EXPERT
REVIEW
OF PROTEOMICS
informa

Expert Review of Proteomics

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/ieru20

De novo sequencing of proteins by mass spectrometry

Rui Vitorino , Sofia Guedes , Fabio Trindade , Inês Correia , Gabriela Moura , Paulo Carvalho , Manuel Santos & Francisco Amado

To cite this article: Rui Vitorino , Sofia Guedes , Fabio Trindade , Inês Correia , Gabriela Moura , Paulo Carvalho , Manuel Santos & Francisco Amado (2020): *De novo* sequencing of proteins by mass spectrometry, Expert Review of Proteomics, DOI: 10.1080/14789450.2020.1831387

To link to this article: https://doi.org/10.1080/14789450.2020.1831387

Accepted author version posted online: 05 Oct 2020.

Publisher: Taylor & Francis & Informa UK Limited, trading as Taylor & Francis Group

Journal: *Expert Review of Proteomics*

DOI: 10.1080/14789450.2020.1831387

De novo sequencing of proteins by mass spectrometry

Rui Vitorino[1,2,3] , Sofia Guedes[1], Fabio Trindade[3], Inês Correia[2], Gabriela Moura[2], Paulo Carvalho[4], Manuel Santos[2], Francisco Amado[1]

[1] QOPNA & LAQV-REQUIMTE, Departamento de Química, Universidade de Aveiro, Aveiro, Portugal

[2] iBiMED, Department of Medical Sciences, University of Aveiro, Aveiro, Portugal

[3] Unidade de Investigação Cardiovascular, Departamento de Cirurgia e Fisiologia, Faculdade de Medicina, Universidade do Porto, Porto, Portugal

Correspondence:

Rui Vitorino

QOPNA & LAQV-REQUIMTE, Departamento de Química, Universidade de Aveiro, Aveiro, Portugal

Email:

Abstract

Introduction

Proteins are crucial for every cellular activity and unravelling their sequence and structure is a crucial step to fully understand their biology. Early methods of protein sequencing were mainly based on the use of enzymatic or chemical degradation of peptide chains. With the completion of the human genome project and with the expansion of the information available for each protein, various databases containing this sequence information were formed.

Areas covered

De novo protein sequencing, shotgun proteomics and other mass-spectrometric techniques, along with the various software are currently available for proteogenomic analysis. Emphasis is placed on the methods for *de novo* sequencing, together with potential and shortcomings using databases for interpretation of protein sequence data.

Expert opinion:

As mass-spectrometry sequencing performance is improving with better software and hardware optimizations, combined with user-friendly interfaces, *de-novo* protein sequencing becomes imperative in shotgun proteomic studies. Issues regarding unknown or mutated peptide sequences, as well as, unexpected post-translational modifications (PTMs) and their identification through false discovery rate searches using the target/decoy strategy need to be addressed. Ideally, it should become integrated in standard proteomic workflows as an add-on to conventional database search engines, which then would be able to provide improved identification coverage at controlled false discovery rates.

Keywords: algorithms, database search, de novo, mass spectrometry, MS/MS, proteomics, sequence tags, sequencing, tandem MS

Article highlights

- Peptide sequencing is an essential tool in biomedical and pharmaceutical research, among other fields.
- The introduction of search algorithms and the development of protein databases has enhanced the quality of *de novo* sequencing.
- The review discusses the advantages and limitations of traditional *de novo* sequencing methods, and the improvements brought about by the introduction of modern technologies.
- The potential role of proteogenomics in improving the applicability of traditional sequencing methods has also been discussed.

1. Introduction

Advancements in the field of medicine are a result of the spirit of inquiry of researchers toward human physiology and various associated biological processes. At the molecular level, proteins are one of the four important macromolecules necessary for critical physiological functions, and they are constituted by unique sequences of amino acids [1]. A thorough understanding of protein biology is crucial in biomedical sciences. When scientists discovered that amino acids are the building blocks of proteins, they were unaware of the precise arrangement of amino acids that forms the basis of protein structure [2]. Researchers have since attempted to determine protein sequence and structure. Molecular biology methods have enabled researchers to study biological molecules in detail. The initial methods used for protein sequence identification were manual and labor-intensive. The advent of high-throughput sequencing techniques gave rise to the multiple fields of omics [3,4]. Sequence data generated were deposited in several databases, which were subsequently converted into repositories containing detailed information on the sequences and the molecular features of biological molecules. At present, these databases are an indispensable resource for bioinformaticians, who employ *in-silico* methods to predict the function or behaviour of unknown molecules. However, before the development of sequencers and mass spectrometers, researchers mostly depended on *de novo* sequencing for sequence and structure prediction [5]. Currently, despite the growing availability of high-resolution mass spectrometers and the development of high-throughput software, *de novo* sequencing remains a mandatory process for the identification of new post-translational modifications (PTMs), annotation of translational errors, or identification without database searching owing to the absence of reference sequences. Herein, we have discussed the methods used for *de novo* sequencing, the tools and software associated with it, the role of protein sequence databases in the analysis of mass spectrometry data, and the advantages and disadvantages of peptide-based approaches. Through this review, we aim to elucidate the current status of the applicability of peptide-based sequencing approaches, and the advancements required to ensure that these techniques can be applied with high accuracy.

2. Shotgun proteomics

Shotgun proteomics is likely the strategy of mass spectrometry (MS) used most frequently for *de novo* sequencing. Shotgun proteomics is a bottom-up protein analysis approach, in which proteins are subjected to proteolytic digestion and the peptides generated are subjected to liquid chromatography-mass spectrometry (LC-MS) analysis [6][7]. While it was earlier dependent only on gel-free chromatography techniques for separation, such as strong cation exchange or reverse phase chromatography, techniques such as sodium dodecyl sulfate-polyacrylamide gel electrophoresis and isoelectric focusing are presently used for separation of proteins in shotgun proteomics studies [8]. Peptides are identified by comparing the experimental spectra against those generated from an existing sequence database [6]. The false discovery rate (FDR) is often controlled by including sequence decoys in the database, which are generally acquired by reversing the target sequences and using these decoy identification scores to model probabilistic functions [9]. Comparison to theoretical mass spectra was observed to be a rapid and less error-prone method compared to the method of using sequence interferences without comparison to sequences from a database; therefore, the former has been widely adopted as the typical approach for shotgun proteomics.

Shotgun proteomics has become a relatively high-throughput technique, provided the steps 2D gel-based separation and proteolytic in gel-digestion can be omitted by working directly with sample solutions [10]. Shotgun proteomics is popularly used for proteome profiling, protein quantification, and analysis of protein modifications and protein-protein interactions [8]. For instance, the proteome of soybean root hair cells was analysed by shotgun proteomics using 1D-PAGE-LC and multidimensional protein identification technology (MudPIT) [11,12]. In another study, the genes involved in light regulation in maize seedling leaves were studied using label-free quantitative proteomics, which led to the identification of several important proteins involved in the process [13]. In yet another study, the anti-inflammatory effects of rCC16 (Clara cell protein) were assessed using shotgun proteomics, where 12 proteins were identified, and their functional roles were studied using bioinformatics network analysis [14]. A shotgun proteomics workflow was designed where affinity chromatography was used for the separation of affinity-purified proteins (IPs), and proteins were identified in a single LC-MS/MS run on a Linear Trap Quadrupole (LTQ) Orbitrap instrument [15]. This method ensured the removal of background noise and facilitated the identification of unique protein interactions. The development of such workflows is crucial for the effective use of shotgun proteomics in protein identification and protein interactions studies.

The success of shotgun proteomics methods was based on accurate database search. Several databases such as UniProt/Swiss-Prot, UniProt/TrEMBL, RefSeq, Ensemble, IPI, and Entrez Protein are used in shotgun proteomic analysis [10]. Data interpretation and validation of the identified sequences remain a challenge in high-throughput techniques, and the use of these methods requires the development of large databases [10]. Further, the process of digestion of proteins to peptides without prior separation makes it challenging to match peptides with their respective proteins in complex samples containing multiple homologous proteins, such as those from higher eukaryotes. Therefore, shotgun proteomics has limited applicability in *de novo* sequencing of samples with proteins encoded by paralogous genes and with high sequence redundancy. Another major problem with this approach is its inability to correctly identify amino acids with identical masses, such as Ile and Leu, when low accuracy mass instruments, such as ion traps, are used. However, this problem can be solved using high-accuracy instruments. Although considered obsolete, MS-based sequencing coupled with enhanced fragmentation techniques, high-resolution spectrometers, and enhanced computational speed led to the establishment of spectral sequencing as a key approach in the era of modern proteomics [16,17].

3. Matrix-assisted laser desorption ionization (MALDI) vs electrospray ionization (ESI)

For a successful MS experiment, the molecules in the sample first need to be charged to separate them based on the respective mass/charge (m/z) ratio in the mass analyzer [18]. Initial experiments of *de novo* sequencing were considerably limited by harsh ionization methods, which may destroy several peptides. Hence, the development of soft ionization techniques, such as ESI and MALDI, was a major breakthrough in this field [19].

