014, *13*, 84-98.

[151] Wang, X., Slebos, R. J., Chambers, M. C., Tabb, D. L.*, et al.*, proBAMsuite, a Bioinformatics Framework for Genome-Based Representation and Analysis of Proteomics Data. *Mol Cell Proteomics* 2016, *15*, 1164-1175.