Technology Research and Development Project: Universal Tools for Peptide Identification and Sequencing (Bandeira, Pevzner)

1 Specific Aims

Mass spectrometry (MS) instruments and experimental protocols have greatly advanced over the last decade. Several new fragmentation technologies like ETD [55] and HCD [47] emerged and high-precision mass spectrometers became widely available. While trypsin remains a dominant protease in proteomics studies, digesting proteins with diverse proteases is becoming popular [53]. Empowered by these changes, MS researchers now have diverse choices with respect to the questions: "what fragmentation method to use?", "how accurate should be the measurements of the mass-to-charge (m/z) ratios?", "what proteases to use?", and "what post-translational modification (PTM) to focus on (e.g. phosphorylation)?". Depending on these choices, the resulting spectra vary in fragmentation propensities and precision. Therefore, unlike in the past when low-precision CID spectra of tryptic peptides dominated the field, spectral datasets generated today are very diverse.

Unfortunately, popular MS/MS database search tools such as SEQUEST [11] and Mascot [49] have not kept pace with the increased diversity of the data because they are largely optimized for low-precision CID spectra of tryptic peptides [43]. Reflecting this concern, Noble and MacCoss pointed out in a recent review that "the field (of MS) is still missing a generic analysis platform that can be adapted automatically and in a principled fashion to handle spectra produced by any given fragmentation protocol" [46]. Moreover, since different laboratories employ different combinations of tools (see Figure 1), even for the same data, capabilities of analyzing the data vary widely and results obtained in one laboratory are often impossible to reproduce in another laboratory [61]. This creates a proteomics version of the Tower of Babel when the tools for solving the basic proteomics problems may become so diverse that different laboratories may loose ability to speak the same computational language. Moreover, the emergence of specialized tools for analyzing specific modifications (e.g., phosphorylation, ubiquitination, etc.) make the situation even more complex raising the question whether, lets say, three different phosporylation-specific tools should be developed for CID, ETD, and HCD spectra.

We advocate using *universal* database search tools that perform well for diverse types of spec-

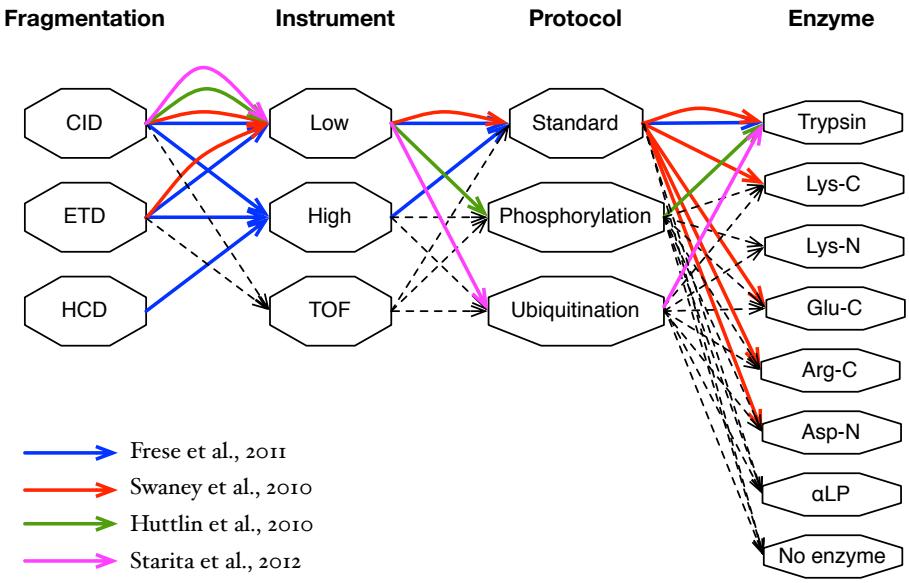


Figure 1: Spectral types as paths in the graph representing possible choices of the fragment method (Fragmentation), the instrument measuring product ion m/z (Instrument), the protocol used to prepare a sample (Protocol), and the enzyme used to digest proteins (Enzyme). 'Low' in Instrument indicates low-resolution instruments, 'High' indicates high-resolution instruments, and 'TOF' indicates time-of-flight instruments. A path in the graph represents a spectral type. For example, the green path (CID,Low,Phosphorylation,Trypsin) represents low-precision CID spectra of trypsin digests generated from a sample enriched for phosphopeptides. The blue, red, green, and magenta paths represent spectral types of the datasets used in recent studies by Frese et al. [14], Swaney et al. [53], Huttlin et al. [24], and Starita et al. [51]. Different combinations of analysis tools were used for different studies. Frese et al. used an in-house tool for peak filtering, de-isotoping, and charge deconvolution, Mascot for database search, Percolator for re-scoring, and RockerBox [59] for peptide-level FDR control. Swaney et al. used an in-house tool for peak filtering, OMSSA [16] for database search, and an in-house tool for both peptide- and protein-level FDR control. Huttlin et al. used an in-house tool for re-calibrating peak masses, SEQUEST for database search, an in-house tool for re-scoring, and peptide- and protein-level FDR control. Starita et al. used the Trans-Proteomics Pipeline [10] along with SEQUEST for database search.

tral datasets. To address this need, we propose to develop peptide identification tool UniQuest and peptide *de novo* sequencing too UniNovo that work well for spectra generated using diverse configurations of MS instruments and experimental protocols. We emphasize that we do not intend to customize these tools for specific spectral datasets but rather to develop a robust probabilistic model that works well across all datasets. Moreover, in contrast to the previous funding cycles (when we invested into development of modification-specific tools) we are not against developing any modification-specific tools. Instead, we will develop a universal tool that works well across *all* modifications.

1.1 Aim 1: Universal peptide identification tool

We aim to develop a universal MS/MS database search tool UniQuest that will *automatically* derive scoring parameters from a training sample of PSMs *without* prior knowledge of the type of the spectra [35]. While the four recent studies highlighted in Figure 1 are very difficult to reproduce (due to multitude of pre-processing and post-processing techniques used in each of these studies), UniQuest will use the same code base (that does not require additional pre– and post-processing) for different spectral types and will be very easy to reproduce. For each spectral type, UniQuest will *automatically* divide the PSMs into subgroups (depending on the precursor ion charge and m/z). and will *automatically* learn a different set of scoring parameters depending on the spectral type and the subgroup. UniQuest should be able to train scoring parameters for any spectral type (including spectral types not specified in Figure 1) or use pre-trained scoring parameters. UniQuest will thus address the goal of designing modification-specific MS/MS database searches for dozens of different modifications. It wil also provide an user interface, taking over the authority to train scoring parameters to the users and making training as easy as running a database search.

1.2 Aim 2: Universal peptide sequencing tool

While existing *de novo* sequencing tools perform well on certain types of spectra (e.g., CID spectra of tryptic peptides), their performance often deteriorates on other types of spectra, such as ETD, HCD spectra, or spectra of non-tryptic digests. Thus, rather than developing a new algorithm for each type of spectra, we propose to develop a *universal de novo* sequencing algorithm UniNovo that works well for all types of spectra or even for spectral pairs (e.g., CID/ETD spectral pairs). The

scoring function of UniNovo will be easily trainable using a small training dataset of PSMs. While, in difference from peptide *identification* tools, peptide *sequencing* tools have rarely been subjected to a rigorous statistical significance analysis in the past, UniNovo will estimate the error rate of the *de novo* reconstructions. UniNovo will generate *gapped* peptide reconstructions that will be used as seeds for time-consuming blind MS/MS database searches using MS-BPM software developed at CCMS.

1.3 Aim 3: Universal tool for identification of cross-linked peptides

2 Significance

2.1 Significance and challenges in peptide identification

While peptide identification tools represent a workshorse of modern proteomics, the algorithms behind these tools are far from being perfect and require further developments. All these tools tools use a scoring function $\text{Score}(P, S)$ to evaluate a PSM formed by a peptide P and a spectrum S and further compute statistical significance (e.g. E-values) of the resulting PSMs. Let P_S be a (correct) peptide that generated S. A scoring function is *adequate* for S (with respect to a protein database $ProteinDB$) if the correct peptide attains the maximal score in the database, i.e., $\max_{P \in ProteinDB} \text{Score}(P, S) = \text{Score}(P_S, S)$. A "good" scoring function should satisfy the following conditions:

(a) It should be adequate for the great majority of spectra (*adequate* scoring),

(b) the algorithm for PSM scoring should be fast (*efficient* scoring),

(c) the algorithm for computing statistical significance of individual PSMs should be fast and accurate (*statistically sound* scoring),

When CCMS started, there were no statistically sound peptide identification tools. In 2008-2012, CCMS invested significant efforts in developing adequate, efficient, and statistically sound tool MS-GF [**?**, **?**, 33, 18, 36, 19]. Most CCMS projects use a very simple dot-product scoring $Score(P, S) = P^* \cdot S^*$ after converting peptide P and spectrum S into vectors P^* and S^* referred

to as *peptide vector* and *spectral vector*, respectively. This scoring approach (implemented in MS-GF and other CCMS software software tools) recently gained popularity outside CCMS and is now being actively used at PNNL and other laboratories.

Conversion of a spectrum S into a spectral vector S^* in MS-GF uses a probabilistic model that ensures that the resulting dot-product scoring is adequate [31] (condition (a)). At the same time, it makes scoring and computing accurate E-values fast [32] (condition (b) and (c)). This simple "spectral vector" scoring model contrasts with many database search [11, 16, 7, 6] and re-scoring [29, 28] tools, using sophisticated scoring functions that often make it difficult to satisfy the conditions (b) and/or (c). However, MS-GF approach currently has limitations (e.g., it does not satisfy conditions b) and/or c) for modified peptides and high-precision spectra) that we plan to address in the new cycle. In particular, UniQuest, in addition to conditions (a), (b), and (c), will satisfy the following conditions on the scoring function:

(A) it should be automatically trainable for *all* spectral types (*universal* scoring),

(B) it should be statistically sound with respect to both unmodified and modified peptides (*modification-aware* scoring),

(C) it should be statistically sound with respect to both low and high precision spectra (*precision-aware* scoring).

While it may appear that extending the generating function approachfrom from (i) unmodified to modified peptides, and from (ii) low to high precision spectra, is a simple matter of modifying the parametric dynamic programming algorithm from [32] and tuning parameters that control the error tolerance, the situation is much more complex (see the "Approach" section) Since neither of existing peptide identification tools satisfies the conditions (A), (B), and (C), addressing them in UniQuest will address the important goal of making MS/MS searches reproducible and using the same code base across diverse technologies and experimental protocols. We hope that these developments will help to prevent the "Tower of Babylon" in proteomics and will contribute to developing universal tools that can be seemlessly used across various datasets and laboratories.

Below we briefly address the significance of these new developments.

2.2 Significance and challenges in peptide sequencing

De novo peptide sequencing is often viewed as a somewhat obscure alternative to MS/MS database search in the case when the peptide of interest is not present in the database. In contrast, CCMS views *de novo* peptide sequencing as an extemeley important area that affect nearly all CCMS tools. In particular, it is a crucial part of MS-GF and InsPect (restrictive MS/MS database search), MODa and MS-Align (blind MS/MS database search), MS-SpecNets (spectral networks), MS-SPS (shotgun protein sequencing), and other tools. Thus, our developments of *de novo* tools have downstream effect on many other CCMS tools.

In contrast to peptide identification that utilizes the information from proteome, the *de novo* peptide sequencing approach attempts to identify peptides only using the information from the input spectrum. Hence, most *de novo* sequencing algorithms are based on the prior knowledge of the fragmentation characteristics (e.g., ion types and their propensities) of MS/MS spectra [41, 13, 12]. While the fragmentation characteristics of CID have been well studied [26, 60, 4, 56, 22, 2] (PEAKS [41] and PepNovo+ [13, 12], are the state of the art sequencing tools for CID spectra), the existing knowledge about fragmentation characteristics of recently introduced fragmentation methods, such as ETD and HCD, have not been fully utilized in existing peptide sequencing tools yet.

The recently emerged fragmentation methods like ETD and HCD have a great potential for *de novo* sequencing. For example, for highly charged spectra, ETD provides better fragmentation and thus is better suited for *de novo* sequencing than CID [63, 52]. Also, more complete fragmentation (especially in low mass regions) in HCD provides a better chance to obtain more accurate *de novo* reconstructions than CID [**?**, 5]. Furthermore, modern mass spectrometers allow the generation of paired spectra (e.g., CID/ETD or HCD/ETD spectral pairs). Since CID, HCD, and ETD spectra provide complementary information for peptide sequencing [50, 9, 20], such spectral pairs enable more accurate *de novo* sequencing.

Several recently proposed *de novo* sequencing algorithms take advantage of new fragmentation methods. For instance, PEAKS [39] and pNovo [5] include *de novo* sequencing algorithm for ETD and HCD spectra, respectively. Recently, [9] and [20] presented Spectrum Fusion and ADEPTS *de novo* sequencing algorithms for CID/ETD spectral pairs. While the above tools perform well for the

spectra generated from the fragmentation method(s) that each tool targeted, they often generate inferior results for the spectra from other fragmentation methods. Moreover, if alternative proteases (e.g., LysC or AspN) are used for protein digestion, these tools often produce suboptimal results because different proteases often generate peptides with different fragmentation propensities [**?**].