ESI is used to analyze samples in solution [20]. The sample solution is transferred to a gas phase via three sequential steps, beginning with the dispersal of a fine spray of charged droplets, followed by the evaporation of the solvent, and the ejection of ions from these charged droplets under a strong electric field [18,21]. The sample is passed through a high potential (2.5-6 kV) needle, which results in the formation of a spray of charged droplets by nebulization. Following this, the droplets shrink in size under high temperature and upon passing through a drying gas (nitrogen). During this process, the surface charge density on the droplets keeps increasing till it reaches a point that is energetically favourable for the ejection of droplets into the gaseous phase [18]. The emitted ions are then analysed using a mass analyzer and can be further fragmented and analysed using tandem-in-time MS in case of instruments such as Orbitrap or FTICR, or in a second mass analyzer using tandem-in-space MS with instruments such as Q-TOF to obtain detailed structural information (ESI-MS/MS). ESI allows the analysis of ionic species with increased sensitivity and can also be used to evaluate neutral compounds by converting them to ionic species using protonation or cationization. This technique is a sensitive and robust method that allows the analysis of thermolabile and non-volatile molecules as well [21]. The combination of ESI with high-performance liquid chromatography (HPLC) allows the analysis of low or high molecular weight molecules having negative, positive, or neutral charge [18]. The nanospray technology also revolutionized this technique, as it facilitated the analysis of biological molecules with minimal sample quantities at a considerably low flow rate with a range of nanolitres per minute. [21]

MALDI is another popular soft ionization technique introduced in 1988 [22,23]. In this technique, samples are first mixed with a suitable solvent and an organic, energy-absorbing compound known as the matrix. The sample then co-crystallizes with the matrix as the solvent dries [21,24]. This is followed by laser-induced desorption and ionization of the analytes present in the sample. These charged ions are then accelerated at a fixed potential, and are separated later on the basis of their m/z ratio in the mass analyzer (e.g., time of flight (TOF) analyzer), and are eventually detected by the detector [25]. Initially, MALDI required vacuum conditions for its overall operation; however, it was later modified to operate under atmospheric pressure with respect to the sample holder, which facilitated the simultaneous use of ESI and MALDI in mass spectrometers. The accuracy and sensitivity of MALDI depends on the choice of matrix used, which is influenced by the nature of the analyte and the charge imparted during ionization [21]. MALDI is also useful for studying biomolecules such as DNA, lipids, and glycoconjugates [21].

Both these methods are commonly used. Yet, there are advantages and limitations associated with their use. Samples with low concentration, such as in the picomolar range, can be analysed using both ESI and MALDI [21]. ESI offers high instrument flexibility owing to the use of sample solvation, and exhibits compatibility with various mass analyzers. Advancements in both these techniques have widened the scope of these techniques in some way; for example, in case of ESI, the use of nanospray ionization has significantly improved the sensitivity of the instrument, whereas in case of MALDI, the recent addition of imaging applications allows the user to study spatial large-scale proteomics. MALDI also allows sample reanalysis [26]. In a study performed using human pancreatic cells for understanding the differences between and the limitations of ESI and MALDI, a GeLC-MS workflow was used. MALDI is primarily dependent on the gas-phase basicity of the analyte [26], and the peptides ions are mostly singly charged, whereas ESI also relies on the hydrophobicity of the molecule [37], with the formation of multiply charged ions [27], which results in the biased identification of peptides [36]. Notwithstanding, the ESI technique proved to be better for detection of modifications, whereas MALDI was observed to have limited applicability owing to time constraints [26]. Additionally, it was observed that both techniques could be used to identify only a low number of peptides, and the limited detection of cysteine-, tryptophan- and methionine-containing peptides was demonstrated specifically. Eventually, the use of both techniques was proposed for robust and accurate prediction of peptide sequences. With respect to peptide quantification, measurements are more difficult to perform using MALDI owing to heterogeneous sample crystallization [21], while ESI is more susceptible to the presence of contaminants in the sample.

4. *De novo* sequencing

In the postgenomic era, *in-silico* protein sequencing strategies using MS-based computational tools can be made feasible using protein sequence databases. However, *de novo* peptide sequencing is still the preferred method for identification of novel proteins and peptide sequences involved in drug design [28]. *De novo* peptide sequencing is also useful for studying novel proteoforms generated as a result of mutations or PTMs. The multiple approaches for *de novo* sequencing include Edman degradation, MS, and ladder sequencing. While Edman degradation is time-intensive, MS-based techniques are economical and efficient [28]. Essentially, there are two approaches for protein identification by MS, namely, the top-down and bottom-up proteomics approaches. In the bottom-up approach, proteins are first digested and then identified using MS analysis, whereas in top-down approaches, intact proteins are analysed [29]. In the former, the sample complexity increases due to proteolytic digestion. The peptides are digested using enzymes (for instance, trypsin) directly in solution or after separation by gel electrophoresis methods. Afterwards, the digested proteins are subjected to ionization before MS analysis. Bottom-up approaches allow the quantification of proteins and provide information on the location of PTMs [29]. In the latter, protein sequence information including PTMs, truncations, and variations in the sequence are preserved depending on the fragmentation used in the mass analyzer [30]. Since the sequencing of labile PTMs presents a challenge in terms of identification and sequence localization, ETD and high-energy collision dissociation (HCD) ensure better performance in protein PTMs analysis. MS-based approaches were first used to sequence peptides such as glutaredoxin or calcitonin and parathyroid hormone using collision induced decomposition (CID) or Fast Atom Bombardment, where overlapping peptides were generated and identified based on the MS spectra generated [31,32]. These initial attempts at *de novo* protein sequencing were followed by the use of shotgun proteomics techniques. Whole genome sequence assemblies were also used for protein sequencing based on homology search. Various strategies have been adopted for *de novo* sequencing using combinations of bottom-up and top-down approaches. The collected MS/MS spectra work as a bar plot, where each fragment ion forms a peak corresponding to its respective m/z ratio [33]. The MS/MS spectra consist of b-ions corresponding to N-terminal peptide fragments and y-ions corresponding to C-terminal peptide fragments [34]. In *de novo* sequencing, the peaks generated during the peptide sequencing process are compared against the reference peaks for the 20 standard amino acids. The difference in the corresponding masses of two consecutive peaks yields the mass of a single amino acid, and this can, in principle, lead to the determination of the peptide sequence. The mass spectra also show modifications that occur in the peptide due to inherent changes or variations during sample processing [33]. Even though the sequence can be predicted using MS/MS spectra, some of the peaks may be missing; therefore, the use of specific algorithms was proposed for robust and rapid sequence prediction. PAAS was one of the first algorithms developed for sequence determination from fragmented peptides [35]. This algorithm generates all possible combinations of amino acids, the masses of which could add up to the required peptide mass, and compares it with theoretical spectra, thereby making the process slow and computationally expensive [5]. Another method used for sequencing is based on prefix pruning; however, this method has limited applicability in cases in which the prefixes are poorly represented [36]. A third approach is based on sub-sequencing, where short sequences from one of the terminals are tested against the fragmented ions, following which the best match is considered [5]. More recently, a method of computer-assisted manual interpretation was developed. In this method, the mass difference between the fragmented ions are analysed on the graph to establish a link between them. The most common and popular algorithms are based on graph theory [33,34]. In this method, a spectral graph is generated for each MS/MS spectrum which has MS/MS peaks as the nodes, representing the masses of fragmented peptides. Edges are formed when the difference in nodes corresponds to the mass of an amino acid [33]. The peptide sequence is determined by connecting the edges and on the basis of the graph score. These graph-based methods can be used to accurately identify the correct sequence from a combination of all possible sequences. Other algorithms based on machine learning, dynamic programing, linear programing, and hidden markov models (HMM) have also been used for *de novo* sequencing [33]. Various software based on different algorithms, such as SEQUEST, PeptideSearch, Lutefisk, Sherenga, PEAKS, and PepNovo, were developed for *de novo* sequencing of peptides [5].

5. Database-based interpretation of MS data

In the postgenomic era, the field of medical sciences was flooded with genomic information, which became an important tool for performing *in-silico* analysis. This extensive genomic information was stored collectively in a single platform in the form of a database, such as UniProt. The success of MS techniques is also dependent on the robustness and accuracy of databases [37]. LC-MS/MS is the most commonly used strategy for proteomic analysis, where LC is used for the removal of contaminants and the separation of compounds from the sample, and MS is used for calculating the m/z ratio. Tandem mass spectrometry (MS/MS) approaches for the identification of peptides and their sequences mostly depend on automated *de novo* sequencing, and manual sequencing methods are used only in case of unavailability of reference genome database or to study modifications unique to a system. The most common method for obtaining complete peptide sequences is based on database search either using experimental spectral libraries or peptide sequence databases [38,39]. In the first strategy (spectral libraries), existing spectra generated from actual peptide sequences are used as references for newly generated spectra [38]. This method omits the use of theoretical spectral data, and hence, allows analysis using data obtained through experiments. To achieve complete coverage and accuracy in protein sequencing, researchers need to generate several thousand spectra, which need to be analysed individually to identify the best match for a protein sequence [38,40]. The creation of spectral libraries with experimental data is an ongoing process, and therefore, at present, *in-silico* protein identification using peptide databases is the preferred technique.

