Algorithms for the *De Novo* Sequencing of Peptides from Tandem Mass Spectra

Jens Allmer
Molecular Biology and Genetics, Izmir Institute of Technology, Urla, Izmir, 35430, Turkey
Phone: 00902327507517, Fax: 00902327507509, Email: jens@allmer.de

Abstract

Proteomics is the study of the proteins, their time and location dependent expression profiles as well as their modifications and interactions. Mass spectrometry is useful to investigate many of the questions asked in proteomics. Typically database search methods are employed to identify proteins from complex mixtures. However, often databases are not available or despite their availability some sequences are not readily found therein. To overcome this problem *de novo* sequencing can be used to directly assign a peptide sequence to an MS/MS spectrum. Many algorithms have been proposed for *de novo* sequencing and a selection of them is detailed in this review. Although a standard accuracy measure has not been agreed upon in the field, relative algorithm performance is discussed. The current state of the *de novo* sequencing is assessed thereafter and finally, examples are used to construct possible future perspectives of the field.

Running title: *De novo* sequencing of MS/MS spectra
Keywords: mass spectrometry, *de novo*, sequencing, tandem MS, MS/MS, sequence tags, database search, algorithms, proteomics

Introduction

Proteomics aims to investigate the complement of a genome, the proteome. The proteome is the entirety of proteins that can be expressed by a genome, including alternatively spliced and modified products. Today mass spectrometry (MS) is the tool of choice for protein identification and quantification [1]. For many forms of quantitation an initial identification of peptides and proteins is necessary; see for example [2]. Therefore, the correct assignment of an amino acid sequence to each tandem MS (MS/MS, MS^2) spectrum is crucial for many scenarios in mass spectrometry based proteomics. There are basically two methods to assign a sequence to MS/MS spectra. One widespread method is dependent on the availability of sequences in a database. Simplified, all sequences in the database can be scored against the spectrum and the best scoring sequence is then accepted as the precursor of the MS/MS spectrum. Sequest [3], Mascot [4], Omssa [5], and X!Tandem [6] are just a few examples of the many algorithms available in the field of database searching. The downside of database searching is that sequences are not always in the database [7]; even in the six frame translation of a genomic database they may not be present. Alternative splice forms, short proteins, wrong (even if just slightly)

gene predictions, and many other factors prevent the correct assignment of sequence to spectrum. This further depends on the complexity of the sample, the mass spectrometer used, and the settings for false discovery rate. Unfortunately, some tandem mass spectra may fit well to an unrelated sequence in the database and thus lead to false identifications. Another big problem is posed by post translational modifications (PTMs) of amino acids which shift the mass of the precursor and thus renders an assignment of sequence to spectrum impossible, unless the modification is anticipated. Peptidomics is an emerging field interested in bioactive peptides which, unlike peptides found in high-throughput studies, do not result from tryptic cleavage and are usually heavily modified and therefore may pose problems to database search algorithms [8]. Similar problems apply when the target sequences are antibodies [9], hormones, or toxins.

De novo sequencing, which aims to assign a sequence to MS/MS spectra without the need for a sequence database, is the other method used. Although there is no sequence database used in this approach, the search space can be represented in the same way (see naïve approach). The theoretical search space for *de novo* sequencing algorithms encompasses the entirety of amino acid sequences that can account for the precursor mass, within a certain mass tolerance, dictated by the mass spectrometer. The complete search space can only be investigated for very small peptides and with increasing length quickly poses a problem which cannot be solved in a timely manner, if at all. Many algorithms have been developed trying to solve the *de novo* sequencing problem. Among them are naïve approaches, theory graph models, dynamic programming, probabilistic and combinatorial algorithms.

In the following different *de novo* sequencing algorithms are introduced and a list of available tools is given for quick reference. Thereafter, the accuracy and available comparisons among popular *de novo* sequencing tools are assessed. Finally, guidelines for the usage of *de novo* sequencing tools and possible pitfalls are discussed followed by an outlook on the development of the field in the near future.

The *de novo* Sequencing Problem

The de novo sequencing problem is the need to assign an amino acid sequence to an MS/MS spectrum without the use of a sequence database or other additional information. *De novo* sequencing, of MS/MS spectra, is possible due to the fact that a peptide can be fragmented into predictable parts with current methods like collision-induced dissociation, CID, or collisionally activated dissociation, CAD, [10,11]. Other methods like electron transfer dissociation, ETD, [12] electron capture dissociation, ECD, [13], and many more achieve the same result.

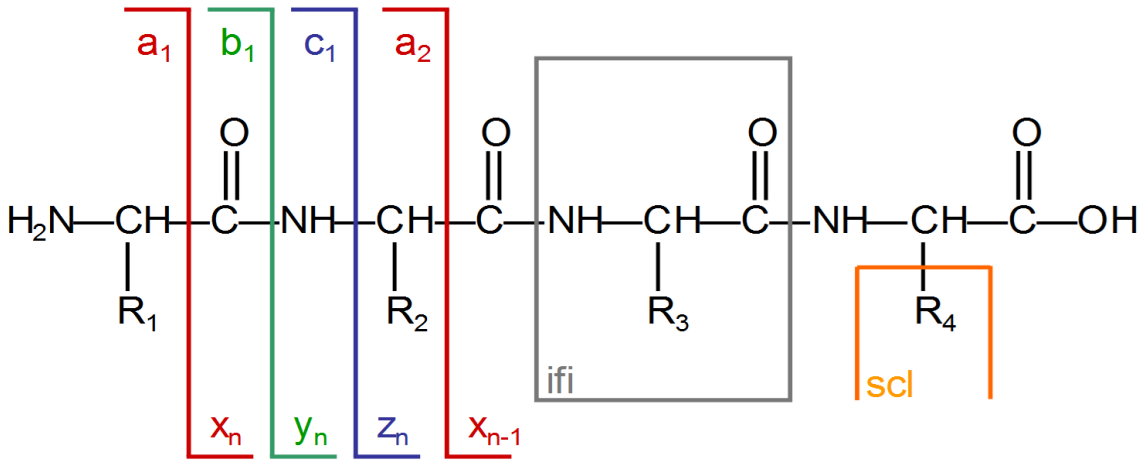

Figure 1: Theoretically a peptide can fragment at any chemical bond in the molecule. Some frequently observed ions are named within the figure (ifi: internal fragment ion, scl: side chain loss). The names of the n-terminal ions are indicated on the left of the respective fragmentation border whereas the c-terminal portions are named on the right. A subscript indicates their position within the sequence in respect to their mass.

Possible fragments have been named (Figure 1) for example *a*, *b*, and *c* for N-terminal fragments resulting from backbone cleavage of the amino acid sequence and *x*, *y*, and *z* for their complementing C-terminal fragments [14,15]. In the resulting MS/MS spectrum the ion types are not known *a priori* and they cannot easily be determined from basic principles for each peak. Thus the ion-types of the peaks in an MS/MS spectrum are unknown. Several fragments may further lead to similar mass to charge (m/z) ratios while neutral losses of H2O and NH3, for example, complicate the spectrum further.
One of the challenges in *de novo* sequencing is thus assigning an ion type to each peak. Unfortunately, not all ions from the theoretical fragmentation are found in MS/MS spectra and typically, for low accuracy but abundantly available mass spectrometers, the low m/z (<300 Dalton, Da) and high m/z (>1800 Da) are not well populated with diagnostic peaks. Seidler and his colleagues reviewed the fragmentation process for several instrumentations and fragmentation methods in respect to *de novo* sequencing in more detail [16].
Any algorithm for *de novo* sequencing needs to assign an ion-type to the abundant peaks in an MS/MS spectrum and needs to establish consecutive ions of the same type, differing in the mass of an amino acid (subsequently referred to as ion ladders) which then can be assigned an amino acid sequence. This is possible since there should be peaks that differ in the mass of an amino acid which indicates that they are of the same ion-type, forming an ion-ladder. However, due to the number of peaks many spurious assignments may be produced in this step. Furthermore, many peaks may be missed thus forming a complete ion-ladder to assign a sequence may not be possible.
Derivatization of peptides to discern n-terminal from c-terminal fragments has been used successfully to facilitate *de novo* sequencing [17,18]. Keough and colleagues found that derivatization helped them to get long un-interrupted ion ladders [19]. Derivatization methods that simplify MS/MS spectra and fix the precursor charge to one of the fragments have been described [20,21]. Although derivatization may aid in *de novo*