Since *de novo* sequencing tools are tightly coupled with a multitude of CCMS tools, it is impractical for CCMS to develop specialized tools for every existing (let alone newly emerging) fragmentation technique or an experimental protocol. Thus, availability of a universal *de novo* sequencing tool is an important future goal for CCMS. As with any universal tool, there is a concern whether it will be able to compete with specialized tools that are aimed at specific fragmentation techniques. To address this concern, we implemented a pilot version of UniNovo and conducted an initial benchmarking with state-of-the-art specialized tools PepNovo+, PEAKS, and pNovo. The results show that the performance of UniNovo is superior to other tools for ETD spectra and superior or comparable to others for CID and HCD spectra.

3 Innovation

3.1 Innovation in peptide identification

UniQuest will address the following limitations of most peptide identification tools (including MS-GF [32] and MS-GFDB [35]): inability to estimate E-values accurately for PSMs formed by modified peptides and limited ability to take advantage of accurate m/z values in high-precision MS/MS spectra. In addition, it will greatly reduce the running time by turning the traditional "Spectra against Peptides" searches into "Peptides against Spectra" searches (see Figure 2). The advantage of this approach is that it allows one to use the *suffix arrays* [42], a powerful combinatorial pattern matching technique. UniQuest will also feature an improved usability due to ProteoSAFe, a user-frendly interface for searches, reports and data management.[1] UniQuest will support the HUPO Proteomics Standard Initiative standard file formats [27].

The key innovation in designing the precision-aware scoring is a new generative model that describes how a peptides generates a spectrum. Kim et al., 2009 [31] introduced an abstract model

[1]To-ju Huang, Claudiu Farcas, Jeremy Carver, Natalie Castellana, Ari Frank, Sangtae Kim, Jian Wang, Pavel A. Pevzner, Vineet Bafna, Ingolf Krüger, and Nuno Bandeira. ProteoSAFe: A Scalable, Accessible, and Flexible Software Environment for Proteomics Analysis, *in preparation*.

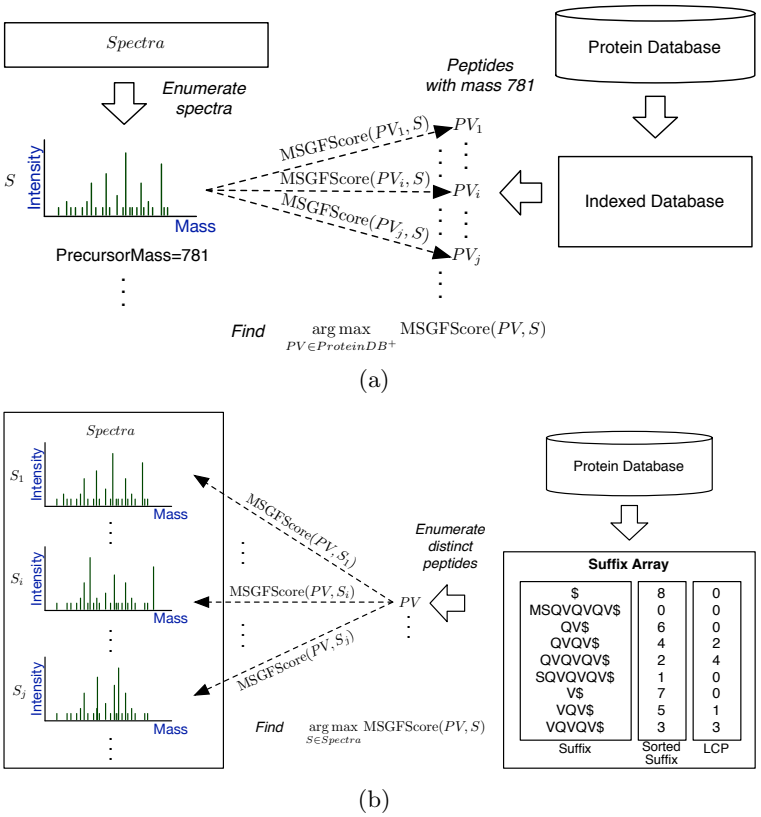


Figure 2: Two approaches for comparing spectra against a protein database. (a) Traditional spectrum-based approach compares each spectrum against all peptides. (b) UniQuest uses the suffix array to compare each peptide against all spectra with the same precursor mass.

that described a probabilistic process of transforming a Boolean string (peptide) into another Boolean string (spectrum). This simple model was further extended to real spectra and proved to be valuable for designing MS-GF peptide identification tool. However, this model, while adequate for low-precision spectra, needs to be modified for high-precision spectra. UniQuest models a peptide as a boolean string (as before) but model a spectrum as a directed acyclic graph (DAG) called *G-spectrum* and further applies a transformation of a Boolean string into a *G*-spectrum DAG for scoring real PSMs. In this G-spectrum representation, vertex labels encode the information on the *intensities* of individual peaks, and the edge labels encode the information on the *mass errors* of pairs of peaks assuming they represent consecutive peaks of the same ion type.

3.2 Innovation in peptide sequencing

UniNovo will represent the first universal *de novo* peptide sequencing tool that will work across *all spectral types* (e.g., the combination of the fragmentation method, the instrument type, and the protease used to digest sample proteins). The scoring function of UniNovo will be easily trainable using a small training dataset consisting of a few thousands of annotated spectra (PSMs). All information needed for *de novo* sequencing will be learned from the training dataset.

One of the biggest challenges in *de novo* sequencing is to estimate the error rate of the resulting *de novo* reconstructions. Unlike peptide identification tools *de novo* peptide sequencing tools have rarely been subjected to a statistical significance analysis in the past. Several *de novo* sequencing tools report the error rate of amino acid predictions (e.g. confidence scores in PEAKS), but this is often not sufficient because the overall quality of the sequence cannot be easily determined by the error rates of individual amino acid predictions. Estimating the probability that the entire reconstructed peptide (rather than individual amino acids) is correct is crucially important for downstream applications (e.g., blind database search) since a single error in the reconstruction may lead to a failure of such an application.

To our knowledge, only PepNovo+ [12] reports the empirical probability that the output peptide is correct (using logistic regression with multiple features of the reconstructions). However, PepNovo+ does not include an automated training procedure (that would allow to easily extend PepNovo+ for newly emerging MS approaches) and is currently trained only for CID. Since PepNovo+ was not originally designed as a universal tool, extending PepNovo+ beyond CID spectra requires training complex boosting-based models for predicting peak ranks and rescoring peptide candidates. Currently, PepNovo+ training includes several manual steps and requires large corpus of training spectra (a few hundred thousands of PSMs). Thus, in case of non-CID fragmentation methods, it remains unclear how to obtain accurate error rate estimation for *de novo* reconstructions using PepNovo. We thus decided to develop a principally new universal tool UniNovo rather than to extend PepNovo into a truly universal tool. UniNovo will estimate the probability that each reported reconstruction is correct, using simple statistics that are readily obtained from a small training dataset.

4 Approach

4.1 Aim 1: Peptide identification

UniQuest takes a spectral dataset *Spectra* and a protein database *ProteinDB* as an input and outputs a set of scored PSMs along with statistical significance estimates. The workflow of MS-GF+ comprises the following 4 steps: generating spectral vectors, searching a protein database, computing E-values of PSMs, and estimating FDRs.

To make UniQuest statistically sound, we adopt the generating function approach to rigorously compute E-values of PSMs using the score distribution of *all peptides* [32]. The scores of PSMs reported by existing peptide identification tools are often poorly correlated with their statistical significance (e.g., E-values) [17]. It is important to rank PSMs based on their statistical significance, because such ranking (rather than ranking based on "raw scores") often dramatically increases the number of identified spectra [32, 37]. This observation explains why it is important to make UniQuest statiscally sound (condition (c)). The spectral vector scoring model described below is essential here, because the generating function approach is easily applicable to the scoring functions that can be represented as a dot-product of vectors [19].

Below we describe each step and explain how to address the challenge of implementing universal as well as modification-aware and precision-aware MS/MS database search tool.

4.1.1 Generating spectral vectors

Rescaling mass values A spectral vector [35, 57] of a spectrum S is an M-dimensional vector with integer values where M is the nominal parent mass of S. Transformation of experimental spectra into spectral vectors requires *binning*. To support this transformation, the generating function approach was designed under the assumption that amino acid masses are integers. While this enables efficient computing of scores, it causes rounding errors because peaks in the spectrum correspond to *real* rather than *nominal* masses of amino acid sequences. To minimize the rounding errors, we *rescale* every mass m into $0.9995m$ [31, 58, 3]. This dramatically reduces the rounding errors, so one can estimate the nominal mass of a peptide (or a peak) of mass m by simply taking $[0.9995m]$ where $[x]$ represents the closest integer to x. We investigated all 188 million unique peptides of length up to 20 in the human IPI database and the estimation was inaccurate only for

874 peptides (4.6×10^{-6} %).

Peptide variants A (non-modified) *peptide* is defined as a string over the alphabet $\mathcal{A}$ of 20 standard amino acids. UniQuest is a restrictive MS/MS database search tool that allows a user to specify a set of allowed modifications. Let $\mathcal{A}^+$ be an *extended* amino acid set containing both unmodified and modified amino acids. For an (unmodified) amino acid $a \in \mathcal{A}$, let $\text{Mod}(a) \subset \mathcal{A}^+$ be the *set* of both unmodified and modified amino acids associated with a. For example, if T (Thr) and T^* (phosphorylated Thr) are in $\mathcal{A}^+$, $\text{Mod}(T) = \{T, T^*\}$. Given a peptide $P = a_1 \dots a_k$, define $PV = pv_1 \dots pv_k$ as a *variant* of P if $pv_i \in \text{Mod}(a_i)$ for all i $(1 \leq i \leq k)$.

Incorporating ion types and ranks into scoring The conversion from an experimental spectrum to a spectral vector proceeds as follows. A *spectrum* $S = \{(mz_1, rank_1), \dots, (mz_l, rank_l)\}$ is represented as a set of *ranked* peaks where the ith highest intensity peak gets rank i (mz_j and $rank_j$ represent m/z and rank of jth peak, respectively). An *ion type* is represented as a triplet of integers *charge, offset*, and *sign*, where sign represents whether the ion type is a prefix ion ($sign = 1$) or a suffix ion ($sign = -1$). For example, singly-charged b-ions and y-ions correspond to ion types (1, 1, 1) and (1, 19, -1), respectively. Given an ion type $ion = (charge, offset, sign)$, one can turn a spectrum S of nominal parent mass M into $S_{ion} = \{(mass_1, rs_1), \dots, (mass_l, rs_l)\}$ using the following transformation:

$$mass_j = \begin{cases} [mz_j \cdot charge \cdot 0.9995] - offset & \text{if } sign = 1 \\ M - \big([mz_j \cdot charge \cdot 0.9995] - offset\big) & \text{if } sign = -1 \end{cases}$$

$$rs_j = \text{RankScore}(ion, rank_j),$$

where $\text{RankScore}(ion, rank)$ is a pre-computed function that takes an ion type ion and an integer $rank$ and returns a probabilistic log-likelihood score defined in [35, 31]. Assume that $\mathcal{I}$ is a set of ion types contributing to scoring. The spectral vector of S (denoted by $\vec{S} = (s_1, \dots, s_M)$ is computed as follows:

$$s_i = \sum_{ion \in \mathcal{I}} \max(\{rs \mid (mass, rs) \in S_{ion} \text{ and } mass = i\} \cup \text{RankScore}(ion, \infty)),$$

where RankScore(ion, ∞) represents the score given when ion is missing.

We also define a peptide vector of a variant as follows. Let Mass(a) be the nominal mass of a (possibly modified) amino acid a. For example, Mass$(T) = 101$ and the mass of phosphorylated Thr is Mass$(T^*) = 181$. Given a variant $PV = pv_1 \dots pv_k$, define the mass of PV as Mass$(PV) = \sum_{i=1}^{k}$ Mass(pv_i). Given a variant $PV = pv_1 \dots pv_k$ of mass M, we define its peptide vector (denoted by $\vec{PV}$) as a 0-1 vector $(m_1, \dots, m_M)$ with $(n-1)$ 1s, such that $m_i = 1$ if i equals to Mass$(pv_1) + \dots + Mass(pv_j)$ $(1 \leq j \leq k)$.

Score of a PSM (PV, S) is defined as MSGFScore$(PV, S) = \vec{PV} \cdot \vec{S}$ if Mass(PV) = ParentMass(S) and $-\infty$ otherwise. The score represents the log likelihood ratio described in [31].