Various databases and search algorithms were developed for easy and accurate identification of peptide sequences. A unique workflow was designed to identify pathogenic bacteria using tandem MS analysis [41]. Researchers constructed a proteomic database by automated analysis using the existing sequences of 87 bacterial genomes available in public repositories. MS/MS peptide-spectral matching was performed using the software SEQUEST [41]. ProteomeGenerator, which is another workflow, was developed based on mRNA sequencing data and proteogenomic peptide-spectral matching for proteomic analysis [42]. In this method, the pipeline assembles proteomes of actively transcribed genes using mRNA expression and identifies non-canonical peptides. An algorithm considers and compares the mass spectra generated by various software, such as SEQUEST HT, MaxQuant, Byonic, and PEAKS, and generates the best output. Lastly, complete computational environments for handling quantitative shotgun proteomics data should be considered as well; some examples of these are PatternLab for proteomics and TransProteomic Pipeline [43,44].

The database-based interpretation of protein sequences has certain drawbacks: i) data redundancy, which may prevent the identification of novel protein sequences, ii) the datasets, their matching score, and the accuracies of various databases are variable, and no single platform exists that can represent a comprehensive and robust database with uniform identifiers and file formats, iii) search algorithms and workflows for each database are different, and the final output of all databases are mostly represented in a probabilistic manner, iv) proteomic homology searches suffer from a bias of enzymatic digestion and degeneracy in the genetic code, which can prevent the accurate prediction of protein sequences from a fragmented peptide, v) the m/z values and the errors associated with them may result in variations in sequence interpretation [40].

5.1 Limitations of cross-species protein identification using MS-based sequence similarity searches

It is widely known that MS-based protein sequencing approaches are database-dependent, which is useful in cases where the whole genome sequence of an organism is available in the database. Although advanced genomic techniques and next-generation sequencing technology allow easy and economical retrieval of whole genome sequence information, the number of genomes available in public repositories is limited, considering the wide diversity of living organisms [45]. This poses a challenge in the identification of proteins in species for which sequence information is unavailable [27]. Minor genomic changes, such as single nucleotide variations, also limit cross-species identification owing to the limitations of databases. At times, minor changes in protein sequences may hinder their detection owing to changes in their chemistry. Furthermore, the occurrence of species-specific PTMs also prevents cross-species protein identification [27]. Early methods of cross-species protein identification were based on peptide mass fingerprinting (PMF), which is aimed at the determination of amino acid composition and estimation of the intact protein mass and pin values of unsequenced genomes [46]. The study showed that the combination of PMF with sequence composition data enhanced the identification of proteins. Later studies have showed that beyond PMF analysis, it is crucial to determine the phylogenetic distance between the query and reference species, such that the sequence identity is more than 80% [47]. These studies also demonstrated that the accuracy of sequence prediction was highly dependent on sequence identity between the query and reference genomes [27]. This led to the use of homology search methods against specific databases containing MS data. However, database accuracy also poses a problem in the use of these methods. While database search algorithms such as FASTA or BLAST are optimized to search queries comprising sequences with more than 35 amino acid residues, the peptide sequences obtained through tandem MS are approximately of 10-15 amino acids [45]. MS BLAST, which is a modified BLAST tool specifically used for peptide sequences generated via tandem MS, is used for identifying protein sequences in unsequenced genomes [48]. PepExplorer is an example of one of the latest developments within an integrated proteomic data analysis environment [49]. These tools can directly use the output of the tandem MS data from LC-MS/MS or MALDI-MS/MS experiments among others. However, this approach also has limited applicability in the characterization of complete proteomes. These search methods are dependent on the threshold values of identity, as well as on length, which will indicate the similarity percentage of proteins [45]. A detailed workflow based on the combination of PMF, tandem MS data, and *de novo* sequencing was developed for accurate cross-species protein identification [38,50].

5.2 PTMs

Extensively studied PTMs include phosphorylation, methylation, and acetylation, all of which play important roles in signaling events [51,52]. The analysis of these PTMs in the proteome has been facilitated through proteomic studies. UStag, a *de novo* sequencing based tool, was developed for the identification of PTMs by using Fourier transform tandem MS data of yeast [53]. The sequences were filtered and compared to UStags, and noisy sequences were excluded to obtain the final list of sequences with PTMs [53]. Although known PTMs can be identified using western blot, MS is the only option to identify novel or undetected PTMs and determine its location within the sequence, as it can detect mass changes occurring in the protein owing to any kind of modification [29]. However, bottom-up proteomics approaches for detection of PTMs cannot ensure the robust identification of modified amino acids against their unmodified counterparts. Different approaches, such as affinity-based enrichment of modified peptides before MS analysis, have also been used to detect modifications beforehand [29]. These methods also have constraints, such as limited sequence coverage and loss of connectivity. Conversely, top-down approaches are useful as they analyze intact proteins with all known modifications; nevertheless, spectral interpretation is considerably more difficult. Therefore, MS-based techniques are used frequently for the quantification and identification of PTMs; however this process is not fully comprehensive, as the search algorithms used in

MS/MS analysis are unable to indicate the significant diversity in these PTMs. Besides, this process is also highly time-consuming [54]. To overcome the limitations of incorrect assignment of PTMs, and to reduce the time spent in the process, a computational tool named TagGraph was developed to search MS/MS datasets without any restrictions on protein number, PTMs, or protease specificities. TagGraph uses an optimized probabilistic model for the identification of PTMs. This tool has expanded the knowledge on PTM modifications and has facilitated the rapid characterization of PTMs [54].

6. Commercial vs non- commercial software

There are various tools and software that assist researchers in the *de novo* sequencing of peptides. Most of these tools are freely available, whereas some are available under commercial licenses (Table 1). A peptide sequencing program based on the identification of positive ion peptides generated as a result of FAB was developed as early as 1986 [55]. **PAAS 3** was one of the first programs to allow the MS-based identification of peptides and determination of their sequence information [35]. The tool is freely available. The program first generates all possible combinations of amino acids for a particular sequence, which are matched and searched against the reference spectrum. **Lutefisk** was one of the first *de novo* sequencing algorithms, which scans protein databases based on tandem mass spectra of trypsin-digested peptides [56]. Using the database, the algorithm applies graph theory and identifies several fragmented peptides that act as the query for a subsequent homology-based search. **Mascot** and **SEQUEST** are the two most widely used algorithms for MS/MS data interpretation. While SEQUEST is more commonly associated with ion trap MS/MS analyzers, Mascot is more often used in TOF analyzers [57]. Mascot uses a probabilistic model to assess the chances that a particular fragment is associated with the observed spectrum, while SEQUEST scores the observed and predicted spectra using correlation measures [58].

The use of spectral graphs for sequence identification takes into account the best match; however, this match may not necessarily represent the actual sequence. Therefore, a suboptimal algorithm considering tandem mass spectra as a matrix was developed [59].

AUDENS was developed for automated *de novo* peptide sequencing using MS/MS data [60]. The tool uses a dynamic programming algorithm and distinguishes between actual peaks and real signals in the spectrum. This algorithm assigns values to each of the observed peaks and constructs a sequence path using the best scored peaks to reveal the complete sequence. The algorithm is freely available as an open source [60]. An HMM-based algorithm developed for *de novo* peptide sequencing using Bayesian framework was named **NovoHMM [61]**. The algorithm uses a graphical model and factorial HMM for improved peptide identification. This is also freely available at present. **PepNovo** was one of the first scoring methods developed for *de novo* sequencing using tandem MS data [62]. The scoring method is based on a probabilistic network, where the probability of occurrence of a peak in mass spectrum is evaluated as a chance or actual event. This is a freely available software. Another algorithm using MS/MS data is **MSNovo**, which is freely available and is compatible with LCQ and LTQ spectrometers [63]. This algorithm can accurately predict peptide sequences, as well as sequence tags, besides handling high-resolution data. **Vonode**, a *de novo* sequencing algorithm, was developed for the analysis of high-resolution tandem mass spectra [64]. This program is free to use and has shown better performance than PepNovo v2.0. To overcome the shortcomings of using specific algorithms unique for CID, HCD, or ETD (electron transfer dissociation) spectra, a universal tool with the ability to use MS/MS data from all spectra was developed [65]. This tool was named **UniNovo** and is freely available. The tool uses an improved scoring function based on a probabilistic module. A neural model based-peptide sequencing algorithm was developed and named **DeepNovo** [66]. This algorithm uses convolutional and recurrent neural networks to characterize MS/MS spectra in detail. These networks are then integrated using dynamic programming to perform *de novo* sequencing. This algorithm was tested using a wide range of data and was found to achieve up to 99.5% accuracy [66]. The software is freely available. A *de novo* peptide sequencing approach extending beyond an m/z ratio of 1600 was developed utilizing TOF-TOF data [67]. This method, known as **LIPCUT** (length incremented peptide composition lookup table), performs exhaustive analysis of MS/MS spectra using single molecule decomposition, where the peaks generated in a MALDI TOF/TOF spectra are search against a reference table with detailed compositional information on amino acids combinations, which are iterated in each step [67]. **PEAKS** is a *de novo* sequencing software that provides amino acid information by using MS/MS spectrum without referring to existing information from databases [68]. This algorithm provides the peptide sequences that match best to the peaks in the spectrum, and the output shows the confidence scores of amino acids and detailed sequence information. PEAKS performs classic *de novo* peptide sequencing, in which an amino acid sequence is derived from a mass spectrum without the referring to a sequence database. This is in contrast to the method use for “database search”, which is another popular peptide identification approach, which searches a given database to identify the largest peptide. *De* *novo* peptide sequencing is the only available option when sequence databases are unavailable. This makes PEAKS the preferred method for identifying novel peptides and proteins from organisms with unsequenced genomes. The ability to combine *de novo* peptide sequencing results with those of a database search is unique to PEAKS. *De novo* peptide sequences are aligned with protein database entries to provide additional information about PTMs, mutations, homologous peptides, and novel peptides. This software is one the earliest commercial software used in *de novo* peptide sequencing and is updated regularly [68].