sequencing, such methods may not be widely adapted since they incur additional cost and labor. Thus, *de novo* sequencing algorithms need to be fast, reliable and shall not incur additional cost or experimental preparation steps. Instead of labeling the peptides natural isotopes can be used in a similar manner [22].
De novo sequencing can be solved in polynomial time when only n-terminal and c-terminal fragments are considered. Xu and Ma showed that it becomes NP-complete (not solvable in linear time) when further fragment types are incorporated [23].
Another issue is the mass accuracy of mass spectra. Mass spectra with higher mass accuracy from high precision mass spectrometer lead to better *de novo* predictions [24-27]. The fragmentation of precursor ions is not always successful, depending on the precursor abundance, how much of the precursor has been fragmented and with which energy. Therefore, it has been tried to establish the spectral quality prior to performing further analyses such as *de novo* sequencing [28,29]. Spectrum quality is here used in the sense of how well the precursor ion is fragmented into expected ions and how much their intensity is above potential random noise. It has been suggested that high-energy collisional dissociation (high-energy CID) spectra could aid in *de novo* sequencing [30] and it has been found that up to 80% of the high-energy CID spectra contain full or almost full ion ladders [31].
One problem, for both database search and *de novo* sequencing, which has not yet been properly addressed, is co-fragmentation of peptides with similar mass and retention time although there have been approaches in database searching [32].

De Novo Sequencing Algorithms

A large number of *de novo* sequencing algorithms have been proposed and a non-comprehensive list of algorithms, which provide an implementation, is presented in Table 1. Another list of algorithms, which do not provide an implementation, can be found in Supplementary Table 1. Both tables contain an algorithm column which can be used to roughly group the *de novo* sequencing algorithms. Therefore, *de novo* sequencing tools using the same basic algorithm are discussed together.

Table 1: Non comprehensive list of published *de novo* sequencing algorithms with available implementation. Information about availability, main algorithmic features, scoring function, and additional comments are provided. DP: dynamic programming, CO: commercial, OS: open source, EXE executable available, NA: listed source not available at time of writing. The table is sorted by decreasing citation count.

Name	Algorithm	Scoring	Comment	Implementation	Citation
PEAKS	Generation of 10^5 candidate sequences, DP	Peak abundance, mass fit, fragment complementarity	Commercial software, algorithm not fully disclosed	http://www.bioinfor.com:9999/, CO	[33]
Lutefisk	Spectrum graph	Sum of b-ion probabilities during subseqeuncing	Rescoring of prediction with several measures	http://www.hairyfatguy.com/lutefisk/, OS	[34]
PepNovo	Spectrum graph	Likelihood ratio hypothesis testing in respect to random model	Only few learned models available	http://cseweb.ucsd.edu/groups/bioinformatics/software.html, OS	[35]
PepNovo	Spectrum	As Sherenga but	High precision mass	http://peptide.usc	[24]

	graph	additionally uses peak ranks and fragment dependencies	spectrometric data needed	d.edu, OS	
Unnamed	Matrix spectrum graph	Ion abundance ratio	Tree searching to find all suboptimal solutions	http://hto-c.usc.edu:8000/msms/menu/denovo.htm, NA	[36]
NovoHMM	Hidden Markov model	Bayesian posterior probabilities for amino acids	Tested on 1252 spectra and compared to other algorithms	http://people.inf.ethz.ch/befische/proteomics/, EXE	[37]
SeqMS	Spectrum graph	Ion abundance, fragment complementarity	Originally for HCD spectra later adapted for low energy CxD	http://www.protein.osaka-u.ac.jp/rcsfp/profiling/Seqms/SeqMS.html, EXE	[18,38]
EigenMS	Spectral graph partitioning	Mass fit, ion abundance, probability to observe ion	Usage of two graphs,	http://sourceforge.net/projects/eigenms/, OS	[39]
AuDeNs	Spectrum graph, DP	Internally calculated sum of peak relevance	Assigns relevance to peaks during preprocessing	http://www.ti.inf.ethz.ch/pw/software/audens/, EXE	[40]
MSNovo	DP, mass array spectrum representation	Probabilistic distribution of mass tolerance	LCQ/LTQ Charges 1-3	http://msms.usc.edu/supplementary/msnovo, NA	[41]
MAARIAN	Exhausitive enumeration of peptide composition	Sum of peak abundance matched by sequence	MALDI, unimolecular decomposition, small example set, accuracy not assessed	Available upon request, EXE	[42]
PFIA	Exhaustive listing of all possible amino acid compositions	-	Ability to aid sequencing of cyclic peptides	http://hodgkin.mbu.iisc.ernet.in/~pfia/, NA	[43]
Vonode	Spectrum graph	Based primarily on mass accuracy but in part also on fragment abundance	Dependent on high mass accuracy, Also makes sequence tags	http://compbio.ornl.gov/Vonode, EXE	[44]

Naïve Approach

One of the first approaches used to assign a sequence to a tandem MS spectrum is the brute force, or naïve, approach. In brief, all amino acid sequences approximately matching the measured precursor mass are generated and scored against the spectrum. The sequence with the highest score is then accepted as the correct solution [45,46]. With increasing precursor mass the number of possible sequences increase exponentially restricted only by the precursor mass accuracy which prohibits the use of this approach above a low mass cutoff [47]. This can be offset with increasing mass accuracy which

may make this approach viable since the number of possible sequences is more restricted as shown in an approach termed composition based sequencing developed by Spengler in 2004 [27]. See Zubarev and Mann for a disambiguation of mass accuracy and its proper usage [48]. Determining the sequence composition of candidates from a database, enumerating possible sequences, can also limit the number of sequences that need to be explored [42]. The commercial software, PEAKS, also uses exhaustive listing of sequences but restricts the amount of sequences to a subset of 10000. Unfortunately, their algorithm has not been fully disclosed [33]. The rescoring of candidate sequences has later been improved in PEAKS-RM [23].
Instead of generating full sequences, subsequences can be determined by, for example extending short seed sequences determined directly from the MS/MS spectrum [49,50]. The DeNovo Explorer by Applied Biosystems, another example for subsequencing, first determines all subsequences and then transforms them into theoretical spectra and scores them against the experimental spectrum using percentage of matched peak intensity. More information is available on the web site of Applied Biosystems: http://www3.appliedbiosystems.com/cms/groups/psm_marketing/documents/generaldocuments/cms_040353.pdf. Often prefixes cannot be correctly determined and thus correct candidates are filtered [51]. Another alternative example for restricting the number of candidates determines sequence tags from the spectrum and builds a sequence database from these which is then scored against the spectrum [52].

Spectrum Graph

A spectrum graph is the transformation of a peak list into a graph where each m/z value is represented by a node in the graph. Nodes are connected by edges if their m/z values differ by the mass of an amino acid. As a side effect, random noise in the spectrum is reduced since it should not create more additional edges than real fragment ions. Unexplained ion types, which may be considered noise by some algorithms, however, contribute with additional edges that complicate the spectrum graph. Figure 2 shows a spectrum graph where one forward ion-series and one reverse series are shown.