Searching a protein database with suffix arrays We define $ProteinDB^+$ as the set of all variants (with respect to an extended amino acid set $\mathcal{A}^+$) derived from $ProteinDB$. We aim to solve the following problem: given a spectral dataset $Spectra$ and a protein database $ProteinDB$, for each spectrum $S \in Spectra$ find a variant $PV_{S,ProteinDB}$ such that

$$PV_{S,ProteinDB} = \underset{PV \in ProteinDB^+}{\arg\max} \text{MSGFScore}(PV, S).$$

Solving this problem involves the following three steps: (1) for every spectrum $S \in Spectra$, computing $\vec{S}$, (2) for every variant $PV \in ProteinDB^+$, computing $\vec{PV}$, and (3) for every pair of (PV, S) where Mass(PV) = ParentMass(S), computing MSGFScore$(PV, S) = \vec{PV} \cdot \vec{S}$. To execute these steps efficiently, one may simply execute the step (1) and (2), store all $\vec{S}$ and $\vec{PV}$ in the main memory and execute the step (3). But this is often infeasible because the number of variants is usually too large to fit all $\vec{PV}$ in the main memory. Alternatively, one may consider executing the step (2) on the spot for each spectrum, but this is prohibitively slow.

Instead of storing both $\vec{S}$ and $\vec{PV}$, we propose to store only $\vec{S}$ for all spectra in the main memory, and index them by parent masses. Since spectral vectors compactly represent experimental spectra, they can be stored in a memory-efficient manner (4GB per 200,000 spectra). Rather than finding the best scoring peptide for each spectrum, we propose to find the best scoring spectrum for each variant. Formally, UniQuest solves a slightly different problem: for each variant $PV \in$

$ProteinDB^+$, find a spectrum $S_{PV,Spectra}$ such that

$$S_{PV,Spectra} = \arg\max_{S \in Spectra} \text{MSGFScore}(PV, S) = \arg\max_{S \in Spectra_{\text{Mass}(PV)}} \text{MSGFScore}(PV, S), \quad (1)$$

where $Spectra_{\text{Mass}(PV)}$ represents the set of spectra $S \in Spectra$ with ParentMass(S) = Mass(PV). This problem can be solved efficiently by enumerating variants $PV \in ProteinDB^+$ one by one, generating $\vec{PV}$ *on the spot*, and computing MSGFScore$(PV, S) = \vec{PV} \cdot \vec{S}$ for all *pre-computed* $\vec{S}$ where $S \in S_{\text{Mass}(PV)}$.

Similar to pFind [62], we will use a *suffix array* [42] of $ProteinDB$ to further optimize the database search. Instead of searching peptides according to their ordering in the original database file, UniQuest will search peptides according to their ordering in the suffix array, and use the longest common prefix data structure [42] to score each unique peptide only once (Figure 2 (b)).

4.1.2 Designing modification-aware scoring

While Kim et al., 2008 [32] described the generating function algorithm for computing *exact* p-values in the case of unmodified peptides, it is unclear how to extend this approach to modified peptides.

Given a spectrum S, a score threshold t, an extended set of amino acids $\mathcal{A}^+$, and a database size N, we define E-value$(S, t, \mathcal{A}^+, N)$ as the expected number of variants PV (as defined by $\mathcal{A}^+$) with MSGFScore$(PV, S) \geq t$ in a random protein database of size N. To compute E-value$(S, t, \mathcal{A}^+, N)$, we first compute *spectral E-value* E-value$(S, t, \mathcal{A}^+)$, the expected number of variants PV with MSGFScore$(PV, S) \geq t$ given a *single random peptide*. A single random peptide models a random peptide starting at a fixed position in a random protein database.

We consider a set of all possible (unmodified) peptides of length k (where k is a large number) and select a random peptide uniformly from this set (i.e. the probability of selecting a peptide is $\frac{1}{20^k}$). We say that a peptide P *produces* a variant PV if PV is a variant of a prefix of P. For example, $PEPT^*$ and $PEPTI$ are produced by $PEPTIDE$. Given a spectrum S, let $\mathcal{PV}(t)$ be the set of all variants PV with MSGFScore$(PV, S) \geq t$. For every variant PV, there are $20^{k-|PV|}$ peptides of length k producing a variant PV ($|PV|$ stands for the number of amino acids in PV).

Therefore, expected number of variants per random peptide with a score equal or better than t is

$$\text{E-value}(S, \mathcal{A}^+, t) = \sum_{PV \in \mathcal{PV}(t)} \frac{20^{k-|PV|}}{20^k} = \sum_{PV \in \mathcal{PV}(t)} 20^{-|PV|}.$$

Since a variant is a string over the alphabet $\mathcal{A}^+$, this expression can be computed using the generating function approach [32]. Given a spectrum S with $\vec{S} = s_1 \dots s_M$, consider an *amino acid graph* $G(V, E, \mathcal{A}^+)$ with $V = \{0, \dots, M\}$ and $E = \{(i, j) | j - i \in \text{Mass}(a) \text{ for } a \in \mathcal{A}^+\}$, where the *score* of a vertex i is defined as s_i, the *probability* of an edge is defined as $\frac{1}{20}$, the score of a path is defined as the sum of scores of its vertices, and the *probability* of a path is defined as the product of probabilities of its edges. A path in an amino acid graph represents a variant. Therefore, E-value$(S, \mathcal{A}^+, t)$ equals to the sum of probabilities of all paths from 0 to M with scores equal or better than t, and can be computed using parametric dynamic programming [32, 31].

While spectral E-values are useful for evaluating statistical significance of individual PSMs (independently of the database), they need to be transformed into E-value$(S, t, \mathcal{A}^+, N)$ to take into account the fact that the database search represents "multiple testing" where multiple variants (arising from different database peptides) are scored against a spectrum [45]. E-values can be approximated as follows:

$$\text{E-value}(S, t, \mathcal{A}^+, N) \approx \text{E-Value}(S, t, \mathcal{A}^+) \cdot N,$$

where N is the size of the database.

Existing MS/MS database search tools output a set of PSMs and estimate the FDR of this set using the *Target-Decoy Approach* (TDA). Using TDA to computing FDR for *each type* of modification encountered in the search remains an open problem. In addition to computing the FDR via TDA, our approach also provide a possibility to estimate the FDR via E-values of PSMs without using TDA [19].

4.1.3 Designing precision-aware scoring

Extending the generating function approach to high-precision spectra Mass spectrometers are usually divided into High-precision (denoted by H) and Low-precision (denoted by L)

instruments. Depending on whether the precursor and product ions are measured with Low or High-precision, the spectra are divided into LL, LH, HL, and HH spectra (LH spectra are hardly ever used in proteomics studies).

The key part of the generating function approach is the assumption that amino acids have integer masses [32]. However, rounding amino acid masses into integers introduces errors. These rounding errors reduce after rescaling by 0.9995 making them appropriate for LL and HL spectra. However, for HH spectra, the rounding errors remain too large even after rescaling, prohibiting UniQuest from benefiting from precise product ion peaks. A possible solution to this problem is to change the constant used for rescaling. For example, for a scaling constant 274.335215 (e.g. $\text{mass}(G) = 57.021464 \times 274.335215 = 15642.995586 \approx 15643$), the rounding error is bounded by 2.5 parts per million (ppm), which is appropriate for analyzing HH spectra. However, since the time complexity of the generating function algorithm is proportional to 1 over the rescaling constant, this makes computing E-values prohibitively slow. UniQuest will use a new scoring algorithm taking advantage of accurate product ion masses while not substantially increasing the running time.

Database search of high-precision spectra Let $\text{RMass}(a)$ be the real mass of an amino acid a. For a variant $PV = pv_1 \dots pv_k$, let $\text{RMass}(PV) = \sum_{i=1}^{k} \text{RMass}(pv_i)$, and $\text{RParentMass}(S)$ be the real parent mass of a spectrum S. We previously defined $\text{MSGFScore}(PV, S) = \vec{PV} \cdot \vec{S}$ if $\text{Mass}(PV) = \text{ParentMass}(S)$ and $-\infty$ otherwise. Note that the condition $\text{Mass}(PV) = \text{ParentMass}(S)$ while appropriate for LL spectra is weak for HL and HH spectra, because it may be satisfied even when real mass $\text{RMass}(PV)$ significantly deviates (e.g. up to 0.5 Da) from $\text{RParentMass}(S)$. To take advantage of accurate parent masses in HL and HH spectra, this condition has to be redefined to $\text{RMass}(PV) - \Delta < \text{RParentMass}(S) < \text{RMass}(PV) + \Delta$, where Δ is the precursor mass tolerance. To solve the database search problem for this modified definition of MSGFScore, the equation 1 should be changed as follows:

$$S_{PV,Spectra} = \underset{S \in Spectra}{\arg\max}\ \text{MSGFScore}(PV, S) = \underset{S \in Spectra_{\text{RMass}(PV)}}{\arg\max}\ \text{MSGFScore}(PV, S), \qquad (2)$$

x \ y	0	1
0	θ	$1-\rho$
1	$1-\theta$	ρ

(a)

x \ y,z	0,0	0,1	1,0	1,1
0	β_1	β_2	β_3	$1-\alpha$
1	$1-\beta_1$	$1-\beta_2$	$1-\beta_3$	α

(b)

Figure 3: **(a)** Probability $\text{Prob}_{\text{V}}(x|y)$ of a peptide character y generating a vertex label x. **(b)** Probability $\text{Prob}_{\text{E}}(x|y,z)$ of peptide characters y and z generating an edge label x.

where $Spectra_{\text{RMass}(PV)}$ represents the set of spectra $S \in Spectra$ satisfying $\text{RMass}(PV) - \Delta < \text{RParentMass}(S) < \text{RMass}(PV) + \Delta$.

From spectral vectors to G-spectra Kim et al., 2009 [31] introduced an abstract model that described a probabilistic process of transforming a Boolean string (peptide) into another Boolean string (spectrum). This simple model was further extended to real spectra and proved to be valuable for designing MS-GF peptide identification tool. However, this model, while adequate for low-precision spectra, needs to be modified for high-precision spectra. Below, we model a peptide as a boolean string (as before) but model a spectrum as a directed acyclic graph (DAG) called *spectral DAG* and further apply a transformation of a Boolean string into a spectral DAG for scoring real PSMs.

Let $P = p_1 \dots p_M$ be a Boolean string called a *peptide*. Let $G_S = (V, E, \mathcal{A}^+)$ be a labeled DAG called a *G-spectrum* with vertices $V = \{0, \dots, M\}$, and edges $E = \{(i,j) | j - i \in \text{Mass}(a) \text{ for } a \in \mathcal{A}^+\}$. We define the Boolean label of vertex i as v_i and edge (i,j) as $e_{i,j}$. The probability of a peptide P generating a G-spectrum G_S is defined as follows:

$$\text{Prob}(G_S|P) = \prod_{i \in V} \text{Prob}_{\text{V}}(v_i|p_i) \cdot \prod_{(i,j) \in E} \text{Prob}_{\text{E}}(e_{i,j}|p_i, p_j),$$

where $\text{Prob}_{\text{V}}(x|y)$ is a 2×2 matrix representing the probability of a peptide character (0 or 1) generating a vertex label, and $\text{Prob}_{\text{E}}(x|y,z)$ is a 2×4 matrix representing the probability of a pair of peptide characters generating an edge label (Figure 3). In practice, $\beta_1 \approx \beta_2 \approx \beta_3$.

When applying this model for scoring a peptide P and a G-spectrum G_S, we consider a test comparing two hypotheses: one assuming G_S is generated by P and the other assuming G_S is generated by a string consisting of all zeros (denoted by O). The score of (P, G_S) (denoted by

$\text{Score}(P, G_S))$ is defined as follows.

$$\begin{aligned}
\text{Score}(P, G_S) &= \log \frac{\text{Prob}(G_S|P)}{\text{Prob}(G_S|O)} \\
&= \log \frac{\prod_{i \in V} \text{Prob}_{\text{V}}(v_i|p_i) \cdot \prod_{(i,j) \in E} \text{Prob}_{\text{E}}(e_{i,j}|p_i, p_j)}{\prod_{i \in V} \text{Prob}_{\text{V}}(v_i|0) \cdot \prod_{(i,j) \in E} \text{Prob}_{\text{E}}(e_{i,j}|0,0)} \\
&= \sum_{i \in V} \log \frac{\text{Prob}_{\text{V}}(v_i|p_i)}{\text{Prob}_{\text{V}}(v_i|0)} + \sum_{(i,j) \in E} \log \frac{\text{Prob}_{\text{E}}(e_{i,j}|p_i, p_j)}{\text{Prob}_{\text{E}}(e_{i,j}|0,0)} \\
&\approx \underbrace{\sum_{i \in \{i | i \in V, p_i = 1\}} \underbrace{\log \frac{\text{Prob}_{\text{V}}(v_i|1)}{\text{Prob}_{\text{V}}(v_i|0)}}_{\text{VertexScore}(i)}}_{\text{vertex scoring}} + \underbrace{\sum_{(i,j) \in \{(i,j) | (i,j) \in E, p_i = 1, p_j = 1\}} \underbrace{\log \frac{\text{Prob}_{\text{E}}(e_{i,j}|1,1)}{\text{Prob}_{\text{E}}(e_{i,j}|0,0)}}_{\text{EdgeScore}(i,j)}}_{\text{edge scoring}}
\end{aligned} \tag{3}$$

Converting a spectrum into a G-spectrum We now explain how to convert a spectrum S into a G-spectrum given $\mathcal{A}^+$ and $\mathcal{I}$. Vertex and edge sets are constructed as described earlier. For simplicity, suppose that $\mathcal{I} = \{(1, 0, 1)\}$ (i.e. only singly charged prefix ions with an offset zero contribute to the scoring). Given a constant δ called a *fragment mass tolerance*, two peaks of S with m/z x and y form a *duo* if $y - x$ is approximately equal to a mass of an amino acid, i.e., $\text{RMass}(a) - \delta < y - x < \text{RMass}(a) + \delta$ for $a \in \mathcal{A}^+$. The vertex label v_i and the edge label $e_{i,j}$ of G_S are defined as follows: $v_i = 1$ if there exists a peak of mass x satisfying $[0.9995 \cdot x] = i$ and $v_i = 0$ otherwise; $e_{i,j} = 1$, if there exists a duo of peaks with masses x and y such that $[0.9995 \cdot x] = i$ and $[0.9995 \cdot y] = j$, and $e_{i,j} = 0$ otherwise.