Other software programs, such as Probed, EigenMS, PFIA, MAARIAN, and SeqMS, which are based on unique algorithms, have also been developed for *de novo* sequencing [69-71]. Performance evaluation of five *de novo* sequencing algorithms revealed that these algorithms failed to achieve more than 50% identification of peptide sequences using both QSTAR and LCQ datasets, which indicates the need to expand existing spectral data sets for enhancing the performance of these algorithms [72]. A web-based tool **DeNovoID** was also developed for *de novo* peptide sequencing, which uses degenerate amino acid sequences and MS-based mass data to perform search in a peptide database [73]. The hallmark of this algorithm is that it is independent of the protein sequence and is only dependent on the composition of amino acids. This algorithm is useful in cases where the spectra do not generate high-quality matches. **Novor**, a real-time peptide sequencing software, was developed to sequence novel peptides from MS/MS data [74]. The tool uses a machine learning approach to enhance the efficiency and speed of sequencing. The tool is more rapid than any existing software, and its speed surpasses that of MS, which facilitates robust protein sequence identification by analysis of 300 MS/MS spectra within a second [74].

7. Advantages of peptide-based approaches

Although the exact number of human protein products remains unknown to date, over 20,000 protein-coding genes exist in the genome [75]. Given the high number of possible variations at transcriptional and translational levels, more than 100,000 encoding proteins can be expected to exist [76,77]. In addition, a primary source of protein complexity is ubiquitously represented in PTMs, with more than 200 PTMs known [76]. As a result, genes act as precursors for a variety of structurally variant products, and even small structural changes can alter protein function [77].

In bottom-up experiments, all types of protein digestions commence with backbone cleavage at specific sites, which yields more complex peptide mixtures [76]. The products released by the lysis step are easily identified by MS. In particular, LC can separate peptides better than proteins. Most proteins will produce soluble peptides (even if the parent protein itself is insoluble), under conditions favorable for ionization [78]. Peptides are more efficiently segmented in the mass spectrometer when linked sequentially, which produces serial mass spectra [76]. Peptides are then separated using LC and analysed using MS/MS. The MS/MS spectra obtained in each experiment are scanned against a database of simulated MS/MS spectra generated from the computerized digestion of directly inserted proteins or proteins sequenced from DNA [78]. A score is assigned for each match between an experimental spectrum and a theoretical spectrum, and the peptide sequence with the best score (above a predetermined threshold) in the database is usually considered valid [78]. Generally, if the threshold is not reached, then no sequence assignment is made. Once the experimental peptide sequence is set alongside the theoretical peptide sequences [79], a database of known proteins is scanned to identify the proteins that may contain the peptides. This is a powerful strategy and there are only a limited number of alternatives. The process is time-consuming, expensive, and complex to the extent of being impractical, if not impossible to perform [78]. Likewise, the *de novo* interpretation of the ion spectral part or sequencing using sequence markers will be incompatible with the several thousand ion spectra produced per hour by the modern mass spectrometer [76,79]. Therefore, the combination of peptide-centered proteins with an automated sequencing database provides a practical option that can help identify 1000-2000 proteins in biological samples, and if proteins and peptides are separated, up to 4000-8000 proteins can be identified in each sample and sorted comprehensively [80]. However, this method has inherent limitations with respect to the loss of intact protein information, and the collective aspects of protein modification cannot be addressed [78].

Peptide-centered strategies can also be used to quantify individual proteins in a mixture [81]. Stable peer-marked peptides can be added to the sample mix to quantify specific proteins; this facilitates relative and absolute quantification with high accuracy [82]. Alternatively, tag-free methods such as spectral count and ion current measurements can provide less accurate yet differential (or comparative) estimates of peptide levels [81,82].

Usually, numerous MS/MS spectra are generated during protein-centered peptide analysis for a single sample, and these spectra are automatically matched with computer-generated trypsin peptides in relevant databases [81,82]. High-quality spectra of unmodified peptides are often matched without interference from other precursor ions; however, these spectra remain mismatched and unallocated [82]. Improvements in the instrument, especially for the measurement of high-precision precursors, can increase the proportion of assigned spectrum. However, the MS/MS spectrum may not be recognized for the following reasons [83]: i) the mass spectrum is generally highly populated, and there is insufficient information regarding the particle ions; ii) the segmented precursors are not peptides [82,83]; the peptides are modified in a manner that the search algorithm cannot detect them [84]; iii) the peptides are not present in the searched database; iv) multiple precursor ions can be detected in a specific precursor window and split simultaneously to generate a complex spectrum [84]. Most importantly, spectra derived from dipeptides, or spectra that contain residues modified by processes such as oxidation, reduction, nitration, and phosphorylation often remain unmatched [82-84]. Unrecognized modifications are of particular importance and occur for a variety of reasons [82]. Although researchers can choose to include modifications in their search strategies, in most cases, several modifications are not considered after translation [81], because once the statistical criteria are applied, the required search space increases significantly along with the search time, in addition to the decrease in the number of valid IDs. Besides, the physical properties of the modified residue (its quality or ionizing efficiency) may hinder its detection [83]. However, if the focus is on defining specific modifications in pure proteins, the peptide-centered approach [76] will offer certain advantages, because the effect of modification on quality is more pronounced at the peptide level than at the protein level [82]. It is difficult to detect and quantify minor changes in multiple sites at the protein level; however, it is easier and more accurate to determine the nature and number in peptide identification [83].

7.1. Incomplete databases

Another basic assumption is that the selected protein database is complete and contains all protein structures and variables in the target sample [85,86]. Such databases are rare (if at all present). Several variables are not described or documented. In addition, there are several restricted databases, and each database has defects and errors that affect the search results, and there is no consensus on the ideal database to be used or the method for selecting or modifying the minimum requirements [84,86].

It is clear that the matching strategy may only be a robust as the database [87]. For example, if the genome and proteins of an organism are not well defined, even high-quality spectra derived for the organism will not have an accurate match [87]. The search tool always provides the "best match" between experimental and stored data [84,86]. However, the challenge is to objectively assess the quality and accuracy of the output result. There are tools for performing this, and alternative methods are being developed as well [87]. Specifically, a mixed statistical model is usually used to distinguish between real matching and fake peptides from a false-positive peptide spectrum [86], and this model combines a set of factors in a single degree of distinction or statistics using decoy strategies [84,87]. Problems can occur even with a large population database [86]. If only a single peptide entry is appropriate for experimental data, there is no guarantee for correct designation, and when multiple database entries are similar or nearly similar to the experimental data, the selection process becomes subjective [87].

7. 2. Protein inference issues

A group of peptides can be common to several protein sequences. Therefore, regardless of the quality of the analytical work, it is often impossible to select the specific protein being studied [83]. As proteins are cleaved into peptides in the first step of peptide-centric analysis, there is no direct method to restore the link between peptides and their parent proteins [87]. As such, it is usually impossible to determine the number of proteins identified, as only the number of peptides is reported. To address such limitations, besides reporting all possible protein sequences, the best practice is to report a list of proteins according to maximum parsimony, that is, the minimum number of proteins that correspond to all identified peptides [88].

7. 3. Incomplete data obtained

It is important to note that only a part of the peptides that form the backbone of a complete protein are recovered, and these peptides are used only for protein identification [89]. If there are gaps in the sequence, "filling" of the lost amino acids can be performed by assuming that the lost amino acids are exactly the same as those specified in the database entry. For low-abundance proteins, fewer peptides and distributed ancestors are recovered [83,87]. Therefore, more inferences are needed. However, assuming that the modifications themselves can negatively affect their chances of being discovered, it is fair to assume there are no modifications or mutations in the mismatched regions [83,87]. For example, a one-point amino acid mutation within a trypsin peptide will prevent matching, and the presence of most subsequent translation modifications will prevent matching as well [83,87,89].

7. 4. Sample flow and collection

Peptide-centered analysis of a sample can require one hour or more, thereby reducing the analysis throughput [90]. As a result, research aimed at the identification of heterogeneous populations is often insignificant. Clusters or subsets of multiple samples could improve this; however, this would hide natural variance by averaging protein data [91].