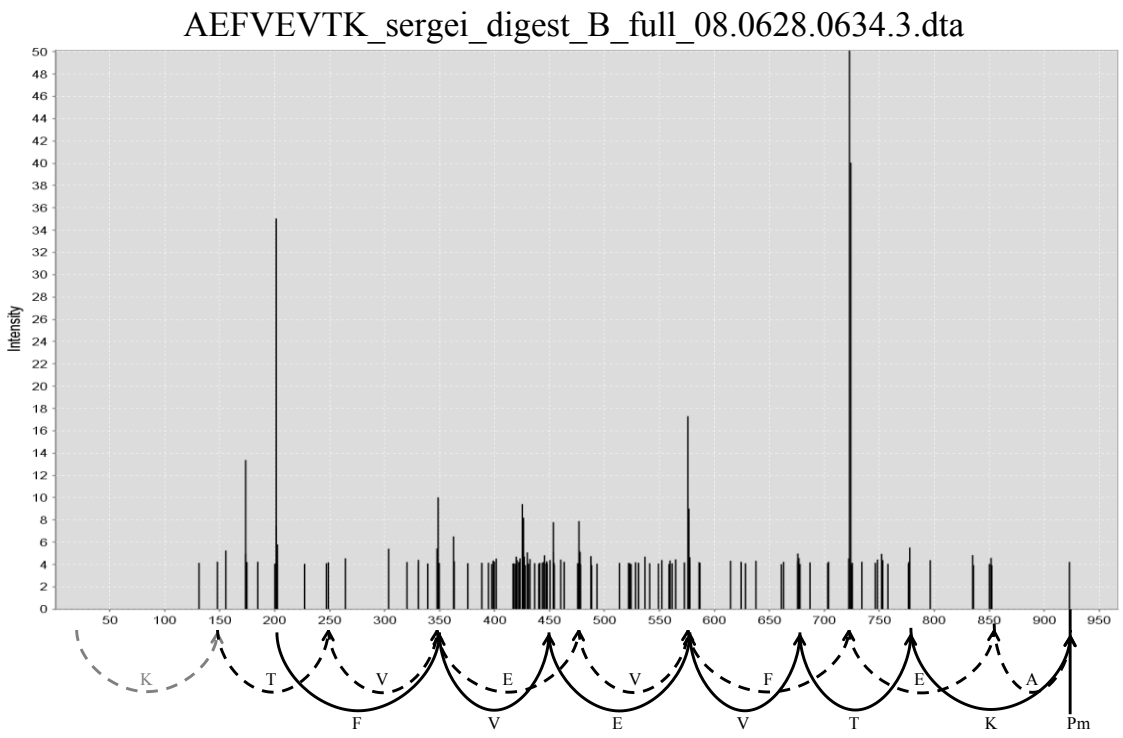

Figure 2: The figure shows the MS/MS spectrum sergei_digest_B_full_08.0628.0634.3.dta from the Keller dataset [53]. Ions of the singly charged b- and y-ion series are connected by edges with the amino acids as labels on the edges. Other node pairs that differ by the mass of an amino acid are not connected to avoid overcrowding of the figure. Pm indicates the precursor mass solid arrows show the b-ion series and dashed arrows show the y-ion series. Gray lines and amino acids are interpolated and have only partial evidence in the spectrum. The spectrum image was made with DtaViewer (http://www.biolnk.com, Allmer, unpublished).

The nodes of the spectrum graph can be drawn on one line as it is done in Figure 2 where the value on the m/z axis is taken as the node center. The edges shown only represent the b- and y-ion series but many more edges can be drawn. Abundance can be encoded as an edge or node weight. In Figure 2 both series are internally uninterrupted and thus lead to a correct *de novo* prediction when the spectrum graph is traversed from either the first node or the last node. More information on how a spectrum graph can be traversed can be found in the following sections. Note that, the b-ion series gives the sequence in forward fashion whereas the y-ion series produces its reverse. In this case the graph can be drawn directed since the sequence is known but in general the resulting spectrum graph is undirected. Instead of only drawing edges if the mass of an amino acid fits between two nodes, the use of different edge types, connecting b-ions, y-ions, as well as interconnecting them, can add vital information to the spectrum graph [44]. Yan and colleagues took a similar approach and thus increased the information contained in their spectrum graph [54]. Bern and Goldberg have further constructed several graphs and have used graph partitioning [39]. A spectrum graph is used in further algorithms in Table 1 that have not been listed here individually [18,24,31,35,36,38,55-58].

Dynamic Programming

Dynamic programming is a technique that can be applied if a problem can be broken into smaller problems which are able to solve the larger problems, once solved themselves. Furthermore, the ability to build upon intermediate results is necessary for dynamic programming, ensuring no recalculations, which makes it especially suitable for finding optimal paths through a graph (e. g.: spectrum graph), although faster heuristics exist. Therefore, several algorithms make use of this technique for finding a path through the spectrum graph which may represent the peptide sequence of the precursor. Dynamic programming guarantees to find the optimal result but due to the complexity and not fully understood nature of MS/MS data the optimal sequence may not represent the correct sequence. In *de novo* sequencing, dynamic programming can be used to find the path potentially representing the correct peptide sequence in a spectrum graph [40,59-61]. A mass array encoding all possible peptide sequences can also be traversed by dynamic programming [41].

Sequence Optimization

Instead of analyzing the MS/MS spectrum, it is also possible to optimize amino acid sequences to determine the one that best fits to the spectrum in respect to a scoring function. Genetic algorithms have been used for sequence optimization which is in essence quite similar to the naïve approach with the difference that not all possible amino acid sequences need to be generated. Instead, a small pool of amino acid sequences is generated and then optimized to best fit the MS/MS spectrum. This heuristic comes at the cost that finding the optimal result cannot be guaranteed. Moreover, in subsequent runs of the same algorithm with the same MS/MS spectrum different sequences may be reported as the best result. An early approach probably failed due to the fitness function which was not discriminative enough [62]. Heredia-Langner and colleagues, using a different fitness function, including shared peak count were more successful on the few spectra that they had at their disposal [63]. They compared the results with Lutefisk and reported to have similar sequence correctness.

Other Algorithms

Some approaches like the creation of spectrum graphs are shared among algorithms and are thus discussed in some more detail. Other algorithms like divide and conquer, for instance, are not in wide spread use. Zhang has used the divide and conquer algorithm for splitting the spectrum into successively smaller sub-spectra until they were solvable by the naïve approach and are then recursively re-integrated to solve the input spectrum [64]. The use of hidden Markov models, as an example for machine learning algorithms, has been used to tackle the *de novo* sequencing problem [37]. A pattern based algorithm has been introduced by Hines and colleagues [65]. Another recent approach seems to employ exhaustive listing of subsequences to aid manual interpretation of MS/MS spectra [43]. Most de novo sequencing algorithms consider all fragments to be singly charged but sequencing accuracy can be increased when using multiple charges as shown by Chong and colleagues [66].

Integrative Approaches

A tandem MS spectrum has limited information and several MS^2 spectra may be used to add confidence to the information while MS/MS spectra generated with different fragmentation models can add information and add additional confidence at the same time.

Bandeira and colleagues clustered MS/MS spectra and were thus able to derive more sequence information [67]. Another way of making use of the combination of multiple spectra is to collect MS^3 spectra from selected fragment ions in MS/MS spectra and then combining the information to yield more meaningful data [68,69].

Zhang and Reilly used a combination of two MALDI fragmentation methods, post source decay, PSD, [70] and photodissociation [71]. They reported a sequence prediction success of about 91% and the ability to differentiate leucine and isoleucine, alas only on a very small dataset of 31 peptides [72]. Their approach makes use of x-ions to derive the peptide sequence and then uses y-, v-, and w-ions for further analysis. Datta and Bern combined the information contained in CID and ETD fragmentation spectra using an algorithm that not only combines the data but also separates n-terminal from c-terminal fragments [73]. They reported sequencing accuracies between 17% and 65%.

Horn and colleagues discerned n-terminal and c-terminal ions using a combination of CID and electron capture dissociation, ECD, [13] spectra [74] while Savitsky and colleagues used the same fragmentation methods but integrated the information therein to yield a higher confidence for their predictions [26]. Zubarev and colleagues also investigated how ECD and CID fragmentation methods can be used synergistically and further included ETD fragmentation in their study [75]. Li and colleagues investigated how the combination of ECD and ETD spectra can aid in *de novo* sequencing [76].

Algorithm Comparison

In the absence of comprehensive datasets that cover the wide range of instrumentation and fragmentation possibilities as well as measurement settings it is not possible to proclaim one algorithm to be better than any other algorithm except for a particular instance of the problem [77]. One public dataset generated with an LCQ mass spectrometer from Thermo Electron published by Keller and colleagues [53], has been used in several studies to determine the quality of *de novo* prediction algorithms. It is highly desirable to have similar datasets for all types of mass spectrometers to enable a more complete picture about the qualities of *de novo* sequencing algorithms.

Generally, datasets are created as a by product of other studies and are based on the identifications of database search programs which cannot guarantee correct identification; see recent reviews on database searching algorithms [78,79]. For benchmarking *de novo* sequencing algorithms this is not acceptable and all spectra should be prepared from purified peptides and synthetic peptides to ensure correct sequence assignment.