In practice, we generate multiple *G-spectra* for a single spectrum, one for each $ion \in \mathcal{I}$. To generate a G-spectrum for $ion = (charge, offset, sign)$ with a real offset $roffset$, (e.g. real offset of the singly-charged b-ion is 1.008), we first convert $S = \{(mz_1, rank_1), \ldots, (mz_l, rank_l)\}$ into $S' = \{(mass_1, rank_1), \ldots, (mass_l, rank_l)\}$ using the following transformation:

$$mass_j = \begin{cases} mz_j \cdot charge - roffset & \text{if } sign = 1 \\ \text{RParentMass}(S) - \big(mz_j \cdot charge - roffset\big) & \text{if } sign = -1 \end{cases}$$

Each peak of S representing ion corresponds to a peak of this *converted spectrum* S' representing an ion type $(1, 0, 1)$. Therefore, the vertex and edge labels of the G-spectrum for ion are defined as outlined before, but using S' instead of S (Figure 4).

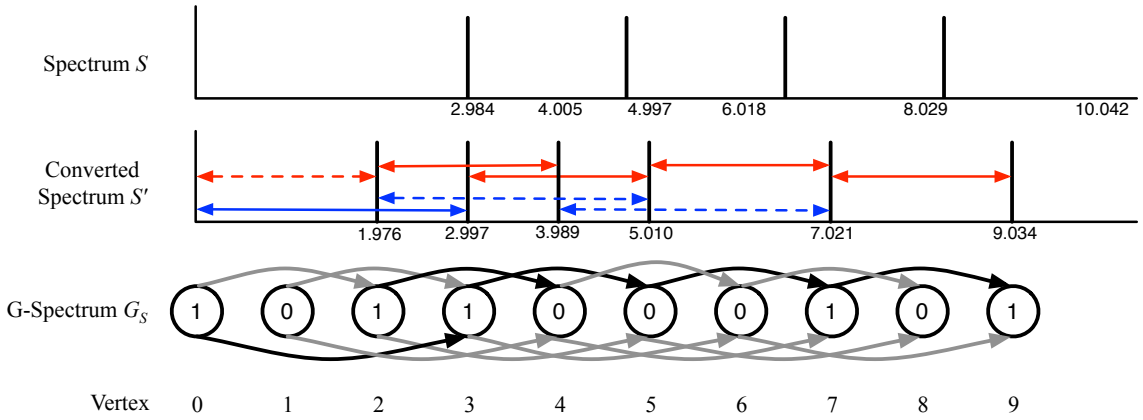


Figure 4: Constructing a G-spectrum in the case of a simplified amino acid model (only two amino acids with real masses 2.012 and 2.996) is used. Assume that only singly-charged b-ion with a real offset 1.008 contributes to the scoring. The spectrum S is converted into S' by shifting each peak by 1.008 to the left. Each arrowed line in S' represents a pair of peaks separated approximately by 2 Da (blue) or 3 Da (red) that form a duo (solid) or does not form a duo (dashed) for a fragment mass tolerance 0.01 Da. A G-spectrum G_S is constructed from S'. The number in the vertex represents its label. The color of the edge represents its label (0 for grey and 1 for black).

In reality, integer instead of Boolean values are used for vertex and edge labels of G-spectra. Given a converted spectrum S', first all peaks $(x, rank)$ are removed if there exists another peak $(x', rank')$ where $[0.9995 \cdot x] = [0.9995 \cdot x']$ and $rank > rank'$. The vertex label v_i is defined as follows: $v_i = rank$ if there exists a peak $(x, rank)$ satisfying $[0.9995 \cdot x] = i$ and $v_i = 0$ otherwise. For an integer m, let AminoAcid(m) be the set of amino acids $a \in \mathcal{A}^+$ satisfying Mass$(a) = m$ (e.g. AminoAcid$(128) = \{\text{Gln}, \text{Lys}\}$). The edge label $e_{i,j}$ is defined as follows: $e_{i,j} = [100 \cdot \min_{a \in \text{AminoAcid}(j-i)}(y - x - \text{RMass}(a))]$ if there exists a duo of peaks with masses x and y such that $[0.9995 \cdot x] = i$ and $[0.9995 \cdot y] = j$, and $e_{i,j} = \infty$ otherwise. The constant 100 is multiplied to discretize the real-valued errors into bins of size 0.01 Da.

In this G-spectrum representation, vertex labels encode the information on the *intensities* of individual peaks, and the edge labels encode the information on the *mass errors* of pairs of peaks assuming they represent consecutive peaks of the same ion type.

From spectral vectors to spectral DAGs Given a set of G-spectra of S, we generate a *spectral DAG* (instead of a spectral vector) of S. A spectral DAG is a DAG with a vertex set $V = \{0, \ldots, M\}$ and and edge set $E = \{(i, j) | j - i \in \text{Mass}(a) \text{ for } a \in \mathcal{A}^+\}$, where the score of a vertex i is the *sum* of VertexScore(i) over all G-spectra of S, the score of an edge (i, j) is the *sum* of EdgeScore(i, j) over all G-spectra of S, the probability of an edge is defined as $\frac{1}{20}$, the score of a path is defined as

the sum of scores of its *vertices and edges*, and the probability of a path is defined as the product of probabilities of its edges. Note that a path of a spectral DAG represents a peptide (or a variant), and the score of a path represents the score of the peptide represented by the path. Given a spectral DAG, one can compute spectral E-values of peptides (or variants) using an approach similar to the generating function approach [32].

4.1.4 Preliminary results

As the first step towards full-scale UniQuest implementation, we recently modified MS-GF to incorporate some of the ideas described above We benchmarked this new version (referred to MS-GF+) using diverse spectral datasets: (i) spectra of varying fragmentation methods with either linear ion trap or orbitrap readout; (ii) spectra of multiple enzyme digests; (iii) spectra of phosphorylated peptides; (iv) spectra of peptides with unusual fragmentation propensities. Overall, we used 19 datasets ($\approx$ 2.83 million spectra from human, yeast, mouse, and *S. pombe* corresponding to 17 distinct spectral types shown in Figure 1. For all these datasets, the new tool significantly increased the number of identified peptides compared to state-of-the-art methods for peptide identifications. We emphasize that while MS-GF+ is not specifically designed for any particular experimental set-up, it improved upon the performance of tools specifically designed for these applications (e.g., specialized tools for phosphoproteomics).

Benchmarking universal MS/MS database search We compared the numbers of identified PSMs at 1% FDR for MS-GF+ and Mascot+Percolator (i.e., PSMs reported by Mascot and re-scored by Percolator). Mascot+Percolator represents the state-of-the-art method for peptide identification that greatly increases the number of identifications as compared to Mascot,

For all datasets, MS-GF+ identified significantly more PSMs compared to Mascot+Percolator (Figure 5). Figure 6 (a) shows the benchmarking results for the five human datasets generated with varying fragmentations and instruments [14]. To figure out how various tools benefits from high-precision product ion peaks, we compared MS-GF+, Mascot+Percolator, and Mascot (Figure 6 (b)). Figure 6 (c) shows the comparison for the ten yeast datasets generated with varying fragmentations (CID or ETD) and enzymes (Trypsin, LysC, ArgC, GluC, or AspN) [53]. Again, for all these datasets, MS-GF+ identified significantly more PSMs (34-168%) than Mascot+Percolator (Figure 6

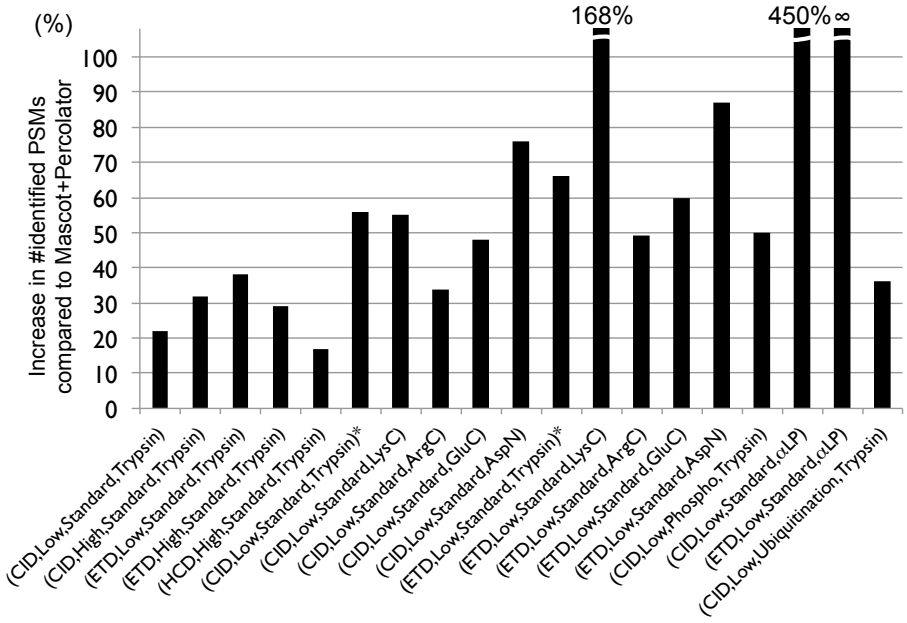


Figure 5: Percent increases in the number of identified PSMs for MS-GF+ compared to Mascot+Percolator for 19 datasets. Each bar represents a spectral dataset of a specified spectral type. For the (ETD,Low,Standard,αLP) dataset, Mascot+Percolator identified no PSM (at 1% FDR).

(c)).

Modification-specific searches To see whether our scoring model can capture the fragmentation propensities specific to phosphopeptides, we generated a scoring parameter set for (CID, Low,Phosphorylation,Trypsin). For the mouse dataset corresponding to (CID,Low,Phosphorylation,Trypsin), we compared the numbers of identified PSMs for MS-GF+ with and without using the phosphorylation-specific parameter set, Mascot+Percolator, and InsPecT [57, 48] equipping with a dedicated scoring model for (CID,Low,Phosphorylation,Trypsin). Interestingly, without phosphorylation-specific scoring parameters, MS-GF+ outperformed both tools, identifying 37% and 44% more PSMs than Mascot+Percolator and InsPecT, respectively. With phosphorylation-specific parameters, MS-GF+ identified 12% more PSMs of phosphopeptides, confirming that our scoring model successfully captures phosphorylation-specific fragmentation propensities.