8. Proteogenomics for the integration and identification of new peptide sequences/mutations

Proteogenomics is an amalgamation of the fields of proteomics and genomics that is used for the identification of novel peptides by using both MS-based proteomics data and sequenced genomic and transcriptomic data [92]. This approach can help by unifying sequence data with other detailed information of a protein. These proteins can also be unified in specific databases representing their functional roles. This technology assists in the collection of information on various peptides associated with gene mutations or genome reorganization. Proteogenomics enables improved annotation of novel genome sequences, such as in case of the backbone of *Bombyx mori*, a silkworm species used in the sericulture industry, or the lake trout *Salvelinus namaycush*, an indicator of environmental pollution, as well as in other organisms, such as fungi (*Sordaria macrospora*) or peanut (*Arachis hypogaea* L.) [93-96]. Improved techniques and novel approaches in proteogenomics can also help reduce data redundancy by facilitating the integration of large-scale genomic data to generate specific datasets [97]. Proteogenomic studies have important applications in cancer studies and have been widely used for the detection of cancer-related mutations, chromosomal hotspots, and cancer markers [98,99]. Various proteogenomic workflows and computational tools have also been designed to identify cancer-related prognostic markers and novel peptides [100-102]. The use of proteogenomics is also reported in immunopeptidome profiling for the detection of antigenic variations and identification of neoantigens [103-105]. Proteogenomic studies using LC-MS/MS or any other tandem MS method assist the identification of disease biomarkers that play a crucial role in diagnosis and prevention of disease [106-109]. The combination of proteogenomic approaches and *de novo* peptide sequencing is useful for the identification of novel genes and genomic features, including PTMs [95]. This advanced technology is a breakthrough in the field of characterization of human proteome variation [110].

9. Expert opinion

Mass spectrum sequencing is a technique that is completely independent of an external protein database resource. This unbiased approach is efficient and accurate, as the quality of spectra produced by high-accuracy and high-resolution mass spectrometers improves further. In principle, sequencing algorithms can be used to retrieve previously unidentified or mutated peptide sequences, as well as unexpected PTMs. This approach can be used complementarily with database searches in fully integrated environments. One of the primary issues that has rarely been discussed in literature is the evaluation of the quality of sequencing matches and the estimation of a reliable FDR, as performed for database searches using the target/decoy strategy. This can be especially challenging when evaluating matches containing sequence mutations. Owing to advancements in sequencing techniques, a wide range of genome sequences are being added to databases, the quality of which cannot be estimated using bioinformatics approaches alone. Modern computational power and the use of computer clusters allow the integration of mass spectrum sequencing into any proteomics workflow; however, achievement of complete accuracy in peptide identification cannot be guaranteed yet. As mass spectrum sequencing performance is improving with enhanced software and hardware optimizations, as well as with the introduction of user-friendly interfaces, the application of this promising technique will surely increase in shotgun proteomic studies. Ideally, the process should be integrated in standard proteomic workflows as an add-on to conventional database search engines, which could then be used to provide improved identification coverage at controlled FDRs. By relying on brief, pre-defined protein sequence databases, traditional research algorithms perform poorly when analysing mass spectra derived from completely unrecognized protein products [111]. In contrast, *de novo* peptide sequencing algorithms can interpret mass spectra without relying on a reference database [112]. However, because there is no method to automatically verify the results of *de novo* sequencing, it is difficult to apply these algorithms to complex protein mixtures [113].

Therefore, future indicators for evaluation of the performance of *de novo* sequence algorithms in the interpretation of large-scale protein datasets and methods for accurate calibration of the FDR will be important aspects of *de novo* sequencing applications [111]. Sequence identification has significantly improved with the latest optimization tools, as indicated by the high accuracy of Novor and PEAKS data compared to HCD and CID segment data [114]. Notably, Novor performs better in terms of sensitivity, and significantly reduces uptime [111]. Although there is no evaluation algorithm that accurately evaluates full-length peptide sequences from experimental data, it is still possible to obtain accurate values when matching short peptide markers [113].

Additionally, promising results on simulated data indicate that improvements in data quality may increase recognition [112]. Although algorithms show improved performance in the evaluation of high-resolution data, long-sequenced peptides (for example, those caused by skewed multiple divisions) continue to pose challenges that cannot be appropriately addressed [113]. These results indicate that there is an urgent need to improve algorithms that can make *de novo* sequencing a practical alternative to the common recognition workflow based on database inquires [115]. In this case, providing reliable references or baseline data will improve the assessment of current and newly developed algorithms [116]. In addition to the necessary improvements implemented by the bioinformatics community, MS tool vendors must improve their understanding of the true potential of this technology, which should form the basis for further development. Furthermore, the combination of proteogenomic approaches and *de novo* peptide sequencing will be useful for the identification of novel genes and genomic features, including PTMs and/or mutations. Indeed, proteogenomic strategies should be on data sets derived from deep, shotgun-like and top-down sampling of the transcriptome and proteome. Though both techniques can deeply sample fragments, a major drawback is the inability to know with certainty the sequence of the intact transcript or protein from which these fragments were derived. There are, however, still many challenges in proteogenomics including the limited coverage and dynamic range of MS based proteomics, and the difficulty of generating proteogenomic data sets with large numbers of samples because of cost and sample accessibility. The synergistic relationship between nucleotide sequencing and proteomics will continue to evolve and will be key for the complete characterization of the human proteome in the coming decades.

The advent of MS for protein identification, and its combination with *de novo* sequencing has revolutionized modern proteomics with the development of the nascent field of proteogenomics. The ability to identify novel peptides, their sequences, mutations, and modifications using these advanced techniques has broadened the understanding of molecular biology, particularly of proteins. Unravelling the genomic features of a protein is crucial for their characterization and understanding their functional role. There are various advancements in *de novo* sequencing of proteins, and new software are being developed continuously for robust and accurate identification. The Human Genome Project has paved the way for *in-silico* studies that will save time and optimize the use of resources. The limitations of such techniques are being studied to improve the usefulness of this approach in protein sequencing.

Funding

This work was supported by Portuguese Foundation for Science and Technology (FCT), European Union, QREN, FEDER and COMPETE for funding the Unidade de Investigação Cardiovascular (UID/IC/00051/2019), iBiMED (UIDB/04501/2020, POCI-01-0145-FEDER-007628) and the project LAQV-REQUIMTE (UIDB/50006/2020) research units and the research project R.V. are supported by individual fellowship grants (IF/00286/2015). Paulo Carvalho thanks Inova Fiocruz for financial support.

Declaration of interest

The authors have no relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript. This includes employment, consultancies, honoraria, stock ownership or options, expert testimony, grants or patents received or pending, or royalties.

Reviewer disclosures

Peer reviewers on this manuscript have no relevant financial or other relationships to disclose**.**

References

1. National Research Council (US) Committee on Research Opportunities in Biology. Opportunities in Biology. Washington (DC): National Academies Press (US); 1989. 3, Molecular Structure and Function. Available from: https://www.ncbi.nlm.nih.gov/books/NBK217812/.).
2. Peptide Sequencing by Edman Degradation. In: *eLS.*
3. Hoy MA. DNA Sequencing and the Evolution of the -Omic" (Ed.^(Eds) (2013)
4. James P. Protein identification in the post-genome era: the rapid rise of proteomics. *Quarterly reviews of biophysics*, 30(4), 279-331 (1997).
5. Ma B, Johnson R. <em>De Novo</em> Sequencing and Homology Searching. *Molecular &amp; Cellular Proteomics*, 11(2), O111.014902 (2012).
6. Zhang Y, Fonslow BR, Shan B, Baek MC, Yates JR, 3rd. Protein analysis by shotgun/bottom-up proteomics. *Chemical reviews*, 113(4), 2343-2394 (2013).
7. Sadygov RG, Cociorva D, Yates JR, 3rd. Large-scale database searching using tandem mass spectra: looking up the answer in the back of the book. *Nature methods*, 1(3), 195-202 (2004).
8. Matallana-Surget S, Leroy B, Wattiez R. Shotgun proteomics: concept, key points and data mining. *Expert review of proteomics*, 7(1), 5-7 (2010).
9. Elias JE, Gygi SP. Target-decoy search strategy for mass spectrometry-based proteomics. *Methods in molecular biology (Clifton, N.J.)*, 604, 55-71 (2010).
10. Nesvizhskii AI, Aebersold R. Interpretation of shotgun proteomic data: the protein inference problem. *Molecular & cellular proteomics : MCP*, 4(10), 1419-1440 (2005).
11. Brechenmacher L, Lee J, Sachdev S *et al.* Establishment of a protein reference map for soybean root hair cells. *Plant physiology*, 149(2), 670-682 (2009).
12. Wolters DA, Washburn MP, Yates JR, 3rd. An automated multidimensional protein identification technology for shotgun proteomics. *Analytical chemistry*, 73(23), 5683-5690 (2001).
13. Shen Z, Li P, Ni RJ *et al.* Label-free quantitative proteomics analysis of etiolated maize seedling leaves during greening. *Molecular & cellular proteomics : MCP*, 8(11), 2443-2460 (2009).

14. Pan C, Park BH, McDonald WH *et al.* A high-throughput de novo sequencing approach for shotgun proteomics using high-resolution tandem mass spectrometry. *BMC bioinformatics*, 11, 118 (2010).

15. Goto-Silva L, Maliga Z, Slabicki M, Murillo JR, Junqueira M. Application of shotgun proteomics for discovery-driven protein-protein interaction. *Methods in molecular biology (Clifton, N.J.)*, 1156, 265-278 (2014).