Accuracy Definitions

In order to make *de novo* predictions comparable, an accuracy measure has to be selected. Since an accuracy measure to evaluate *de novo* sequencing algorithms has not been agreed upon, different studies use different quality measures.

Xu and Ma used two measures, 1) the number of correctly predicted amino acids divided by the number of amino acids in the real peptide, and 2) the number of correctly predicted amino acids divided by the number of amino acids in the prediction [23]. It also needs to be agreed upon when an amino acid is said to be correctly predicted. Xu and Ma state that the amino acid must be at the same mass position in the prediction as in the correct peptide [23]. But other definitions like the one by Pitzer and colleagues who use the notion of longest common subsequence to evaluate the accuracy of sequence predictions [80], may equally well define similarity. Another similarity measure was given by Pevtsov and colleagues who extended the edit distance algorithm with modifiers explicitly modeling expectable problems in *de novo* predictions [81]. They used relative sequence distance as their accuracy measure. Mo and colleagues define the prediction accuracy as the ratio of correctly predicted amino acids over total amino acids in the predicted peptide (similar to Xu and Ma) and recall as the number of correctly predicted residues over the total number of residues in the correct peptide [41]. Bringans and colleagues use the former method for their accuracy assessment of *de novo* sequencing algorithms [82]. An all or nothing score, either the sequence is correctly predicted and the best prediction of the algorithm, or otherwise is false, may be too harsh a criterion with current instrumentation and *de novo* sequencing algorithms. Nonetheless, it is one of the scores employed in a study by DiMaggio and Floudas [55]. They further report the prediction accuracy in respect to sequence distance similar to but not exactly like Pevtsov and colleagues. They further incorporate percentage of matched amino acid residues and correctly predicted subsequences in their study.
For the future, it would be desirable to agree upon a number of accuracy measures that need to be reported in a study comparing different *de novo* sequencing algorithms.

Comparisons of Multiple Algorithms

Currently, there is a trend that each newly developed *de novo* sequencing algorithm compares its results with a selection of other algorithms. Unfortunately, independent researchers rarely assess the quality of de novo sequencing algorithms which would remove possible biases. In a crude assessment we checked the overlap of sequence assignments of several *de novo* sequencing algorithms with database search results. Spectra that had agreeing sequence assignments by Sequest, OMSSA, and X!Tandem, were used for this comparison. The outcome was most disappointing for the *de novo* sequencing algorithms which is likely due to the usage of spectra with low mass accruracy from an LCQ mass spectrometer (Boz and Allmer, unpublished).
Xu and Ma used three small datasets to compare PEAKS, their extension to PEAKS (PEAKS-RM), and PepNovo and concluded that their new method was better than the other methods, with them performing similarly on the datasets that they investigated. For their dataset they achieved an accuracy between 6 to 7 out of 10 amino acids [23]. Often, as also in this case, *de novo* sequencing algorithms are tested on rather small datasets comprised of high quality spectra [80]. While Pitzer and colleagues improved on this and incorporated spectra from several instruments, they still only used a subset of available instrumentation and algorithms. They compared Lutefisk and PepNovo and observed that PepNovo is more successful on ion trap and MALDI data whereas Lutefisk is more successful on QTOF data. From their data it can also be derived that, if an average peptide length of 10 is assumed, the correct prediction of more than 6 out of 10 amino

acids is a rare event. Pevtsov and colleagues compared a larger number of different *de novo* sequencing algorithms, namely Audens, Lutefisk, NovoHMM, PepNovo, and PEAKS [81]. Their dataset was somewhat smaller than the one of Pitzer and colleagues but one ingredient, the measurements of Keller and colleagues are present in both studies as well as in a study by Mo et al. [41]. Pitzer and colleagues found that QStar data, with maximum of 50% correct sequences, lead to better predictions for all algorithms as compared to LCQ data, with a maximum of 18% correctly sequenced spectra. They found that for QStar data PEAKS performs better than Lutefisk and PepNovo, which in turn perform better than Audens and NovoHMM. For LCQ data NovoHMM performed best with PepNovo and PEAKS close behind, followed by Lutefisk and then Audens. Mo and colleagues focused on LCQ and LTQ data and reported that their algorithm performs better than PepNovo and NovoHMM with an accuracy between 12% and 50% depending on the dataset. Looking at the reported recall values MSNovo, their algorithm, does not significantly outperform the other tools and is at times less precise than NovoHMM which could put MSNovo on a level with NovoHMM in the Pevtsov study for LCQ data. Bringans and colleagues found that for a small data set of 4800 MALDI TOF/TOF spectra PEAKS was, with 66%, slightly more accurate than PepNovo and both were more than 10% more accurate than Applied Biosystem's DeNovoExplorer [82]. For QTRAP data PepNovo was slightly better than PEAKS and achieved an accuracy of 65%. Both were significantly better than DeNovoExplorer which only predicted 27% of the residues correctly. DiMaggio and Floudas compared the performance of several *de novo* sequencing algorithms on a very small number of spectra (38) from a QTOF instrument [33] and found that their algorithm PILOT performed best with PEAKS Online, EigenMS, Lutefisk, and LutefiskXP being decreasingly less accurate in terms of fully correct sequence prediction [55]. For ion trap data (36 spectra) they found that PILOT performed best with NovoHMM, PepNovo, EigenMS and PEAKS Online, and Lutefisk XP being decreasingly less accurate.

The studies by Bringans and colleagues, Pevtsov and colleagues, as well as Pitzer and colleagues were independent assessments of prediction accuracy, whereas the other studies mentioned above were a selection of studies that while introducing a new algorithm compare it with existing algorithms.

Determining Sequence Tags

As shown in the previous section, it is difficult to determine full length sequences from MS/MS spectra. In order to map the spectra to their peptide sequence, or several spectra to a protein sequence, it is not absolutely necessary to determine full length *de novo* sequences. Short sub sequences, so called sequence tags, can be successfully mapped to protein sequences if a sequence database exists. This can be viewed as a special case of *de novo* sequencing where the tags don't have to be full length sequences. Or, vice versa, *de novo* sequencing could be a special case of determining sequence tags where the tag happens to be the full length sequence. Determining short sequence tags with high confidence enables the retrieval of peptide sequences from a database that match to this tag; a strategy often referred to as filtration. The problem is whether a tag is forward, derived from an n-terminal ion series, or reverse, derived from a c-terminal ion series. Programs that derive sequence tags from MS/MS spectra are listed in Table 2.

Table 2: A non comprehensive list of algorithms determining a sub sequence (sequence tag) from MS/MS spectra. OS: open source, EXE executable available, NA: listed source not available at time of writing. The table is sorted by decreasing citation count.

Name	Comment	Implementation	Citation
GutenTag	Determines short sequences from MS/MS, ranked by presence of expected ions	http://fields.scripps.edu, EXE	[83]
OpenSea	Treats sequence database and mass spectrum as a sequence of masses	http://libopensea.com/, ?	[84]
PepNovoTag	Uses PepNovo for tag generation and employs a probabilistic approach for tag scoring	http://proteomics.ucsd.edu/, NA	[85]
DirecTag	Evaluates tags in respect to peak intensity, m/z fit, and fragment ion complementarity	http://fenchurch.mc.vanderbilt.edu/bumbershoot/directag/, OS	[86]
Spectral Profiles	Creates gapped sequence tags and spectral profiles from MS/MS spectra	NA	[87]

Short tags usually map to several peptides in a database so that a combinatorial problem needs to be solved. A number of algorithms have been developed to solve this problem (see Table 3). The generation of sequence tags is often a prerequisite for the algorithms listed in Table 3. GutenTag for instance assembles tags of a user specified length from MS/MS spectra and then matches the 25 best scoring tags to a sequence database [83]. Tabb and colleagues later applied more rigorous statistical models to the tag generation in their DirectTag algorithm [86]. Frank and colleagues emphasize the importance of tags to be on a valid global *de novo* path and add a probabilistic filtering step to the tag creation [85]. Instead of using sequence tags Searle and colleagues introduced the notion of mass tags treating both database and tag as a sequence of masses [84].