A similar result was obtained for a (CID,Low,Ubiquitination,Trypsin) dataset. We emphasize that our universl tool does not "know" anything about the peculiarities of the phosphorylation or ubiquitination, and simply trains the scoring parameters in exactly same way it does for other spec-

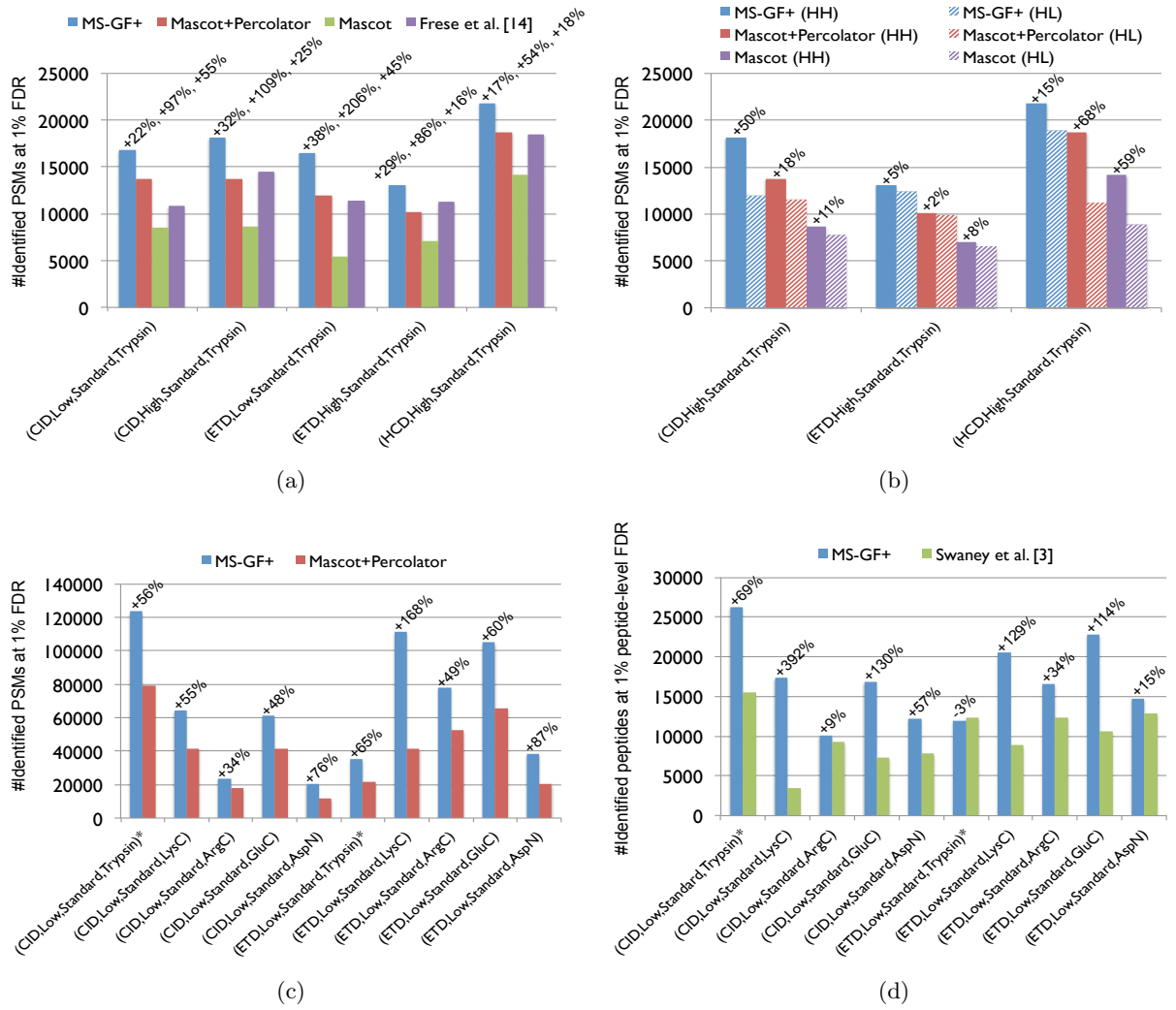


Figure 6: Comparison of various peptide identification tools for diverse spectral types. The numbers of identified PSMs (a-c) or peptides (d) at 1% FDR are shown. Numbers above bars represent the percentages of increase in the number of identifications for MS-GF+ compared to other tools. (a) Results for the human datasets with varying fragmentations and instruments. Percolator greatly increased the number of identifications as compared to Mascot, but MS-GF+ outperformed Mascot+Percolator for all the datasets. (b) Increase in the number of identifications due to the availability of high-precision product ion peaks. For the three human datasets representing HH spectra, MS-GF+, Mascot+Percolator, and Mascot were run using search parameters for HL spectra. The results of these searches (denoted by HL) are compared with the numbers of identifications for the regular searches (denoted by HH). HH searches identified more PSMs than HL searches for every tool and every dataset. (c) Results for the yeast datasets with varying fragmentations and enzymes. MS-GF+ outperformed Mascot+Percolator for all these datasets. (d) Comparison of MS-GF+ and the results in [53] that used OMSSA along with in-house post-processing tools for the yeast datasets.

tral types. This ability to easily train modification-specific scoring parameters for any modification will greatly benefit researchers studying post-translational modifications.

Identifying spectra with unusual fragmentation propensities αLP is a new protease (recently isolated in Elizabeth Komives' laboratory at UCSD) with cleavage specificities somewhat "orthogonal" to trypsin. MS-GF+ was applied to the study of αLP using two *S. pombe* datasets corresponding to (CID,Low,Standard,αLP) and (ETD,Low,Standard,αLP). We ran Mascot+Percolator, OMSSA, and MS-GF+ by specifying 'None' as an enzyme. Since αLP produces peptides with different fragmentation propensities than tryptic peptides, Mascot+Percolator and OMSSA performed very poorly for this novel spectral type. In contrast, MS-GF+ identified 3,535 and 2,829 PSMs from the (CID,Low,Standard,αLP) and (ETD,Low,Standard,αLP) dataset using the scoring parameters for (CID,Low,Standard,Trypsin) and (ETD,Low,Standard,Trypsin), respectively. The poor performance of Mascot+Percolator and OMSSA is because their scoring functions are not adequate for αLP peptides (correct peptide did not attain the maximal score) for most of the spectra due to the large search space (i.e. no enzyme is specified).

Using the identified PSMs by MS-GF+, we trained scoring parameters for (CID,Low,Standard, αLP) and (ETD,Low,Standard,αLP). When these αLP-specific scoring parameters were used, the number of identified PSMs further increased by 35% and 17%, respectively, showing the usefulness of MS-GF+ for studies of new proteases.

Thus, αLP represents a new alternative to trypsin, greatly increasing the protein sequence coverages, but generating spectra with unusual fragmentation propensities. We emphasize that the capabilities of αLP are not obvious when Mascot+Percolator or another tool is used, because it fails to identify αLP peptides.

4.2 Aim 2: peptide sequencing

In the last decade, *de novo* peptide sequencing algorithms have become rather complex and adopted a number of algorithmic and machine learning innovations such as the antisymmetric path approach or the rank-based version of AdaBoost. To simplify the presentation, we start from an idealized model (similar to [34]) that assumes the following:

- the masses of amino acids are integers (e.g., the mass of Gly is 57).

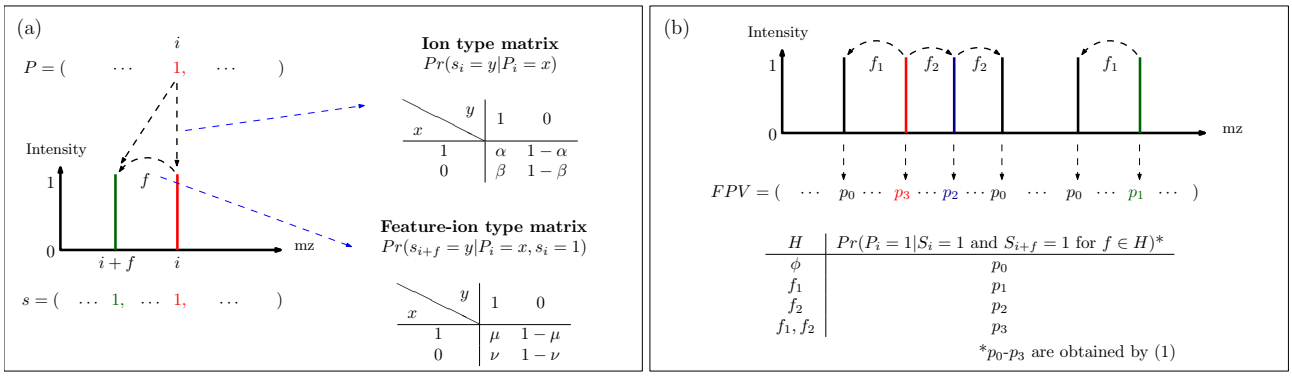


Figure 7: (a) The generation of a partial-spectrum s for P_i. One ion type $\delta = 0$ and one feature f are considered. The probability that $s_i = 1$ is given by α if $P_i = 1$ or by β otherwise. Given that $s_i = 1$, the probability that $s_{i+f} = 1$ (i.e., the peak i satisfies f) is given by μ if $P_i = 1$ or by ν otherwise. The spectrum is generated by taking elementwise OR operation for generated partial-spectra for all elements of P.
(b) The calculation of the fragmentation probability vector FPV from a spectrum S (without knowing the peptide P that generated S). We consider one ion type $\delta = 0$ and two features f_1 and f_2. The events "a peak satisfies f_1" and "a peak satisfies f_2" are assumed to be independent. To derive FPV_i, first we examine which features the peak i satisfies in the spectrum S. Denote the features the peak i satisfies by H. Second, given H, we calculate the probability that $P_i = 1$ (using the probabilities given in ion type matrix and feature-ion type matrix - see the equation (4)).

- the m/z (mass to charge ratio) of peaks (in spectra) are integers.
- the intensity of all peaks is 1.
- only N-terminal charge 1 ions are considered (e.g., b, c, or $b - H_2O$ ions, but not y-ion series).
- the *parent mass* (the mass of the precursor ion) of a spectrum equals to the mass of the peptide that generated the spectrum.

While this model may appear to be too simplistic to capture the complexities of the real spectra, the analysis in [34] illustrates how to extend this idealized model to real spectra. In our implementation of UniNovo, we will follow the approach from [34] but will limit the description in this proposal to the idealized model (the algorithm for a more realistic model becomes too complex to be presented with the page limit for a TRDP).

4.2.1 Peptides and spectra as boolean vectors

Let A be the set of amino acids with (integer) masses $m(a)$ for $a \in A$. A *peptide* $a_1 a_2 \cdots a_k$ is a sequence of amino acids, and the mass of a peptide is the total mass of amino acids in the peptide. We represent a peptide $a_1 a_2 \cdots a_k$ with mass n by a Boolean vector $P = (P_1, \cdots, P_n)$, where $P_i = 1$ if $i = \sum_{t=1}^{j} a_t$ for $0 < j < k$, and $P_i = 0$ otherwise. If $P_i = 1$, we call a mass i a *fragmentation site*.

For example, in the case of two amino acids A and B with masses 2 and 3, the peptide $ABBA$ has the mass of $2+3+3+2=10$ and is represented by a Boolean vector $(0,1,0,0,1,0,0,1,0,0)$. The fragmentation sites of this peptide are, thus, 2, 5, and 8.

A *spectrum* is a list of peaks, where each peak is specified by an m/z. We represent a spectrum of parent mass n by a Boolean vector $S=(S_1,\cdots,S_n)$, where $S_i=1$ if the peak of m/z i (or simply the peak i) is present and $S_i=0$ otherwise.

A *peptide-spectrum match (PSM)* is a pair (P,S) formed by a peptide P and a spectrum S. Given an integer δ called an *ion type* and a PSM (P,S), we say a peak i is a *δ-ion peak* (with respect to P) if $i-\delta$ is a fragmentation site, that is, $P_{i-\delta}=1$. In this model, the ion type can be any integer (for example, the ion types 1 and -27 represent b and a ions, respectively).

Given an integer f called a *feature* and a spectrum S, we say that a peak i *satisfies* f if another peak $i+f$ is present in the spectrum, that is, $S_{i+f}=1$. For instance, a peak 30 satisfies a feature $f=-18$, if $S_{30-18}=1$. For example, a water loss (from any ions) is represented by the feature $f=-18$, and the mass gain from a-ion to b-ion is represented by the feature $f=+27$.

4.2.2 Peptide-spectrum generative model

We model how a peptide P (of mass n) generates a spectrum S. Departing from a 1-step generative model in [1, 34], we introduce a more adequate 2-step probabilistic model in which the dependency between different ions can be described.

Assume that we are given the *ion type set* Δ and (the *feature set* F. For simplicity, we consider the case where only one ion type $\delta=0$ is in Δ and one feature f is in F. Given a peptide P, a *partial-spectrum* s is generated per each element P_i of P as follows: The probability that $s_i=1$ is given by α if $P_i=1$ or by β otherwise (the first generation step). This first step can be characterized by a 2×2 matrix called the *ion type matrix* (Figure 7). Given that $s_i=1$, the probability that $s_{i+f}=1$ (i.e., the peak i satisfies f) is given by μ if $P_i=1$ or ν otherwise (the second generation step). The second step is characterized by the *feature-ion type matrix* (Figure 7). The second step describes the dependency between different ions from the same fragmentation site. If multiple ion types and multiple features are considered, the ion type matrix should be defined per ion type, and the feature-ion type matrix per ion type and per feature. The spectrum S is generated by taking elementwise OR operation for the generated partial-spectra s.

4.2.3 Learning ion type and feature-ion type matrices

Since the ion type matrices and feature-ion type matricies fully describe the generation of a spectrum, in the training step, UniNovo learns these matrices from the *training set* of PSMs. Using the *offset frequency function* introduced in [8], we collect frequently observed ion types and form the ion type set Δ. Likewise, we collect frequently observed features and form the feature set F. From here on, we only consider ion types in the ion type set Δ and features in the feature set F.