16. Mirzaei H, Carrasco MB. Modern Proteomics@ Sample Preparation, Analysis and Practical Applications. In: *Advances in experimental medicine and biology.* (Ed.^(Eds) (2016)

17. Johnson R, Searle BC, Nunn BL *et al.* Assessing protein sequence database suitability using de novo sequencing. *Molecular &amp; Cellular Proteomics*, mcp.TIR119.001752 (2019).

18. Ho CS, Lam CW, Chan MH *et al.* Electrospray ionisation mass spectrometry: principles and clinical applications. *The Clinical biochemist. Reviews*, 24(1), 3-12 (2003).

19. Alves G, Yu YK. Robust accurate identification of peptides (RAId): deciphering MS2 data using a structured library search with de novo based statistics. *Bioinformatics (Oxford, England)*, 21(19), 3726-3732 (2005).

20. Fenn J, Mann M, Meng C, Wong S, Whitehouse C. Electrospray ionization for mass spectrometry of large biomolecules. *Science*, 246(4926), 64-71 (1989).

21. El-Aneed A, Cohen A, Banoub J. Mass Spectrometry, Review of the Basics: Electrospray, MALDI, and Commonly Used Mass Analyzers. *Applied Spectroscopy Reviews*, 44(3), 210-230 (2009).

22. Hillenkamp F, Karas M, Beavis RC, Chait BT. Matrix-assisted laser desorption/ionization mass spectrometry of biopolymers. *Analytical chemistry*, 63(24), 1193a-1203a (1991).

23. Tanaka K, Waki H, Ido Y *et al.* Protein and polymer analyses up to m/z 100 000 by laser ionization time-of-flight mass spectrometry. *Rapid Communications in Mass Spectrometry*, 2(8), 151-153 (1988).

24. Matrix-Assisted Laser Desorption/Ionization Mass Spectrometry in Peptide and Protein Analysis. In: *Encyclopedia of Analytical Chemistry.*

25. Singhal N, Kumar M, Kanaujia PK, Virdi JS. MALDI-TOF mass spectrometry: an emerging technology for microbial identification and diagnosis. *Frontiers in microbiology*, 6, 791 (2015).

26. Nadler WM, Waidelich D, Kerner A *et al.* MALDI versus ESI: The Impact of the Ion Source on Peptide Identification. *Journal of proteome research*, 16(3), 1207-1215 (2017).

27. Hughes C, Ma B, Lajoie GA. De Novo Sequencing Methods in Proteomics. *Methods in molecular biology*, 604, 105-121 (2010).

28. Liu X, Dekker LJ, Wu S *et al.* De novo protein sequencing by combining top-down and bottom-up tandem mass spectra. *Journal of proteome research*, 13(7), 3241-3248 (2014).

29. Gregorich ZR, Chang YH, Ge Y. Proteomics in heart failure: top-down or bottom-up? *Pflugers Archiv : European journal of physiology*, 466(6), 1199-1209 (2014).

30. Quan L, Liu M. CID,ETD and HCD Fragmentation to Study Protein Post-Translational Modifications. *Modern Chemistry & Applications*, 1, 1-2 (2013).

31. Hopper S, Johnson RS, Vath JE, Biemann K. Glutaredoxin from rabbit bone marrow. Purification, characterization, and amino acid sequence determined by tandem mass spectrometry. *The Journal of biological chemistry*, 264(34), 20438-20447 (1989).

32. Whaley B, Caprioli RM. Identification of nearest-neighbor peptides in protease digests by mass spectrometry for construction of sequence-ordered tryptic maps. *Biological mass spectrometry*, 20(4), 210-214 (1991).

33. Muth T, Hartkopf F, Vaudel M, Renard BY. A Potential Golden Age to Come-Current Tools, Recent Use Cases, and Future Avenues for De Novo Sequencing in Proteomics. *Proteomics*, 18(18), e1700150 (2018).

34. Song Y. A new parameterized algorithm for rapid peptide sequencing. *PLoS One*, 9(2), e87476 (2014).

35. Sakurai T, Matsuo T, Matsuda H, Katakuse I. PAAS 3: A computer program to determine probable sequence of peptides from mass spectrometric data. *Biomedical Mass Spectrometry*, 11(8), 396-399 (1984).

36. Ma B, Zhang K, Liang C. An effective algorithm for peptide de novo sequencing from MS/MS spectra. *Journal of Computer and System Sciences*, 70(3), 418-430 (2005).

37. Addona T, Clauser K. De novo peptide sequencing via manual interpretation of MS/MS spectra. *Current protocols in protein science*, Chapter 16, Unit 16.11 (2002).

38. Kim S, Gupta N, Bandeira N, Pevzner PA. Spectral dictionaries: Integrating de novo peptide sequencing with database search of tandem mass spectra. *Molecular & cellular proteomics : MCP*, 8(1), 53-69 (2009).

39. Attila K-F, Beata R, Michael PM, Sandor P. Database Searching in Mass Spectrometry Based Proteomics. *Current Bioinformatics*, 7(2), 221-230 (2012).

40. Cottrell JS. Protein identification using MS/MS data. *Journal of proteomics*, 74(10), 1842-1851 (2011).

41. Dworzanski JP, Snyder AP, Chen R, Zhang H, Wishart D, Li L. Identification of bacteria using tandem mass spectrometry combined with a proteome database and statistical scoring. *Analytical chemistry*, 76(8), 2355-2366 (2004).

42. Cifani P, Dhabaria A, Chen Z *et al.* ProteomeGenerator: A Framework for Comprehensive Proteomics Based on de Novo Transcriptome Assembly and High-Accuracy Peptide Mass Spectral Matching. 17(11), 3681-3692 (2018).

43. Nott TJ, Petsalaki E, Farber P *et al.* Phase transition of a disordered nuage protein generates environmentally responsive membraneless organelles. *Mol Cell*, 57(5), 936-947 (2015).

44. Carvalho PC, Lima DB, Leprevost FV. Integrated analysis of shotgun proteomic data with PatternLab for proteomics 4.0. 11(1), 102-117 (2016).

45. 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. *Molecular & cellular proteomics : MCP*, 3(3), 238-249 (2004).

46. Cordwell SJ, Basseal DJ, Humphery-Smith I. Proteome analysis of Spiroplasma melliferum (A56) and protein characterisation across species boundaries. *Electrophoresis*, 18(8), 1335-1346 (1997).

47. Molloy MP, Phadke ND, Maddock JR, Andrews PC. Two-dimensional electrophoresis and peptide mass fingerprinting of bacterial outer membrane proteins. *Electrophoresis*, 22(9), 1686-1696 (2001).

48. Shevchenko A, Sunyaev S, Loboda A *et al.* Charting the proteomes of organisms with unsequenced genomes by MALDI-quadrupole time-of-flight mass spectrometry and BLAST homology searching. *Analytical chemistry*, 73(9), 1917-1926 (2001).

49. Leprevost FV, Valente RH, Lima DB *et al.* PepExplorer: a similarity-driven tool for analyzing de novo sequencing results. *Molecular & cellular proteomics : MCP*, 13(9), 2480-2489 (2014).

50. Ostrowski M, Fegatella F, Wasinger V, Guilhaus M, Corthals GL, Cavicchioli R. Cross-species identification of proteins from proteome profiles of the marine oligotrophic ultramicrobacterium, Sphingopyxis alaskensis. *Proteomics*, 4(6), 1779-1788 (2004).

51. Silva AMN, Vitorino R, Domingues MRM, Spickett CM, Domingues P. Post-translational modifications and mass spectrometry detection. *Free radical biology & medicine*, 65, 925-941 (2013).

52. Liu H, Duan Y. Effects of posttranslational modifications on the structure and dynamics of histone H3 N-terminal Peptide. *Biophysical journal*, 94(12), 4579-4585 (2008).

53. Shen Y, Tolić N, Hixson KK, Purvine SO, Anderson GA, Smith RD. De novo sequencing of unique sequence tags for discovery of post-translational modifications of proteins. *Analytical chemistry*, 80(20), 7742-7754 (2008).

54. Devabhaktuni A, Lin S, Zhang L, Swaminathan K. TagGraph reveals vast protein modification landscapes from large tandem mass spectrometry datasets. 37(4), 469-479 (2019).

55. Hamm CW, Wilson WE, Harvan DJ. Peptide sequencing program. *Bioinformatics (Oxford, England)*, 2(2), 115-118 (1986).

56. Taylor JA, Johnson RS. Sequence database searches via de novo peptide sequencing by tandem mass spectrometry. *Rapid Communications in Mass Spectrometry*, 11(9), 1067-1075 (1997).

57. Elias JE, Haas W, Faherty BK, Gygi SP. Comparative evaluation of mass spectrometry platforms used in large-scale proteomics investigations. *Nature methods*, 2(9), 667-675 (2005).

58. Eng JK, McCormack AL, Yates JR. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J Am Soc Mass Spectrom*, 5(11), 976-989 (1994).

59. Lu B, Chen T. A suboptimal algorithm for de novo peptide sequencing via tandem mass spectrometry. *Journal of computational biology : a journal of computational molecular cell biology*, 10(1), 1-12 (2003).