Table 3: A non comprehensive list of algorithms that map sequence tags or full *de novo* sequences derived from mass spectra to a sequence database. OS: open source, EXE executable available, NA: listed source not available at time of writing, OV: online version, AR available upon request. The table is sorted by decreasing citation count.

Name	Algorithm	Comments	Implementation	Citation
MSBlast	Accounts for MS specific problems and maps tags to database	Uses WU-BLAST2 for homology searching	http://genetics.bwh.harvard.edu/msblast, OV	[88]
Inspect	Uses spectrum graph for tag generation and trie for database search	Uses filtering of the database to find PTMs	http://proteomics.ucsd.edu, OS	[89]
FASTS	Integrates the combined mapping of multiple short sequences to a sequence database	Based on FASTA, comes in two flavors, FASTS and FASTF	http://fasta.bioch.virginia.edu/fasta_www2/fasta_list2.shtml, OS	[90]
CIDentify	Uses complete *de novo* sequences with mass gaps	Allows for common errors in the *de novo* prediction	http://faculty.virginia.edu/wrpearson/fasta/OLD/CIDentify/, EXE	[91]
ByOnic	Determines likely b- and y-ions for lookup in the subsequent database search	Tests showed that it was more sensitive than database search algorithms	NA	[92]
GenomicPe	Maps *de novo* predictions	Allows analysis of	http://www.allmer.	[93,94]

ptideFinder	error tolerantly to a genomic database	intron exon borders and can find alternative splicing events	de/software/gpf/, AR	
MS-Dictionary	Matches all plausible *de novo* predictions to the sequence database	They report increased sensitivity in comparison to X!Tandem	http://proteomics.ucsd.edu, AR	[95]
IggyPep	Maps *de novo* predictions error tolerantly to an indexed genomic database	Unclear how it improved upon existing methods	http://www.iggypep.org, OV	[96]
Spider	Accounts for errors in sequence tags	Uses editing operations to align sequences	http://bif.csd.uwo.ca/spider, NA	[97]

Shevchenko and colleagues used PredictSequence for generation of sequence tags and used WU-BLAST2 for matching the sequence tags, while accounting for typical MS problems and integrating the resulting database matches [88].
Mackey and colleagues build on the FASTA heuristic and added combinatorics to integrate multiple short hits and also changed FASTA's scoring scheme to alignment probability [90]. A similar approach has been taken by Johnson and Taylor who have used CIDentify, a version of FASTA predating Mackey's improvement [91]. Lu and Chen implemented a search of a sequence database, indexed using a suffix tree, against a spectrum graph [98]. The GenomicPeptideFinder (GPF) uses *de novo* predictions and maps them error tolerantly to the six frame translation of a genomic database. It has been used to investigate *Chlamydomonas reinhardtii* and significantly increased the number of detected peptides as well as suggested new splicing events [93,94]. RAId [99] and MS-Dictionary [87] also use multiple *de novo* sequences to map them to the six frame translation of a genomic database. IggyPep uses an indexed genomic database to map *de novo* sequences or sequence tags. It has been used on the genome of Sea Urchin and helped to find additional neuropetides [8]. It is, however, unclear how IggyPep improves upon GPF, RAId, MSBlast and other methods. Inspect uses sequence tags to filter a sequence database to derive candidates that may contain PTMs [89].
Instead of using sequence tags, or gapped tags, [87] ByOnic uses lookup peaks to filter sequence databases [92].

Data Integration

De novo sequencing provides the complete amino acid sequence of an MS/MS spectrum whereas sequence tags only provide short but positioned subsequences. Both can be used for error tolerant database search while *de novo* sequence predictions can also be used standalone. Integrating information from *de novo* sequencing, database search and additional resources can increase the confidence in each protein identification [100]. OVNIp is an application that has recently been published. It allows the exploitation of *de novo* sequences, in tandem with database search programs, to increase the confidence in protein identification [101]. In the future, it would be good to have similar tools which from *de novo* peptide predictions create *de novo* protein predictions and compare these with known protein sequences.

Application

De novo sequencing has recently gained a lot of attention and numerous studies, making use of it, have been published in the last three years. Due to the low accuracy of *de novo* predictions, it is mostly used in tandem with database search algorithms. In this way, adding confidence to peptide and protein identification, aiding in enlarging the number of MS/MS spectra that can be assigned a sequence, or suggesting MS/MS spectra that could be derived from a peptide precursor with a PTM.

Although more than 60% of amino acid residues can be correctly predicted by some algorithms under specific circumstances, only 30% of the peptides are correctly predicted [82]. The accurate prediction of full-length sequences remains challenging [31]. This is quite limiting when working with unsequenced species or sequences that are not expected to be in any database. This is inline with the view of Zubarev and colleagues, who state that *de novo* sequencing is the answer to interpretation of mass spectra but point out that adequate accuracy can only be achieved with highly accurate mass spectrometers, using a combination of multiple fragmentation methods [75].

Regardless of the limitations, *de novo* sequencing has been used, in high-throughput studies and lead to findings that would not have been possible otherwise [102,103]. The current limitations of sequencing peptides and proteins *de novo* is very well documented in a study that sequenced a beta-defensin of reptilian origin. The authors faced a multitude of problems and resorted to the use of a combination of multiple mass spectrometers, multiple fragmentation methods and different derivatization methods, as well as getting aid from Edman degradation [104], until they were successful [105]. Another study, sequencing a hormone, faced similar problems and came to similar conclusions but further added top-down sequencing while not employing Edman degradation [106].

Outlook

De novo sequencing may be useful in assigning meaning to unidentified, high quality mass spectra, when it is used as a part of an identification pipeline, for example extending the one built by Ning and colleagues [107]. Such multi-pass approaches, in contrast to multi-search analyses, reduce the number of false negatives while not significantly increasing the overall runtime [108]. Therefore it can be expected that *de novo* sequencing will, increasingly, find its way into computational pipelines for MS data analysis, as exemplified in a study by Junqueira and colleagues [109], in the near future.

It has long been recognized in proteomics that standards and standard datasets need to be available for benchmarking of the current state of the art and new methods [110]. This is not different in *de novo* sequencing and it seems essential that current tools be properly evaluated under a wide range of practical conditions. Although amenable, it is unlikely that this will happen within the near future due to constantly changing instrumentation and fragmentation methods.

In general *de novo* sequencing algorithms are easier adaptable to include the detection of post translational modifications than database search algorithms which quickly have to face databases of insurmountable size when multiple PTMs are considered. Although many *de novo* sequencing algorithms include the ability to search for static, variable, multiple, or a combination of PTMs, due to the infancy of the *de novo* prediction

algorithms for PTM detection, they have not been discussed in this review. In the near future more tools will be developed that either predict PTMs directly from MS/MS data or use multi-pass analysis on a collection of spectra to assign PTMs [111-113]. With these methods or a combination of methods, proper assignment of PTMs will become possible.

Expert commentary

One of the central dogmas in biology, one gene leading to one transcript, has been proven false and the real problem seems quite complicated rather than straight forward. This includes alternative splicing, alternative start and stop sites and other means that lead to unexpected proteins. Mass spectrometry has become predominant for many areas of in proteomics. For protein identification database search is being used but this strategy fails when any alternative transcripts have not been annotated and are thus not available in sequence databases. This problem is elevated since most proteins have not been sequenced on the protein level but are mere predictions with large associated errors. Mapping EST sequences can help in this case but currently only for abundant transcripts. An additional problem inherent in database search algorithms is that the precursor mass needs to fit to the sequence within certain bounds which makes post translational modifications a difficult problem.

Therefore, only *de novo* sequencing algorithms can ultimately be used to sequence proteins. This currently comes with the cost that *de novo* sequencing algorithms are not precise enough to provide high confidence in their sequencing results. This may in one part be due to the fact that more effort has been put into the development and assessment of database search algorithms and in another part be due to the need for high quality spectra when performing *de novo* sequencing. Another point is the large abundance of different instrumentation including various fragmentation methods. Peptide fragmentation is still under investigation for even established methods and has not been fully understood. Fragmentation pathways are the basis for many *de novo* sequencing algorithm and if they were clearly defined the prediction quality would increase considerably. At this point it is thus difficult for a user to find the best suited *de novo* sequencing algorithm for their instrumentation.

It is, however, possible to amend database search results with *de novo* sequencing results and the synthesis of this information can already increase the detectable number of proteins and the confidence in their identification.