Next UniNovo learns the ion type and feature-ion type matrices that characterize the generative model of the PSMs in the training dataset. For example, $\alpha = Pr(s_i = 1|P_i = 1)$ can be empirically determined if partial-spectra s are given. However, it is not clear how to decompose a spectrum S into partial-spectra s (since partial spectra may share peaks in the spectrum). As a compromise, we learn $Pr(S_i = 1|P_i = 1)$ for estimation of α. Other probabilities are also empirically determined similarly by substituting the partial-spectra by the spectrum. All the above probabilities can be learned from a small set of PSMs even if there are many ion types in Δ and features in F because each probability is associated to an individual ion type or a combination of an ion type and a feature, not a combination of multiple ion types and multiple features.

Lastly, we compute the probability that a random element of a peptide vector is a fragmentation site, i.e., $Pr(P_i = 1)$. This probability is called the *prior fragmentation probability* and denoted by p.

4.3 How to infer fragmentation sites from a spectrum

Given a spectrum S of parent mass n, our goal is to predict the fragmentation sites of the (unknown) peptide P that generated S. For simplicity, assume that there exists a single ion type $\delta = 0$ is in the ion type set Δ (but multiple features in the feature set F). Given a peak i, define H as the set of features that the peak i satisfies. Then the fragmentation sites are predicted by solving the following Bayesian inference problem.

Fragementation inference problem: Given the set of features H and P_i such that $Pr(P_i = 1) = p$ (the prior fragmentation probability), derive the posterior probability $Pr(P_i = 1|S_i = S_{i+f} = 1 \text{ for } f \in H)$.

Since there is only one ion type, we have only one ion type matrix. On the other hand, per each feature we have a feature-ion type matrix. Let μ_f and ν_f denote μ and ν associated to the feature f, respectively. If we can assume that all features are independent (i.e., the events "$S_i = S_{i+f} = 1$ for f" are independent for $f \in H$), then one can show that

$$Pr(P_i = 1 | S_i = S_{i+f} = 1 \text{ for } f \in H) = \frac{\gamma \cdot \prod_{f \in H} \mu_f}{\gamma \cdot \prod_{f \in H} \mu_f + (1-\gamma) \cdot \prod_{f \in H} \nu_f}. \tag{4}$$

where $\gamma = \frac{p \cdot \alpha}{p \cdot \alpha + (1-p) \cdot \beta}$ Denote the obtained probability in (4) as π_i. We define a *fragmentation probability vector (FPV)* as a vector with n elements such that

$$FPV_i = \begin{cases} \pi_i & \text{if } S_i = 1 \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

for $i = 1, \cdots, n-1$, and $FPV_n := 1$ (see Figure 7 (b)). FPV_i is an estimated probability that $P_i = 1$. We use FPV for the generation of *de novo* reconstructions.

The equation (4) is based on a simplified model in which a single ion type and multiple independent features are used. However, some features are known to be strongly dependent (e.g., a feature describing a single water loss and a double water losses), and usually multiple ion types are present in the ion type set. Thus, in practice, per each peak, UniNovo automatically selects a small number of features that are weakly correlated yet effective to determine the ion type of the peak. Assuming that the selected features are mutually independent, FPV is calculated per ion type using the equation (4), and then the final FPV is given by a weighted summation of the FPV's of different ion types.

4.3.1 Generating *de novo* reconstructions

To generate *de novo* reconstructions, we first construct a *spectrum graph* [8]. Given a spectrum S of parent mass n from an unknown peptide P, the spectrum graph $G(V, E)$ is defined as a directed acyclic graph whose vertex set V consists of 0 (the source), n (the sink), and integers i such that $FPV_i > 0$. Two vertices i and j are connected by an edge (i, j) if $j - i$ equals to the mass of an amino acid or the total mass of multiple amino acids (*a mass gap*). Any path from 0 (the source)

to n (the sink) in a spectrum graph corresponds to a peptide (possibly containing mass gaps). We say that a vertex i is *correct* if $P_i = 1$ and an edge (i, j) is *correct* if both vertices i and j are correct. We also say that a path r is *correct* if all vertices in r are correct. The *length* of a reconstruction is defined by the total number of amino acids and mass gaps in the reconstruction.

To score a *de novo* reconstruction, we use an additive (i.e., the score of a path is the sum of scores of vertices of the path) log likelihood ratio scoring. Given a vertex i, let $FPV_i = x$. The likelihoods of the following two hypothesis for the outcome $FPV_i = x$ are tested: a) the vertex i is correct and b) the vertex i is incorrect. Let $Pr(P_i = 1|FPV_i = x) = x$. Then, we have

$$\frac{\mathcal{L}(P_i = 1|FPV_i = x)}{\mathcal{L}(P_i = 0|FPV_i = x)} = \frac{Pr(FPV_i = x|P_i = 1)}{Pr(FPV_i = x|P_i = 0)} = \frac{x}{1-x} \cdot \frac{1-p}{p}. \tag{6}$$

The score of the vertex i with $FPV_i = x$ is defined by $Score(i) := \left[\log \frac{x}{1-x} \cdot \frac{1-p}{p} \right]$ where $[\,\cdot\,]$ denotes the rounding to the nearest integer.

Since an additive scoring is used, top scoring reconstructions can be efficiently generated using dynamic programming [8]. After generating the reconstructions, a probability that each reconstruction is correct (termed the *accuracy* of the reconstruction) is predicted, using Hunter's bound [23]. Hunter's bound can be calculated from statistics that are readily learned from a small set of PSMs.

4.3.2 Preliminary results

Benchmarking To benchmark UniNovo, we used 13 datasets with diverse fragmentation methods (CID/ETD/HCD), digested with diverse proteases (trypsin, LysC, and AspN), and having diverse charge states. Out of all identified spectra in these datasets, we selected 1,000 spectra (or pairs of spectra) from distinct peptides randomly and formed the 13 datasets listed in Table 1.

We benchmarked UniNovo, PepNovo+ [12], PEAKS [41], and pNovo [5] using the datasets in Table 1. For each tool, we generated N reconstructions per each spectrum for $N = 1$, 5, and 20. We say that a spectrum is *correctly sequenced* if at least one of N reconstructions generated from the spectrum is correct. To evaluate the performance of each tool, the number of correctly sequenced spectra and the average length of correct reconstructions were measured for each tool.

Figure 8 shows the comparison results for different datasets while Figure 9 shows the Venn diagrams of the correctly sequenced spectra. UniNovo found the largest number of correctly sequenced

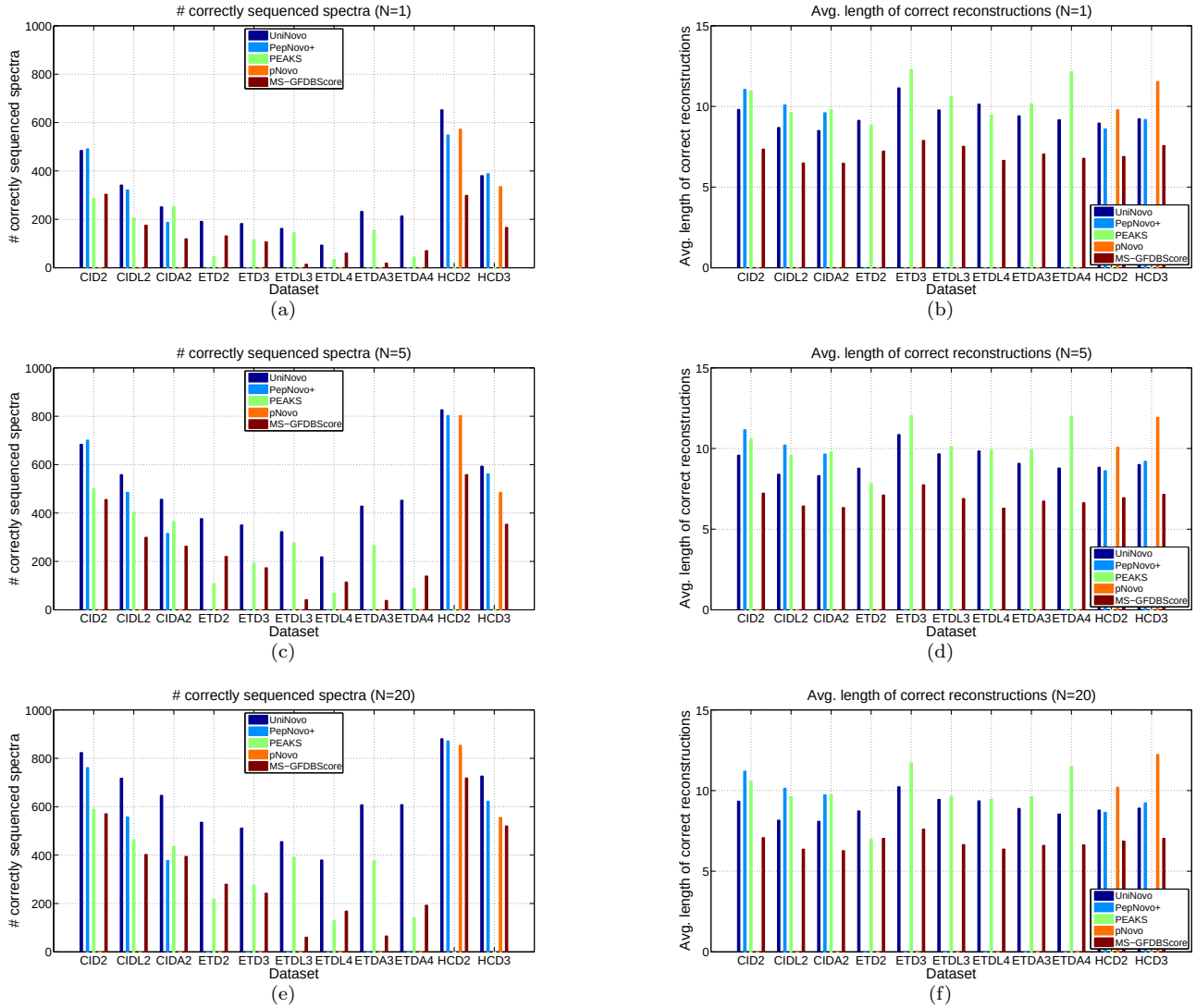


Figure 8: Comparison of *de novo* sequencing tools (as well as a database search tool MS-GFDB [**?**] tweaked for *de novo* sequencing). Per each spectrum, N top scoring reconstructions were generated by UniNovo, PepNovo+ [13, 12], PEAKS [41], pNovo [5], and MS-GFDBScore. The number of reported reconstructions per a spectrum (N) is set to 1, 5, and 20. Figures on the left side ((a), (c), and (e)) show the number of correctly sequenced spectra in each dataset, and figures on the right side ((b), (d), and (f)) show the average length of the correct reconstructions.

Dataset	CID2	CIDL2	CIDA2	ETD2	ETD3	ETDL3	ETDL4	ETDA3	ETDA4	HCD2	HCD3	CID/ETD2	CID/ETD3
Fragmentation	CID	CID	CID	ETD	ETD	ETD	ETD	ETD	ETD	HCD	HCD	CID/ETD	CID/ETD
Charge	2	2	2	2	3	3	4	3	4	2	3	2	3
Enzyme	Tryp	LysC	AspN	Tryp	Tryp	LysC	LysC	AspN	AspN	Tryp	Tryp	Tryp	Tryp
Avg. pep. length	12.6	11.4	12.3	12.5	16.4	12.5	18.7	12.8	18.9	10.5	14.5	12.3	17.1
UniNovo		*	*	*	*	*	*	*	*	*	*	*	*
PepNovo+	*	*	*	N/A	N/A	N/A	N/A	N/A	N/A	*	*	N/A	N/A
PEAKS	*	*	*	*	*	*	*	*	*	*	*	N/A	N/A
pNovo	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	*	*	N/A	N/A

Table 1: Summary of the datasets used for benchmarking. The 13 datasets in the Table (each with 1000 spectra) were selected from 13 large *original* datasets (at least 50,000 spectra) previously analyzed in [**?**, **?**, 15]. All spectra in the original datasets were identified by MS-GFDB [**?**] at 1% peptide-level FDR. Out of all identified spectra, we selected 1,000 spectra (or pairs of spectra) from distinct peptides randomly and formed the 13 datasets listed in the Table. The unselected identified spectra (about 5,000-20,000 spectra depending on the type of spectra) were used for the training of UniNovo. The peptides contained in the training dataset were not contained in the above 13 datasets. While UniNovo is applicable to all datasets, other tools are only applicable to (or optimized for) datasets marked by '*'.