60. Grossmann J, Roos FF, Cieliebak M *et al.* AUDENS: a tool for automated peptide de novo sequencing. *Journal of proteome research*, 4(5), 1768-1774 (2005).

61. Fischer B, Roth V, Roos F *et al.* NovoHMM: a hidden Markov model for de novo peptide sequencing. *Analytical chemistry*, 77(22), 7265-7273 (2005).

62. Frank A, Pevzner P. PepNovo: de novo peptide sequencing via probabilistic network modeling. *Analytical chemistry*, 77(4), 964-973 (2005).

63. Mo L, Dutta D, Wan Y, Chen T. MSNovo: a dynamic programming algorithm for de novo peptide sequencing via tandem mass spectrometry. *Analytical chemistry*, 79(13), 4870-4878 (2007).

64. Pan C, Park BH, McDonald WH *et al.* A high-throughput de novo sequencing approach for shotgun proteomics using high-resolution tandem mass spectrometry. *BMC bioinformatics*, 11(1), 118 (2010).

65. Jeong K, Kim S, Pevzner PA. UniNovo: a universal tool for de novo peptide sequencing. *Bioinformatics (Oxford, England)*, 29(16), 1953-1962 (2013).

66. Tran NH, Zhang X, Xin L, Shan B, Li M. De novo peptide sequencing by deep learning. *Proceedings of the National Academy of Sciences*, 114(31), 8247-8252 (2017).

67. Olson MT, Epstein JA, Yergey AL. De novo peptide sequencing using exhaustive enumeration of peptide composition. *J Am Soc Mass Spectrom*, 17(8), 1041-1049 (2006).

68. Ma B, Zhang K, Hendrie C *et al.* PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. *Rapid communications in mass spectrometry : RCM*, 17(20), 2337-2342 (2003).

69. Zhang N, Aebersold R, Schwikowski B. ProbID: a probabilistic algorithm to identify peptides through sequence database searching using tandem mass spectral data. *Proteomics*, 2(10), 1406-1412 (2002).

70. Bern M, Goldberg D. De novo analysis of peptide tandem mass spectra by spectral graph partitioning. *Journal of computational biology : a journal of computational molecular cell biology*, 13(2), 364-378 (2006).

71. Jagannath S, Sabareesh V. Peptide Fragment Ion Analyser (PFIA): a simple and versatile tool for the interpretation of tandem mass spectrometric data and de novo sequencing of peptides. *Rapid communications in mass spectrometry : RCM*, 21(18), 3033-3038 (2007).

72. Pevtsov S, Fedulova I, Mirzaei H, Buck C, Zhang X. Performance evaluation of existing de novo sequencing algorithms. *Journal of proteome research*, 5(11), 3018-3028 (2006).

73. Halligan BD, Ruotti V, Twigger SN, Greene AS. DeNovoID: a web-based tool for identifying peptides from sequence and mass tags deduced from de novo peptide sequencing by mass spectroscopy. *Nucleic Acids Res*, 33(Web Server issue), W376-381 (2005).

74. Ma B. Novor: real-time peptide de novo sequencing software. *J Am Soc Mass Spectrom*, 26(11), 1885-1894 (2015).

75. Ning Z, Zhang X, Mayne J, Figeys D. Peptide-Centric Approaches Provide an Alternative Perspective To Re-Examine Quantitative Proteomic Data. *Analytical chemistry*, 88(4), 1973-1978 (2016).

76. Cristobal A, Marino F, Post H, van den Toorn HW, Mohammed S, Heck AJ. Toward an optimized workflow for middle-down proteomics. *Analytical chemistry*, 89(6), 3318-3325 (2017).

77. Zhang P, Culver-Cochran AE, Stevens Jr SM, Liu B. Characterization of a SILAC method for proteomic analysis of primary rat microglia. *Proteomics*, 16(9), 1341-1346 (2016).

78. Gao Y, Yates III JR. Protein Analysis by Shotgun Proteomics. *Mass Spectrometry-Based Chemical Proteomics*, 1-38 (2019).

79. Ludwig C, Gillet L, Rosenberger G, Amon S, Collins BC, Aebersold R. Data-independent acquisition-based SWATH-MS for quantitative proteomics: a tutorial. *Molecular systems biology*, 14(8) (2018).

80. Ting YS, Egertson JD, Payne SH *et al.* Peptide-Centric Proteome Analysis: An Alternative Strategy for the Analysis of Tandem Mass Spectrometry Data. *Molecular & cellular proteomics : MCP*, 14(9), 2301-2307 (2015).

81. Ning Z, Zhang X, Mayne J, Figeys D. Peptide-centric approaches provide an alternative perspective to re-examine quantitative proteomic data. *Analytical chemistry*, 88(4), 1973-1978 (2016).

82. Garcia Ln, Girod M, Rompais M, Dugourd P, Carapito C, Lemoine Jrm. Data-independent acquisition coupled to visible laser-induced dissociation at 473 nm (DIA-LID) for peptide-centric specific analysis of cysteine-containing peptide subset. *Analytical chemistry*, 90(6), 3928-3935 (2018).

83. Berg P, McConnell EW, Hicks LM, Popescu SC, Popescu GV. Evaluation of linear models and missing value imputation for the analysis of peptide-centric proteomics. *BMC bioinformatics*, 20(2), 102 (2019).

84. Lyu J, Wang K, Ye M. Modification-free approaches to screen drug targets at proteome level. *TrAC Trends in Analytical Chemistry*, (2019).

85. Yang H, Chi H, Zhou W-J *et al.* Open-pNovo: de novo peptide sequencing with thousands of protein modifications. *Journal of proteome research*, 16(2), 645-654 (2017).

86. Yang Y, Franc V, Heck AJ. Glycoproteomics: a balance between high-throughput and in-depth analysis. *Trends in biotechnology*, 35(7), 598-609 (2017).

87. Prieto G, Vázquez J. Calculation of False Discovery Rate for Peptide and Protein Identification. In: *Mass Spectrometry Data Analysis in Proteomics.* (Springer, 2020) 145-159.

88. Zhang B, Chambers MC, Tabb DL. Proteomic parsimony through bipartite graph analysis improves accuracy and transparency. *Journal of proteome research*, 6(9), 3549-3557 (2007).

89. Riffle M, May DH, Timmins-Schiffman E *et al.* MetaGOmics: a web-based tool for peptide-centric functional and taxonomic analysis of metaproteomics data. *Proteomes*, 6(1), 2 (2018).

90. Kim Y-I, Cho J-Y. Gel-based proteomics in disease research: Is it still valuable? *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics*, 1867(1), 9-16 (2019).

91. Bhosale SD, Moulder R, Kouvonen P, Lahesmaa R, Goodlett DR. Mass spectrometry-based serum proteomics for biomarker discovery and validation. In: *Serum/Plasma Proteomics.* (Springer, 2017) 451-466.

92. Jagtap PD, Johnson JE, Onsongo G *et al.* Flexible and accessible workflows for improved proteogenomic analysis using the Galaxy framework. *Journal of proteome research*, 13(12), 5898-5908 (2014).

93. Ye X, Tang X, Wang X *et al.* Improving Silkworm Genome Annotation Using a Proteogenomics Approach. 18(8), 3009-3019 (2019).

94. Dupree EJ, Crimmins BS, Holsen TM, Darie CC. Developing Well-Annotated Species-Specific Protein Databases Using Comparative Proteogenomics. *Advances in experimental medicine and biology*, 1140, 389-400 (2019).

95. Blank-Landeshammer B, Teichert I. Combination of Proteogenomics with Peptide De Novo Sequencing Identifies New Genes and Hidden Posttranscriptional Modifications. 10(5) (2019).

96. Li H, Zhou R, Xu S *et al.* Improving Gene Annotation of the Peanut Genome by Integrated Proteogenomics Workflow. *Journal of proteome research*, 19(6), 2226-2235 (2020).

97. Woo S, Cha SW, Merrihew G *et al.* Proteogenomic database construction driven from large scale RNA-seq data. *Journal of proteome research*, 13(1), 21-28 (2014).

98. Satpathy S, Jaehnig EJ. Microscaled proteogenomic methods for precision oncology. 11(1), 532 (2020).

99. Ma YS, Huang T, Zhong XM *et al.* Proteogenomic characterization and comprehensive integrative genomic analysis of human colorectal cancer liver metastasis. 17(1), 139 (2018).

100. Ni Y, Stingo FC, Ha MJ, Akbani R, Baladandayuthapani V. Bayesian Hierarchical Varying-sparsity Regression Models with Application to Cancer Proteogenomics. *Journal of the American Statistical Association*, 114(525), 48-60 (2019).

101. Zhu Y, Orre LM, Johansson HJ. Discovery of coding regions in the human genome by integrated proteogenomics analysis workflow. 9(1), 903 (2018).

102. Chakraborty S, Andrieux G, Hasan AMM, Ahmed M, Hosen MI. Harnessing the tissue and plasma lncRNA-peptidome to discover peptide-based cancer biomarkers. 9(1), 12322 (2019).

103. Freudenmann LK, Marcu A. Mapping the tumour human leukocyte antigen (HLA) ligandome by mass spectrometry. 154(3), 331-345 (2018).