Five-year view

Currently, the number of different instruments and fragmentation methods is constantly increasing and this may not stop in the near future making it possible for developers to present ever more niche *de novo* sequencing algorithms targeted to one particular mass spectrometer. This does not help the consolidation of current *de novo* sequencing methods and also complicates the independent comparison of their performance which is seen by the fact that only few independent comparisons have been made but not within the last 3 years.

A trend, further complicating the matter, to combine spectra from multiple fragmentation procedures can be seen in the literature but ultimately it will be necessary to perform

sequencing using one mass spectrometer with a single fragmentation method to enable high throughput analyses and thus these studies may be abandoned in the future.
Post translational modifications (PTMs) pose a great difficulty to mass spectrometry-based proteomics if the PTMs are not anticipated. Some studies claim to accommodate for even unexpected PTMs and in the near future a focus of the field will be to turn this claim into reality for a selected number of high end mass spectrometers. Within the next three years some algorithms will be able to sequence short peptides containing one unexpected PTM successfully.
In general, the support for low cost mass spectrometers is decreasing and the focus will be on new high end machines which offer a greater success rate for *de novo* sequencing algorithms and it is unlikely that this trend will reverse in the future.
Within the next five years fragmentation pathways will not be fully understood and thus algorithms based on this expert knowledge will still be hampered by unexplained peaks within MS/MS spectra.
Within the next five years the number of comparative studies may increase again whenever competing *de novo* sequencing algorithms are proposed for the same mass spectrometer and fragmentation method combination. There will however be very few studies comparing *de novo* sequencing results across platforms a trend which will continue until new mass spectrometers are only marginally better than their predecessors.

Key issues

- Mass spectrometry is the key tool for performing proteomics.
- Database search algorithms, and those using sequence tagging, have inherent problems that prevent them from identifying all possible amino acid sequences.
- De novo sequencing is similar to sequence tagging but provides a full length amino acid sequence; it is dissimilar from it and database search algorithms in that it is independent of any additional information than contained in an MS/MS spectrum.
- De novo amino acid sequencing from MS/MS spectra offers the possibility to sequence any peptide or protein precursor.
- Many de novo sequencing approaches have been proposed using a wide variety of algorithms to solve the de novo sequencing problem.
- It is difficult to compare existing algorithms since they are often targeted to specific experimental setups and since no quality measures have been agreed upon in the field.
- The accuracy of de novo predictions is not yet good enough to solely rely on them for the sequencing of a proteome.
- Although it is possible, for most existing algorithms, to include post translational modifications this is not discussed in this review due to the infancy of the field in this respect.
- Integrative approaches using a combination of methods are able to elevate this problem for already sequenced organisms; thus being able to combine database, sequence tagging, de novo sequencing and other information.

Acknowledgements

The Turkish Academy of Science (TÜBA) supported the preparation and publication of this article.

Annotated references:
1 (Mann): ** Excellent introductory review to the field of mass spectrometry based proteomics
2 (Allmer): * Contains additional information on the impact of post translational modifications not covered in this review.
33 (PEAKS): * One of the most widely used commercial de novo sequencing algorithm
35 (PepNovo): * One of the most widely used free de novo sequencing algorithm
80 (Pitzer): ** Independent comparison of multiple de novo sequencing algorithms
81 (Pevtsov): ** Independent comparison of multiple de novo sequencing algorithms
82 (Bringans): ** Independent comparison of multiple de novo sequencing algorithms

References

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

2 Allmer J. Existing bioinformatics tools for the quantitation of post-translational modification. *Amino Acids* Published Online First: 2010. doi:10.1007/s00726-010-0614-3

3 Eng J, 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 Spectr* 1994;**5**:976–89.

4 Perkins DN, Pappin DJ, Creasy DM, *et al.* Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* 1999;**20**:3551–67.

5 Geer LY, Markey SP, Kowalak JA, *et al.* Open mass spectrometry search algorithm. *J Proteome Res* 2004;**3**:958–64.

6 Craig R, Beavis RC. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* 2004;**20**:1466–7.

7 Standing KG. Peptide and protein de novo sequencing by mass spectrometry. *Curr Opin Struct Biol* 2003;**13**:595–601.

8 Menschaert G, Vandekerckhove TTM, Baggerman G, *et al.* Peptidomics Coming of Age: A Review of Contributions from a Bioinformatics Angle. *Journal of Proteome Research* 2010;**9**:2051–61.

9 Bandeira N, Pham V, Pevzner P, *et al.* Automated de novo protein sequencing of monoclonal antibodies. *Nat Biotechnol* 2008;**26**:1336–8.

10 Wells JM, McLuckey SA. Collision-induced dissociation (CID) of peptides and proteins. *Meth Enzymol* 2005;**402**:148–85.

11 Sleno L, Volmer DA. Ion activation methods for tandem mass spectrometry. *J Mass Spectrom* 2004;**39**:1091–112.

12 Syka JEP, Coon JJ, Schroeder MJ, *et al.* Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. *Proceedings of the National Academy of Sciences of the United States of America* 2004;**101**:9528–33.

13 Zubarev RA, Kelleher NL, McLafferty FW. Electron Capture Dissociation of Multiply Charged Protein Cations. A Nonergodic Process. *Journal of the American Chemical Society* 1998;**120**:3265–6.

14 Biemann K. Appendix 5. Nomenclature for peptide fragment ions (positive ions). *Meth Enzymol* 1990;**193**:886–7.

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

16 Seidler J, Zinn N, Boehm ME, *et al.* De novo sequencing of peptides by MS/MS. *Proteomics* 2010;**10**:634–49.

17 Uttenweiler-Joseph S, Neubauer G, Christoforidis S, *et al.* Automated de novo sequencing of proteins using the differential scanning technique. *Proteomics* 2001;**1**:668–82.

18 Fernandez-de-Cossio J, Gonzalez J, Betancourt L, *et al.* Automated interpretation of high-energy collision-induced dissociation spectra of singly protonated peptides by "SeqMS", a software aid for de novo sequencing by tandem mass spectrometry. *Rapid Commun Mass Spectrom* 1998;**12**:1867–78.

19 Keough T, Lacey MP, Fieno AM, *et al.* Tandem mass spectrometry methods for definitive protein identification in proteomics research. *Electrophoresis* 2000;**21**:2252–65.

20 An M, Dai J, Wang Q, *et al.* Efficient and clean charge derivatization of peptides for analysis by mass spectrometry. *Rapid Commun Mass Spectrom* 2010;**24**:1869–74.

21 Chen W, Lee PJ, Shion H, *et al.* Improving de Novo Sequencing of Peptides Using a Charged Tag and C-Terminal Digestion. *Analytical Chemistry* 2007;**79**:1583–90.

22 Cannon WR, Jarman KD. Improved peptide sequencing using isotope information inherent in tandem mass spectra. *Rapid Commun Mass Spectrom* 2003;**17**:1793–801.

23 Xu C, Ma B. Complexity and scoring function of MS/MS peptide de novo sequencing. *Comput Syst Bioinformatics Conf* 2006;:361–9.

24 Frank AM, Savitski MM, Nielsen ML, *et al.* De Novo Peptide Sequencing and Identification with Precision Mass Spectrometry. *Journal of Proteome Research* 2007;**6**:114–23.

25 Wong J, Sullivan M, Cartwright H, *et al.* msmsEval: tandem mass spectral quality assignment for high-throughput proteomics. *BMC Bioinformatics* 2007;**8**:51.

26 Savitski MM, Nielsen ML, Kjeldsen F, *et al.* Proteomics-grade de novo sequencing approach. *J Proteome Res* 2005;**4**:2348–54.

27 Spengler B. De novo sequencing, peptide composition analysis, and composition-based sequencing: a new strategy employing accurate mass determination by fourier

transform ion cyclotron resonance mass spectrometry. *J Am Soc Mass Spectrom* 2004;**15**:703–14.

28 Bern M, Goldberg D, McDonald WH, *et al.* Automatic quality assessment of Peptide tandem mass spectra. *Bioinformatics* 2004;**20 Suppl 1**:I49–54.

29 Purvine S, Kolker N, Kolker E. Spectral quality assessment for high-throughput tandem mass spectrometry proteomics. *Omics* 2004;**8**:255–65.