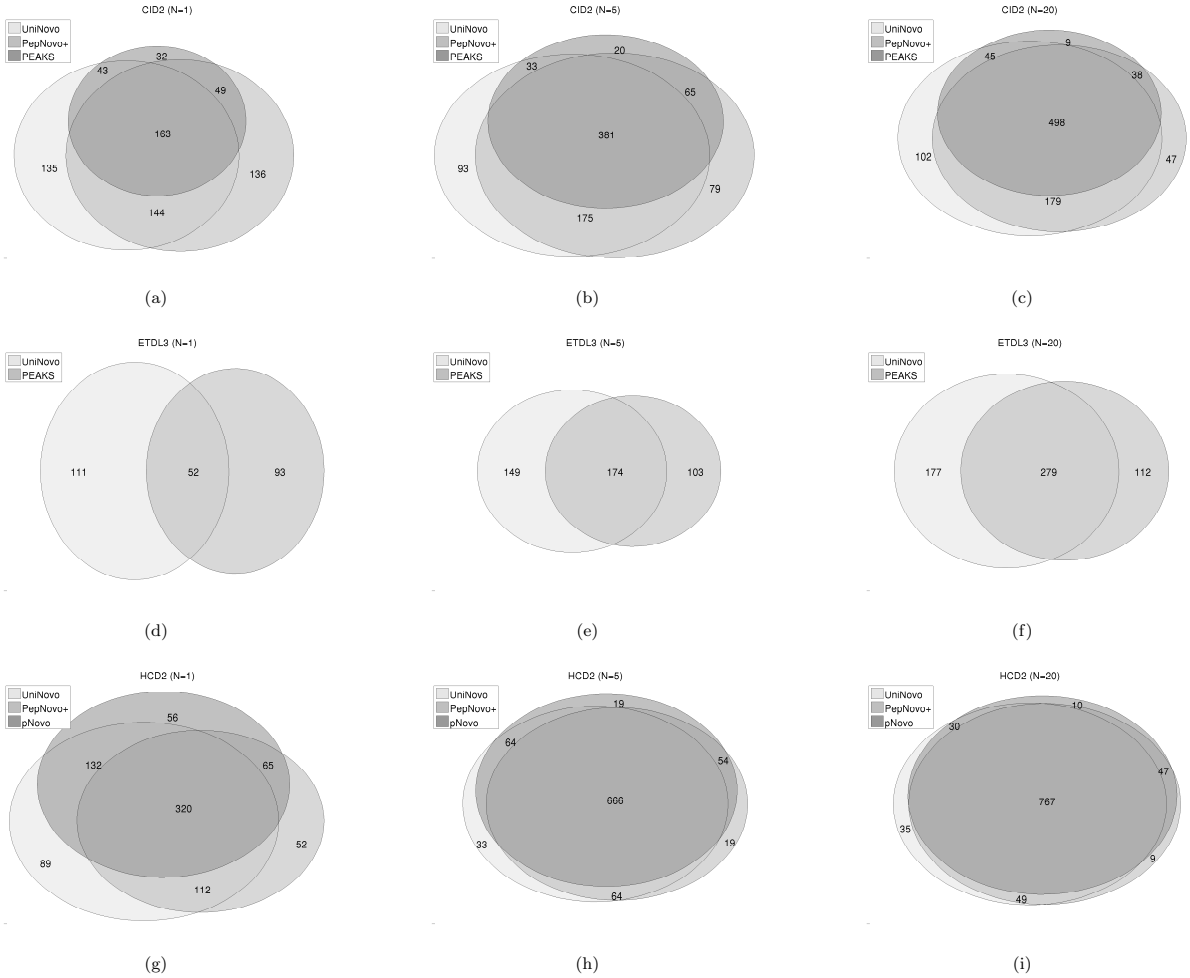


Figure 9: The Venn diagrams of the correctly sequenced spectra for CID2 (a)-(c), ETD3 (d)-(f), and HCD2 (g)-(i) datasets. For all datasets, the overlaps between different tools increase as N grows, as expected. Relatively small overlaps are observed for ETD spectra when compared to CID or HCD spectra.

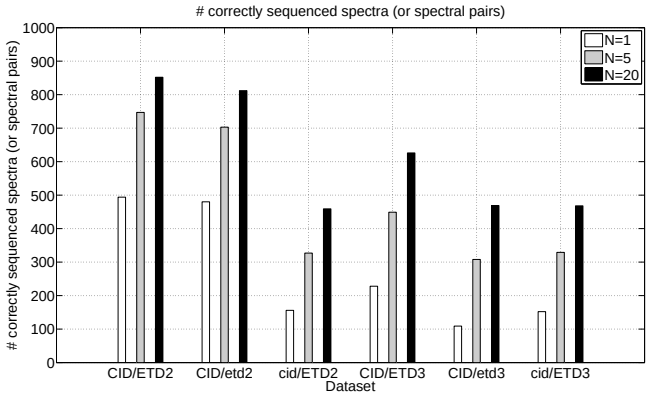
spectra among all the tested tools in most datasets. In particular, for ETD spectra, UniNovo reported significantly more correctly sequenced spectra than PEAKS. For example, in case of ETD2 or ETDL4 dataset, the number of correctly sequenced spectra was more than twice for UniNovo than for PEAKS. For CID spectra, UniNovo and PepNovo+ showed similar results. When $N = 1$, UniNovo and PepNovo+ found about the same number of correctly sequenced spectra in CID2 and CIDL2 datasets, but UniNovo found about 35% more correctly sequenced spectra than PepNovo+ in CIDA2 dataset. The results on HCD spectra also demonstrate that UniNovo finds the largest number of correctly sequenced spectra in general. The reconstructions reported by pNovo were, however, longer than those by UniNovo (and PepNovo+) suggesting that UniNovo still has room for improvement for HCD spectra.

AspN digested peptides generate spectra with distinct fragmentation propensities as compared to trypsin and LysC. While UniNovo worked well for sequencing AspN digested peptides, PepNovo+ showed suboptimal results for the spectra of AspN digested peptides. This is not a criticism of PepNovo+ since it was only trained for CID spectra of tryptic peptides.

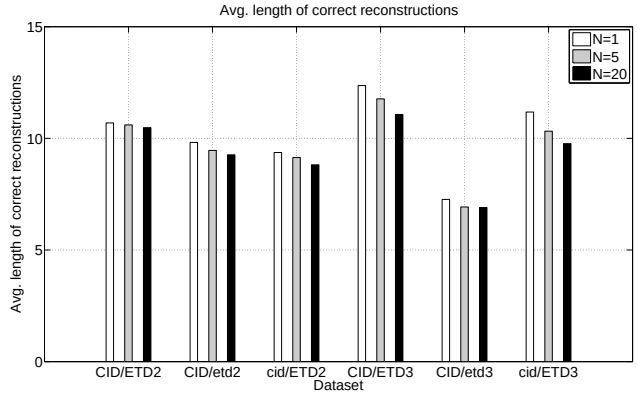
Sequencing of paired spectra Given multiple spectra from the same precursor ion, UniNovo first generates a spectrum graph from each of the spectra and next merges the spectrum graphs into a combined spectrum graph, on which the reconstructions are generated. We benchmarked UniNovo in *de novo* sequencing of paired spectra from CID/ETD2 and CID/ETD3 datasets (Figure 10). When precursor ions were doubly charged, the performance boost from the paired spectra was very modest. In contrast, for triply charged spectra, the use of paired spectra was highly beneficial for generating more accurate reconstructions.

5 Summary

Many efforts have been invested into making peptide identification and sequencing tools compatible with new types of data. For example, several pre-processing and post-processing strategies [54, 21] as well as several statistical modeling tools (e.g., PeptideProphet [29], and Percolator [28]) have been proposed to boost the performance of MS/MS database search tools. These tools do not find new PSMs, but rather re-score PSMs reported by a database search tool using more complex scoring



(a)



(b)

Figure 10: *De novo* sequencing of paired spectra. CID/ETD spectral pairs were analyzed by UniNovo (in CID/ETD2 and CID/ETD3 datasets). To see if the spectral pairs are beneficial for *de novo* sequencing, CID/etd2 (cid/ETD2) dataset was generated from CID/ETD2 dataset by collecting only CID (ETD) spectra in CID/ETD2 dataset. Likewise, CID/etd3 and cid/ETD3 datasets were generated from CID/ETD3 dataset. (a) the number of correctly sequenced spectra (or spectral pairs), (b) the average length of correct reconstructions for each dataset. The spectral pairs resulted in more accurate and longer reconstructions, in particular for triply charged spectral pairs.

and output high-scoring PSMs. While they often improve the performance of a database search tool, their performance is negatively affected when the database search tool fails to find correct PSMs [35]. Another downside of the pre- or post-processing strategies and statistical modeling tools is that since they are often not integrated into database search tools, using them complicates the analysis of MS/MS spectra. Our goal is to develop a tool that does not require any additional pre-processing, post-processing, or statistical modeling tools and thus makes the MS/MS analysis reproducible across different datasets and laboratories.

Our preliminary analysis showed that for diverse types of spectral datasets, the pilot version of our universal tool identifies more PSMs than existing peptide identification tools even when they are coupled with powerful statistical modeling tools like Percolator. In fact, our recent results demonstrated that MS-GF+ has similar discriminating power as the leading spectral library search tool SpectraST [38]. The comparable performance of MS-GF+ and SpectraST indicates that access to previously identified spectra does not necessarily translate into significant improvement in accurate peptide identification. It implies that scoring methods that compare Spectrum-Spectrum Matches (SSMs) are also important (see another TRDP for CCMS efforts to improve spectral library searches). In contrast to the highly sophisticated methods for PSM scoring used in database searches, the library SSM scoring has not matured enough and is largely based on simple spectral cosine scores rather than statistical significance. In the new cycle, CCMS will improve the performance of spectral library searching tools by developing rigorous methods for computing statistical significance of SSMs.

As pointed out by [40], *de novo* sequences not only are valuable for the analysis of the novel peptides that are not present in proteome databases but also for various downstrean application like MS/MS database searches. Since the reconstructions reported by UniNovo contain mass gaps (termed *gapped peptides* [**?**, 25]), MS-BPM algorithm developed at CCMS [44] can be used for superfast MS/MS searches (UniNovo⊕MS-BPM). MS-BPM enables searches against a sequence database using gapped peptides as queries. Currently MS-BPM takes gapped peptides generated by MS-GappedDictionary [25]. However, the reconstructions from UniNovo are usually longer than those from MS-GappedDictionary (8-9 vs. 5-6). Since the search time of MS-BPM strongly depends on the length of gapped peptides - the longer gapped peptides, the shorter search time - the running time of UniNovo⊕MS-BPM is smaller than MS-GappedDictionary⊕MS-BPM by an

order of magnitude in a blind search against the IPI Human proteome database [30].

6 Driving Biomedical Projects

References

[1] Nuno Bandeira, Jesper V. Olsen, Matthias Mann, and Pavel A. Pevzner. Multi-spectra peptide sequencing and its applications to multistage mass spectrometry. *Bioinformatics*, 24(13):i416–i423, January 2008.

[2] Sheila J Barton and John C Whittaker. Review of factors that influence the abundance of ions produced in a tandem mass spectrometer and statistical methods for discovering these factors. *Mass Spectrometry Reviews*, 28(1):177–187, 2009.

[3] M Bern, Y Cai, and D Goldberg. Lookup peaks: a hybrid of de novo sequencing and database search for protein identification by tandem mass spectrometry. *Anal. Chem.*, 79:1393–400, Jan 2007.

[4] Linda A Breci, David L Tabb, 3rd Yates, John R, and Vicki H Wysocki. Cleavage n-terminal to proline: analysis of a database of peptide tandem mass spectra. *Analytical Chemistry*, 75(9):1963–1971, 2003.

[5] Hao Chi, Rui-Xiang Sun, Bing Yang, Chun-Qing Song, Le-Heng Wang, Chao Liu, Yan Fu, Zuo-Fei Yuan, Hai-Peng Wang, Si-Min He, and Meng-Qiu Dong. pNovo: de novo peptide sequencing and identification using HCD spectra. *J. Proteome Res.*, 9(5):2713–2724, 2010.

[6] Jürgen Cox, Nadin Neuhauser, Annette Michalski, Richard A Scheltema, Jesper V Olsen, and Matthias Mann. Andromeda: A peptide search engine integrated into the maxquant environment. *J. Proteome Res.*, 10:1794–805, Feb 2011.

[7] Robertson Craig and Ronald C Beavis. Tandem: matching proteins with tandem mass spectra. *Bioinformatics*, 20(9):1466–7, Jun 2004.

[8] V Dancik, T A Addona, K R Clauser, J E Vath, and P A Pevzner. De novo peptide sequencing via tandem mass spectrometry. *Journal of Computational Biology*, 6(3-4):327–342, 1999.

[9] Ritendra Datta and Marshall Bern. Spectrum fusion: using multiple mass spectra for de novo peptide sequencing. *Journal of Computational Biology*, 16(8):1169–1182, 2009.

[10] Eric W Deutsch, Luis Mendoza, David Shteynberg, Terry Farrah, Henry Lam, Natalie Tasman, Zhi Sun, Erik Nilsson, Brian Pratt, Bryan Prazen, Jimmy K Eng, Daniel B Martin, Alexey I Nesvizhskii, and Ruedi Aebersold. A guided tour of the trans-proteomic pipeline. *Proteomics*, 10(6):1150–9, Mar 2010.

[11] J Eng, A McCormack, and J Yates. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.*, 5:976–89, Jan 1994.

[12] Ari Frank. A ranking-based scoring function for peptide-spectrum matches. *Journal of Proteome Research*, 8(5):2241–2252, 2009.