104. Chong C, Müller M. Integrated proteogenomic deep sequencing and analytics accurately identify non-canonical peptides in tumor immunopeptidomes. 11(1), 1293 (2020).

105. Zhang X, Qi Y, Zhang Q, Liu W. Application of mass spectrometry-based MHC immunopeptidome profiling in neoantigen identification for tumor immunotherapy. *Biomedicine & pharmacotherapy = Biomedecine & pharmacotherapie*, 120, 109542 (2019).

106. Di Meo A, Batruch I, Brown MD *et al.* Identification of Prognostic Biomarkers in the Urinary Peptidome of the Small Renal Mass. *The American journal of pathology*, 189(12), 2366-2376 (2019).

107. Krochmal M, van Kessel KEM, Zwarthoff EC *et al.* Urinary peptide panel for prognostic assessment of bladder cancer relapse. *Scientific reports*, 9(1), 7635 (2019).

108. Huang CH, Kuo CJ, Liang SS *et al.* Onco-proteogenomics identifies urinary S100A9 and GRN as potential combinatorial biomarkers for early diagnosis of hepatocellular carcinoma. *BBA clinical*, 3, 205-213 (2015).

109. Chiou SH, Lee KT. Proteomic analysis and translational perspective of hepatocellular carcinoma: Identification of diagnostic protein biomarkers by an onco-proteogenomics approach. *The Kaohsiung journal of medical sciences*, 32(11), 535-544 (2016).

110. Sheynkman GM, Shortreed MR, Cesnik AJ, Smith LM. Proteogenomics: Integrating Next-Generation Sequencing and Mass Spectrometry to Characterize Human Proteomic Variation. *Annual review of analytical chemistry (Palo Alto, Calif.)*, 9(1), 521-545 (2016).

111. Cafarelli T, Desbuleux A, Wang Y, Choi SG, De Ridder D, Vidal M. Mapping, modeling, and characterization of protein–protein interactions on a proteomic scale. *Current Opinion in Structural Biology*, 44, 201-210 (2017).

112. Klasberg S, Bitard-Feildel T, Mallet L. Computational identification of novel genes: current and future perspectives. *Bioinformatics and Biology insights*, 10, BBI. S39950 (2016).

113. Muhammed MT, Aki-Yalcin E. Homology modeling in drug discovery: Overview, current applications, and future perspectives. *Chemical biology & drug design*, 93(1), 12-20 (2019).

114. Liu B, Yang J, Li Y, McDermaid A, Ma Q. An algorithmic perspective of de novo cis-regulatory motif finding based on ChIP-seq data. *Briefings in bioinformatics*, 19(5), 1069-1081 (2018).

115. Murphy GS, Greisman JB, Hecht MH. De novo proteins with life-sustaining functions are structurally dynamic. *Journal of molecular biology*, 428(2), 399-411 (2016).

116. Gautam R, Kaur P, Sharma M. A comprehensive review on nature inspired computing algorithms for the diagnosis of chronic disorders in human beings. *Progress in Artificial Intelligence*, 1-24 (2019).

117. Tabb DL, Saraf A, Yates JR, 3rd. GutenTag: high-throughput sequence tagging via an empirically derived fragmentation model. *Analytical chemistry*, 75(23), 6415-6421 (2003).

118. Johnson RS, Taylor JA. Searching sequence databases via de novo peptide sequencing by tandem mass spectrometry. *Molecular biotechnology*, 22(3), 301-315 (2002).

119. Mo L, Dutta D, Wan Y, Chen T. MSNovo:□ A Dynamic Programming Algorithm for de Novo Peptide Sequencing via Tandem Mass Spectrometry. *Analytical chemistry*, 79(13), 4870-4878 (2007).

120. Searle BC, Dasari S, Wilmarth PA *et al.* Identification of protein modifications using MS/MS de novo sequencing and the OpenSea alignment algorithm. *Journal of proteome research*, 4(2), 546-554 (2005).

121. Fernandez-de-Cossio J, Gonzalez J, Satomi Y *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*, 21(9), 1694-1699 (2000).

122. Dancík V, Addona TA, Clauser KR, Vath JE, Pevzner PA. De novo peptide sequencing via tandem mass spectrometry. *Journal of computational biology : a journal of computational molecular cell biology*, 6(3-4), 327-342 (1999).

123. Han Y, Ma B, Zhang K. SPIDER: software for protein identification from sequence tags with de novo sequencing error. *Journal of bioinformatics and computational biology*, 3(3), 697-716 (2005).

124. Li C, Li K, Li K, Xie X, Lin F. SWPepNovo: An Efficient De Novo Peptide Sequencing Tool for Large-scale MS/MS Spectra Analysis. *International journal of biological sciences*, 15(9), 1787-1801 (2019).

Tables

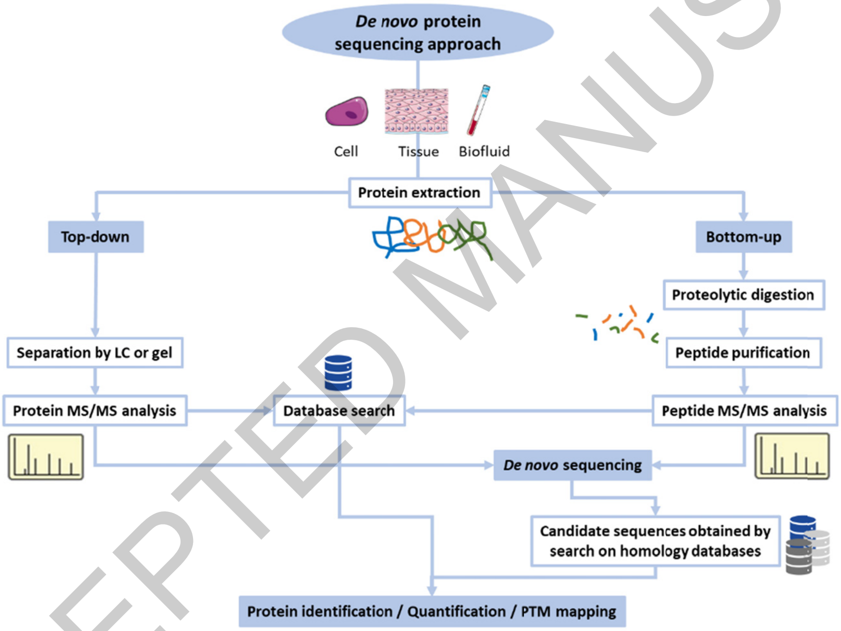
Table 1: Different software for *de novo* sequencing

Software name	**Algorithm**	**Score**	**Comment**	**References**
AUDENS	Spectrum graph	Internally calculated sum of peak relevance	Assigns relevance to peaks during preprocessing	[60]
EigenMS	Spectral graph partitioning	Mass fit, ion abundance, probability to observe ion	Usage of two graphs	[70]
GutenTag	Combination of de novo and database search algorithms	Scores sequence tag	Identifies proteins correctly considering modifications and post translational modifications	[117]
Lutefisk	Spectrum graph	Sum of b-ion probabilities during subsequence	Rescoring of prediction with several measures	[118]
MSNovo DP	mass array spectrum representation	Probabilistic distribution of mass tolerance	LCQ/LTQ Charges 1–3	[119]
NovoHMM	Hidden Markov model	Bayesian posterior probabilities for amino acids	Tested on 1252 spectra and compared with other algorithms	[61]
Novor	Spectrum	Decision tree model	Academic or non-	[74]

	graph		commercial license	
OpenSea	Heuristic algorithm	Partially correct sequence tags with a database to identify the homologous or modified proteins	Accurately locates sequence variation sites and unanticipated posttranslational modifications	[120]
PEAKS	Generation of 105 candidate sequences	Peak abundance, mass fit, fragment complementarity	Commercial software, algorithm not fully disclosed	[68]
PepNovo	Spectrum graph	Likelihood ratio hypothesis testing in respect to random model	Only a few learned models available	[62]
SeqMS	Spectrum graph	Ion abundance, fragment complementarity	Originally for HCD spectra, later adapted for low-energy	[121]
SHERENGA	Spectrum graph	Scoring is based on assigning a probability-based score, taking into account rewards/penalties for fragment ions that are present or missing.	The algorithm will use the highest scoring sequence path from the spectrum graph as the peptide sequence.	[122]
SPIDER	Dynamic programming	The *de novo* cost function, the homology score matrix, insertion/deletion cost	Free software	[123]
SWPepNovo	Spectrum graph	SW26010 many-core processor, namely SWPepNovo, to process the large-scale peptide MS/MS spectra using a	A two-level parallelization mechanism	[124]

		parallel peptide spectrum matches (PSMs) algorithm.		
Vonode	Spectrum graph	Based primarily on mass accuracy but also, in part, on fragment abundance	Dependent on high mass accuracy, also makes sequence tags	[14]

Figure 1: Schematic representation of *de novo* sequencing


De novo protein sequencing approach
Cell
Tissue
Biofluid
Protein extraction
Top-down
Bottom-up
Proteolytic digestion
Peptide purification
Separation by LC or gel
Protein MS/MS analysis
Database search
Peptide MS/MS analysis
De novo sequencing
Candidate sequences obtained by search on homology databases
Protein identification / Quantification / PTM mapping