30 Olsen JV, Macek B, Lange O, *et al.* Higher-energy C-trap dissociation for peptide modification analysis. *Nat Methods* 2007;**4**:709–12.

31 Chi H, Sun R-X, Yang B, *et al.* pNovo: De novo Peptide Sequencing and Identification Using HCD Spectra. *Journal of Proteome Research* 2010;**9**:2713–24.

32 Bern M, Finney G, Hoopmann MR, *et al.* Deconvolution of Mixture Spectra from Ion-Trap Data-Independent-Acquisition Tandem Mass Spectrometry. *Analytical Chemistry* 2010;**82**:833–41.

33 Ma B, Zhang K, Hendrie C, *et al.* PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. *Rapid Commun Mass Spectrom* 2003;**17**:2337–42.

34 Taylor JA, Johnson RS. Sequence database searches via de novo peptide sequencing by tandem mass spectrometry. *Rapid Commun Mass Spectrom* 1997;**11**:1067–75.

35 Frank A, Pevzner P. PepNovo: de novo peptide sequencing via probabilistic network modeling. *Anal Chem* 2005;**77**:964–73.

36 Lu B, Chen T. A suboptimal algorithm for de novo peptide sequencing via tandem mass spectrometry. *J Comput Biol* 2003;**10**:1–12.

37 Fischer B, Roth V, Roos F, *et al.* NovoHMM: a hidden Markov model for de novo peptide sequencing. *Anal Chem* 2005;**77**:7265–73.

38 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* 2000;**21**:1694–9.

39 Bern M, Goldberg D. De novo analysis of peptide tandem mass spectra by spectral graph partitioning. *J Comput Biol* 2006;**13**:364–78.

40 Grossmann J, Roos FF, Cieliebak M, *et al.* AUDENS: a tool for automated peptide de novo sequencing. *J Proteome Res* 2005;**4**:1768–74.

41 Mo L, Dutta D, Wan Y, *et al.* MSNovo: A Dynamic Programming Algorithm for de Novo Peptide Sequencing via Tandem Mass Spectrometry. *Analytical Chemistry* 2007;**79**:4870–8.

42 Olson MT, Epstein JA, Yergey AL. De Novo Peptide Sequencing Using Exhaustive Enumeration of Peptide Composition. *Journal of the American Society for Mass Spectrometry* 2006;**17**:1041–9.

43 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 Commun Mass Spectrom* 2007;**21**:3033–8.

44 Pan C, Park B, McDonald W, *et al.* A high-throughput de novo sequencing approach for shotgun proteomics using high-resolution tandem mass spectrometry. *BMC Bioinformatics* 2010;**11**:118.

45 Hamm CW, Wilson WE, Harvan DJ. Peptide sequencing program. *Comput Appl Biosci* 1986;**2**:115–8.

46 Sakurai T, Matsuo T, Matsuda H, *et al.* PAAS 3: A computer program to determine probable sequence of peptides from mass spectrometric data. *Biological Mass Spectrometry* 1984;**11**:396–9.

47 Allmer J. *GenomicPeptideFinder*. 2006.

48 Zubarev R, Mann M. On the Proper Use of Mass Accuracy in Proteomics. *Molecular & Cellular Proteomics* 2007;**6**:377–81.

49 Siegel MM, Bauman N. An efficient algorithm for sequencing peptides using fast atom bombardment mass spectral data. *Biological Mass Spectrometry* 1988;**15**:333–43.

50 Biemann K, Cone C, Webster BR, *et al.* Determination of the amino acid sequence in oligopeptides by computer interpretation of their high-resolution mass spectra. *J Am Chem Soc* 1966;**88**:5598–606.

51 Lu B, Chen T. Algorithms for de novo peptide sequencing using tandem mass spectrometry. *Drug Discovery Today Biosilico* 2004;**2**:85–90.

52 Sun H, Zhang J, Liu H, *et al.* TVNovo: De novo peptide sequencing for high resolution LTQ-FT mass spectrometry using virtual database searching. In: *3rd International Conference on Biomedical Engineering and Informatics (BMEI)*. Yantai: 2010. 2240–5.

53 Keller A, Purvine S, Nesvizhskii AI, *et al.* Experimental protein mixture for validating tandem mass spectral analysis. *Omics* 2002;**6**:207–12.

54 Yan B, Pan C, Olman VN, *et al.* A graph-theoretic approach for the separation of b and y ions in tandem mass spectra. *Bioinformatics* 2005;**21**:563–74.

55 DiMaggio PA, Floudas CA. De Novo Peptide Identification via Tandem Mass Spectrometry and Integer Linear Optimization. *Analytical Chemistry* 2007;**79**:1433–46.

56 Taylor JA, Johnson RS. Implementation and Uses of Automated de Novo Peptide Sequencing by Tandem Mass Spectrometry. *Analytical Chemistry* 2001;**73**:2594–604.

57 Dancik V, Addona TA, Clauser KR, *et al.* De novo peptide sequencing via tandem mass spectrometry. *J Comput Biol* 1999;**6**:327–42.

58 Bartels C. Fast algorithm for peptide sequencing by mass spectroscopy. *Biological Mass Spectrometry* 1990;**19**:363–8.

59 Goto MA, Schwabe EJ. A Dynamic Programming Algorithm for Finding Highest-Scoring Forbidden-Pairs Paths with Variable Vertex Scores. In: *Bioinformatics Research and Applications*. Berlin, Heidelberg: Springer Berlin, Heidelberg 2008. 171–82.

60 Bafna V, Edwards N. On de novo Interpretation of Tandem Mass Spectra for Peptide Identification. *ACM Press, New York, NY, USA* 2003;:9–18.

61 Chen T, Kao MY, Tepel M, *et al.* A dynamic programming approach to de novo peptide sequencing via tandem mass spectrometry. *J Comput Biol* 2001;**8**:325–37.

62 Stranz DD, Martin III LB. Derivation of Peptide Sequence from Mass Spectral Data using the Genetic Algorithm. *J Biomol Tech* 1999;**9**:1–8.

63 Heredia-Langner A, Cannon WR, Jarman KD, *et al.* Sequence optimization as an alternative to de novo analysis of tandem mass spectrometry data. *Bioinformatics* 2004;**20**:2296–304.

64 Zhang Z. De Novo Peptide Sequencing Based on a Divide-and-Conquer Algorithm and Peptide Tandem Spectrum Simulation. *Analytical Chemistry* 2004;**76**:6374–83.

65 Hines WM, Falick AM, Burlingame AL, *et al.* Pattern-based algorithm for peptide sequencing from tandem high energy collision-induced dissociation mass spectra. *Journal of the American Society for Mass Spectrometry* 1992;**3**:326–36.

66 Chong KF, Ning K, Leong HW, *et al.* Modeling and characterization of multi-charge mass spectra for peptide sequencing. *J Bioinform Comput Biol* 2006;**4**:1329–52.

67 Bandeira N, Tsur D, Frank A, *et al.* A New Approach to Protein Identification. In: *Research in Computational Molecular Biology*. Springer Berlin / Heidelberg 2006. 363–78.

68 Bandeira N, Olsen JV, Mann M, *et al.* Multi-spectra peptide sequencing and its applications to multistage mass spectrometry. *Bioinformatics* 2008;**24**:i416–23.

69 Olsen JV, Mann M. Improved peptide identification in proteomics by two consecutive stages of mass spectrometric fragmentation. *Proc Natl Acad Sci U S A* 2004;**101**:13417–22.

70 Kaufmann R, Kirsch D, Spengler B. Sequencing of peptides in a time-of-flight mass spectrometer: evaluation of postsource decay following matrix-assisted laser desorption ionisation (MALDI). *International Journal of Mass Spectrometry and Ion Processes* 1994;**131**:355–85.

71 Thompson MS, Cui W, Reilly JP. Fragmentation of singly charged peptide ions by photodissociation at lambda = 157 nm. *Angew Chem Int Ed Engl* 2004;**43**:4791–4.

72 Zhang L, Reilly JP. Peptide de Novo Sequencing Using 157 nm Photodissociation in a Tandem Time-of-Flight Mass Spectrometer. *Analytical Chemistry* 2010;**82**:898–908.