[13] Ari Frank and Pavel Pevzner. PepNovo: de novo peptide sequencing via probabilistic network modeling. *Anal. Chem.*, 77(4):964–973, 2005.

[14] Christian K Frese, A F Maarten Altelaar, Marco L Hennrich, Dirk Nolting, Martin Zeller, Jens Griep-Raming, Albert J R Heck, and Shabaz Mohammed. Improved peptide identification by targeted fragmentation using cid, hcd and etd on an ltq-orbitrap velos. *J. Proteome Res.*, 10(5):2377–88, May 2011.

[15] Christian K. Frese, A. F. Maarten Altelaar, Marco L. Hennrich, Dirk Nolting, Martin Zeller, Jens Griep-Raming, Albert J. R. Heck, and Shabaz Mohammed. Improved peptide identification by targeted fragmentation using CID, HCD and ETD on an LTQ-Orbitrap velos. *J. Proteome Res.*, 10(5):2377–2388, 2011.

[16] Lewis Y Geer, Sanford P Markey, Jeffrey A Kowalak, Lukas Wagner, Ming Xu, Dawn M Maynard, Xiaoyu Yang, Wenyao Shi, and Stephen H Bryant. Open mass spectrometry search algorithm. *J. Proteome Res.*, 3(5):958–64, Jan 2004.

[17] Viktor Granholm, William Stafford Noble, and Lukas Käll. On using samples of known protein content to assess the statistical calibration of scores assigned to peptide-spectrum matches in shotgun proteomics. *J. Proteome Res.*, 10(5):2671–8, May 2011.

[18] N Gupta and P Pevzner. False discovery rates of protein identifications: a strike against the two-peptide rule. *J. Proteome Res.*, Jul 2009.

[19] Nitin Gupta, Nuno Bandeira, Uri Keich, and Pavel A Pevzner. Target-decoy approach and false discovery rate: when things may go wrong. *J. Am. Soc. Mass Spectrom.*, 22(7):1111–20, Jul 2011.

[20] Lin He and Bin Ma. ADEPTS: advanced peptide de novo sequencing with a pair of tandem mass spectra. *Journal of Bioinformatics and Computational Biology*, 8(6):981–994, 2010.

[21] Edward J Hsieh, Michael R Hoopmann, Brendan Maclean, and Michael J Maccoss. Comparison of database search strategies for high precursor mass accuracy ms/ms data. *J. Proteome Res.*, 9:1138–43, Nov 2009.

[22] Yingying Huang, Joseph M Triscari, George C Tseng, Ljiljana Pasa-Tolic, Mary S Lipton, Richard D Smith, and Vicki H Wysocki. Statistical characterization of the charge state and residue dependence of low-energy CID peptide dissociation patterns. *Analytical Chemistry*, 77(18):5800–5813, 2005.

[23] David Hunter. An upper bound for the probability of a union. *Journal of Applied Probability*, 13(3):597–603, 1976.

[24] Edward L Huttlin, Mark P Jedrychowski, Joshua E Elias, Tapasree Goswami, Ramin Rad, Sean A Beausoleil, Judit Villén, Wilhelm Haas, Mathew E Sowa, and Steven P Gygi. A tissue-specific atlas of mouse protein phosphorylation and expression. *Cell*, 143(7):1174–89, Dec 2010.

[25] Kyowon Jeong, Sangtae Kim, Nuno Bandeira, and Pavel A. Pevzner. Gapped spectral dictionaries and their applications for database searches of tandem mass spectra. *Molecular & Cellular Proteomics*, 10(6):M110.002220, 2011.

[26] Richard S. Johnson, Stephen A. Martin, Klaus Biemann, John T. Stults, and J. Throck Watson. Novel fragmentation process of peptides by collision-induced decomposition in a tandem mass spectrometer: differentiation of leucine and isoleucine. *Anal. Chem.*, 59(21):2621–2625, 1987.

[27] Andrew R Jones, Martin Eisenacher, Gerhard Mayer, Oliver Kohlbacher, Jennifer Siepen, Simon Hubbard, Julian Selley, Brian Searle, James Shofstahl, Sean Seymour, Randall Julian,

Pierre-Alain Binz, Eric W Deutsch, Henning Hermjakob, Florian Reisinger, Johannes Griss, Juan Antonio Vizcaino, Matthew Chambers, Angel Pizarro, and David Creasy. The mzidentml data standard for mass spectrometry-based proteomics results. *Mol. Cell. Proteomics*, Feb 2012.

[28] Lukas Käll, Jesse D Canterbury, Jason Weston, William Stafford Noble, and Michael J Maccoss. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat. Methods*, 4(11):923–5, Nov 2007.

[29] A Keller, A Nesvizhskii, E Kolker, and R Aebersold. Empirical statistical model to estimate the accuracy of peptide identifications made by ms/ms and database search. *Anal. Chem.*, 74:5383–92, Jan 2002.

[30] Paul J Kersey, Jorge Duarte, Allyson Williams, Youla Karavidopoulou, Ewan Birney, and Rolf Apweiler. The international protein index: an integrated database for proteomics experiments. *Proteomics*, 4(7):1985–1988, 2004.

[31] S Kim, N Gupta, N Bandeira, and P Pevzner. Spectral dictionaries: Integrating de novo peptide sequencing with database search of tandem mass spectra. *Mol. Cell. Proteomics*, 8:53–69, Jan 2009.

[32] S Kim, N Gupta, and P Pevzner. Spectral probabilities and generating functions of tandem mass spectra: A strike against decoy databases. *J. Proteome Res.*, 7:3354–63, Jul 2008.

[33] Sangtae Kim, Nuno Bandeira, and Pavel A Pevzner. Spectral profiles, a novel representation of tandem mass spectra and their applications for de novo peptide sequencing and identification. *Molecular & Cellular Proteomics*, 8(6):1391–1400, 2009.

[34] Sangtae Kim, Nitin Gupta, Nuno Bandeira, and Pavel A. Pevzner. Spectral dictionaries. *Molecular & Cellular Proteomics*, 8(1):53 –69, 2009.

[35] Sangtae Kim, Nikolai Mischerikow, Nuno Bandeira, J Daniel Navarro, Louis Wich, Shabaz Mohammed, Albert J R Heck, and Pavel A Pevzner. The generating function of cid, etd, and cid/etd pairs of tandem mass spectra: Applications to database search. *Mol. Cell. Proteomics*, 9(12):2840–52, Dec 2010.

[36] Sangtae Kim, Nikolai Mischerikow, Nuno Bandeira, J. Daniel Navarro, Louis Wich, Shabaz Mohammed, Albert J. R. Heck, and Pavel A. Pevzner. The generating function of CID, ETD, and CID/ETD pairs of tandem mass spectra: Applications to database search. *Molecular & Cellular Proteomics*, 9(12):2840 –2852, 2010.

[37] Aaron A Klammer, Christopher Y Park, and William Stafford Noble. Statistical calibration of the sequest xcorr function. *J. Proteome Res.*, 8(4):2106–13, Apr 2009.

[38] Henry Lam, Eric W Deutsch, James S Eddes, Jimmy K Eng, Nichole King, Stephen E Stein, and Ruedi Aebersold. Development and validation of a spectral library searching method for peptide identification from ms/ms. *Proteomics*, 7(5):655–67, Mar 2007.

[39] Xiaowen Liu, Baozhen Shan, Lei Xin, and Bin Ma. Better score function for peptide identification with ETD MS/MS spectra. *BMC Bioinformatics*, 11(Suppl 1):S4, 2010.

[40] Bin Ma and Richard Johnson. De novo sequencing and homology searching. *Molecular & Cellular Proteomics*, page O111.014902, 2011.

[41] Bin Ma, Kaizhong Zhang, Christopher Hendrie, Chengzhi Liang, Ming Li, Amanda Doherty-Kirby, and Gilles Lajoie. PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. *Rapid Communications in Mass Spectrometry: RCM*, 17(20):2337–2342, 2003.

[42] U Manber and Gene Myers. Suffix arrays: a new method for on-line string searches. *SIAM J. Computing*, 22:935–48, Jan 1990.

[43] Alexey I Nesvizhskii. A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. *J Proteomics*, 73(11):2092–123, 2010.

[44] Julio Ng, Amihood Amir, and Pavel A. Pevzner. Blocked pattern matching problem and its applications in proteomics. RECOMB 2011, Vancouver, Canada, pages 298–319, 2011.

[45] William S Noble. How does multiple testing correction work? *Nat. Biotechnol.*, 27(12):1135–7, Dec 2009.

[46] William Stafford Noble and Michael J Maccoss. Computational and statistical analysis of protein mass spectrometry data. *PLoS Comput. Biol.*, 8(1):e1002296, Jan 2012.

[47] Jesper V Olsen, Boris Macek, Oliver Lange, Alexander Makarov, Stevan Horning, and Matthias Mann. Higher-energy c-trap dissociation for peptide modification analysis. *Nature Methods*, 4(9):709–712, 2007.

[48] Samuel H Payne, Margaret Yau, Marcus B Smolka, Stephen Tanner, Huilin Zhou, and Vineet Bafna. Phosphorylation-specific ms/ms scoring for rapid and accurate phosphoproteome analysis. *J. Proteome Res.*, 7(8):3373–81, Aug 2008.

[49] D Perkins, D Pappin, D Creasy, and J Cottrell. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, 20:3551–67, Jan 1999.

[50] Mikhail M. Savitski, Michael L. Nielsen, Frank Kjeldsen, and Roman A. Zubarev. Proteomics-Grade de novo sequencing approach. *J. Proteome Res.*, 4(6):2348–2354, 2005.

[51] Lea M Starita, Russell S Lo, Jimmy K Eng, Priska D von Haller, and Stanley Fields. Sites of ubiquitin attachment in saccharomyces cerevisiae. *Proteomics*, 12(2):236–40, Jan 2012.

[52] Danielle L Swaney, Graeme C McAlister, and Joshua J Coon. Decision tree-driven tandem mass spectrometry for shotgun proteomics. *Nature Methods*, 5(11):959–964, 2008.

[53] Danielle L. Swaney, Craig D. Wenger, and Joshua J. Coon. Value of using multiple proteases for Large-Scale mass Spectrometry-Based proteomics. *J. Proteome Res.*, 9(3):1323–1329, 2010.

[54] Steve M M Sweet, Andrew W Jones, Debbie L Cunningham, John K Heath, Andrew J Creese, and Helen J Cooper. Database search strategies for proteomic data sets generated by electron capture dissociation mass spectrometry. *J. Proteome Res.*, 8(12):5475–84, Dec 2009.

[55] John E P Syka, Joshua J Coon, Melanie J Schroeder, Jeffrey Shabanowitz, and Donald F Hunt. Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. *Proc. Natl. Acad. Sci. USA*, 101(26):9528–33, Jun 2004.

[56] David L. Tabb, Yingying Huang, Vicki H. Wysocki, and John R. Yates. Influence of basic residue content on fragment ion peak intensities in Low-Energy Collision-Induced dissociation spectra of peptides. *Anal. Chem.*, 76(5):1243–1248, 2004.

[57] Stephen Tanner, Hongjun Shu, Ari Frank, Ling-Chi Wang, Ebrahim Zandi, Marc Mumby, Pavel A Pevzner, and Vineet Bafna. Inspect: identification of posttranslationally modified peptides from tandem mass spectra. *Anal. Chem.*, 77(14):4626–39, Jul 2005.

[58] J A Taylor and R S Johnson. Sequence database searches via de novo peptide sequencing by tandem mass spectrometry. *Rapid. Commun. Mass Spectrom.*, 11(9):1067–75, Jan 1997.

[59] Henk W P van den Toorn, Javier Muñoz, Shabaz Mohammed, Reinout Raijmakers, Albert J R Heck, and Bas van Breukelen. Rockerbox: analysis and filtering of massive proteomics search results. *J. Proteome Res.*, 10(3):1420–4, Mar 2011.

[60] Vicki H Wysocki, George Tsaprailis, Lori L Smith, and Linda A Breci. Mobile and localized protons: a framework for understanding peptide dissociation. *Journal of Mass Spectrometry*, 35(12):1399–1406, 2000.

[61] John R Yates, Sung Kyu Robin Park, Claire M Delahunty, Tao Xu, Jeffrey N Savas, Daniel Cociorva, and Paulo Costa Carvalho. Toward objective evaluation of proteomic algorithms. *Nat. Methods*, 9(5):455–6, May 2012.

[62] Chen Zhou, Hao Chi, Le-Heng Wang, You Li, Yan-Jie Wu, Yan Fu, Rui-Xiang Sun, and Si-Min He. Speeding up tandem mass spectrometry-based database searching by longest common prefix. *BMC Bioinformatics*, 11:577, Jan 2010.

[63] Roman A. Zubarev, Alexander R. Zubarev, and Mikhail M. Savitski. Electron Capture/Transfer versus collisionally Activated/Induced dissociations: Solo or duet? *Journal of the American Society for Mass Spectrometry*, 19:753–761, 2008.