73 Datta R, Bern M. Spectrum Fusion: Using Multiple Mass Spectra for De Novo Peptide Sequencing. *Journal of Computational Biology* 2009;**16**:1169–82.

74 Horn DM, Zubarev RA, McLafferty FW. Automated de novo sequencing of proteins by tandem high-resolution mass spectrometry. *Proc Natl Acad Sci U S A* 2000;**97**:10313–7.

75 Zubarev RA, Zubarev AR, Savitski MM. Electron capture/transfer versus collisionally activated/induced dissociations: solo or duet? *J Am Soc Mass Spectrom* 2008;**19**:753–61.

76 Li X, Lin C, Han L, *et al.* Charge Remote Fragmentation in Electron Capture and Electron Transfer Dissociation. *Journal of the American Society for Mass Spectrometry* 2010;**21**:646–56.

77 Sreevatsa A, Badrunnisa S, Shaukath A, *et al.* Computational diagnostics based on proteomic data- review on approaches and algorithms. *Int J Binfo Res* 2010;**2**:56–66.

78 Shadforth I, Crowther D, Bessant C. Protein and peptide identification algorithms using MS for use in high-throughput, automated pipelines. *Proteomics* 2005;**5**:4082–95.

79 Kapp EA, Schutz F, Connolly LM, *et al.* An evaluation, comparison, and accurate benchmarking of several publicly available MS/MS search algorithms: sensitivity and specificity analysis. *Proteomics* 2005;**5**:3475–90.

80 Pitzer E, Masselot A, Colinge J. Assessing peptidede novo sequencing algorithms performance on large and diverse data sets. *Proteomics* 2007;**7**:3051–4.

81 Pevtsov S, Fedulova I, Mirzaei H, *et al.* Performance evaluation of existing de novo sequencing algorithms. *J Proteome Res* 2006;**5**:3018–28.

82 Bringans S, Kendrick TS, Lui J, *et al.* A comparative study of the accuracy of several de novo sequencing software packages for datasets derived by matrix-assisted laser desorption/ionisation and electrospray. *Rapid Commun Mass Spectrom* 2008;**22**:3450–4.

83 Tabb DL, Saraf A, Yates JR. GutenTag: high-throughput sequence tagging via an empirically derived fragmentation model. *Anal Chem* 2003;**75**:6415–21.

84 Searle BC, Dasari S, Turner M, *et al.* High-Throughput Identification of Proteins and Unanticipated Sequence Modifications Using a Mass-Based Alignment Algorithm for MS/MS de Novo Sequencing Results. *Analytical Chemistry* 2004;**76**:2220–30.

85 Frank A, Tanner S, Bafna V, *et al.* Peptide sequence tags for fast database search in mass-spectrometry. *J Proteome Res* 2005;**4**:1287–95.

86 Tabb DL, Ma Z-Q, Martin DB, *et al.* DirecTag: accurate sequence tags from peptide MS/MS through statistical scoring. *J Proteome Res* 2008;**7**:3838–46.

87 Kim S, Bandeira N, Pevzner PA. Spectral Profiles, a Novel Representation of Tandem Mass Spectra and Their Applications for de Novo Peptide Sequencing and Identification. *Molecular & Cellular Proteomics* 2009;**8**:1391–400.

88 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. *Anal Chem* 2001;**73**:1917–26.

89 Tanner S, Shu H, Frank A, *et al.* InsPecT: identification of posttranslationally modified peptides from tandem mass spectra. *Anal Chem* 2005;**77**:4626–39.

90 Mackey AJ, Haystead TA, Pearson WR. Getting more from less: algorithms for rapid protein identification with multiple short peptide sequences. *Mol Cell Proteomics* 2002;**1**:139–47.

91 Johnson RS, Taylor JA. Searching sequence databases via de novo peptide sequencing by tandem mass spectrometry. *Mol Biotechnol* 2002;**22**:301–15.

92 Bern M, Cai Y, Goldberg D. Lookup Peaks: A Hybrid of de Novo Sequencing and Database Search for Protein Identification by Tandem Mass Spectrometry. *Analytical Chemistry* 2007;**79**:1393–400.

93 Allmer J, Naumann B, Markert C, *et al.* Mass spectrometric genomic data mining: Novel insights into bioenergetic pathways in Chlamydomonas reinhardtii. *Proteomics* 2006;**6**:6207–20.

94 Allmer J, Markert CH, Stauber E J., *et al.* A new approach that allows identification of intron-split peptides from mass spectrometric data in genomic databases. *FEBS Lett* 2004;**562**:202–6.

95 Kim S, Gupta N, Bandeira N, *et al.* Spectral dictionaries: Integrating de novo peptide sequencing with database search of tandem mass spectra. *Mol Cell Proteomics* 2009;**8**:53–69.

96 Menschaert G, Vandekerckhove TTM, Baggerman G, *et al.* A Hybrid, de Novo Based, Genome-Wide Database Search Approach Applied to the Sea Urchin Neuropeptidome. *Journal of Proteome Research* 2010;**9**:990–6.

97 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.

98 Lu B, Chen T. A suffix tree approach to the interpretation of tandem mass spectra: applications to peptides of non-specific digestion and post-translational modifications. *Bioinformatics* 2003;**19 Suppl 2**:II113–21.

99 Alves G, Yu YK. Robust accurate identification of peptides (RAId): deciphering MS2 data using a structured library search with de novo based statistics. *Bioinformatics* 2005;**21**:3726–32.

100 Allmer J, Kuhlgert S, Hippler M. 2DB: a Proteomics database for storage, analysis, presentation, and retrieval of information from mass spectrometric experiments. *BMC Bioinformatics* 2008;**9**:302–13.

101 Tessier D, Yclon P, Jacquemin I, *et al.* OVNIp: An open source application facilitating the interpretation, the validation and the edition of proteomics data generated by MS analyses and de novo sequencing. *Proteomics* 2010;**10**:1794–801.

102 Naumann B, Busch A, Allmer J, *et al.* Comparative quantitative proteomics to investigate the remodeling of bioenergetic pathways under iron deficiency in Chlamydomonas reinhardtii. *Proteomics* 2007;**7**:3964–79.

103 Tannu N, Hemby S. De novo protein sequence analysis of Macaca mulatta. *BMC Genomics* 2007;**8**:270.

104 Edman P, Högfeldt E, Sillén LG, *et al.* Method for Determination of the Amino Acid Sequence in Peptides. *Acta Chem Scand* 1950;**4**:283–93.

105 Stegemann C, Kolobov A, Leonova YF, *et al.* Isolation, purification and de novo sequencing of TBD-1, the first beta-defensin from leukocytes of reptiles. *Proteomics* 2009;**9**:1364–73.

106 Ma M, Chen R, Ge Y, *et al.* Combining Bottom-Up and Top-Down Mass Spectrometric Strategies for De Novo Sequencing of the Crustacean Hyperglycemic Hormone from Cancer borealis. *Analytical Chemistry* 2009;**81**:240–7.

107 Ning K, Fermin D, Nesvizhskii AI. Computational analysis of unassigned high-quality MS/MS spectra in proteomic data sets. *Proteomics* 2010;**10**:2712–8.

108 Tharakan R, Edwards N, Graham DRM. Data maximization by multipass analysis of protein mass spectra. *Proteomics* 2010;**10**:1160–71.

109 Junqueira M, Spirin V, Balbuena TS, *et al.* Protein identification pipeline for the homology-driven proteomics. *Journal of Proteomics* 2008;**71**:346–56.

110 Domon B, Aebersold R. Challenges and Opportunities in Proteomics Data Analysis. *Molecular & Cellular Proteomics* 2006;**5**:1921–6.

111 Liu C, Song Y, Yan B, *et al.* Fast de novo peptide sequencing and spectral alignment via tree decomposition. *Pac Symp Biocomput* 2006;:255–66.

112 Searle BC, Dasari S, Wilmarth PA, *et al.* Identification of protein modifications using MS/MS de novo sequencing and the OpenSea alignment algorithm. *J Proteome Res* 2005;**4**:546–54.

113 Zhong H, Li L. An algorithm for interpretation of low-energy collision-induced dissociation product ion spectra for de novo sequencing of peptides. *Rapid Commun Mass Spectrom* 2005;**19**:1084–96.