A Pre-computed Probabilistic Molecular Search Engine for Tandem Mass Spectrometry Proteomics

Jeff Jones [1]

[1] SoCal Bioinformatics Inc., Montrose, CA

Keywords: mass spectrometry, proteomics, peptide search, sequence identity, false discovery rate, information theory

Abstract

Mass spectrometry methods of peptide identification involve comparing observed tandem spectra with *in-silico* derived spectrum models. Presented here is a proteomics search engine that offers a new variation of the standard approach, with improved results. The proposed method employs information theory and probabilistic information retrieval on a pre-computed and indexed fragmentation database generating a peptide-to-spectrum match (PSM) score modeled on fragment ion frequency. As a result, the direct application of modern document mining, allows for treating the collection of peptides as a corpus and corresponding fragment ions as indexable words, leveraging ready-built search engines and common predefined ranking algorithms. Fast and accurate PSM matches are achieved yielding a 5-10% higher rate of peptide identities than current database mining methods. Immediate applications of this search engine are aimed at identifying peptides from large sequence databases consisting of homologous proteins with minor sequence variations, such as genetic variation expected in the human population.

Introduction

Advances in mass spectrometry (MS) have propelled proteomics toward the forefront of translational biology, with the possibility of exceeding the breadth and complexity of large-scale whole genome analysis. However, interpretation of proteomics data remains challenging due to the vast number of predicted protein sequences possible from genetic translation, splice variation, and the myriad combinations of potential post translational modifications (PTMs). Hence, common to many liquid-chromatography coupled mass spectrometry (LCMS) proteomic surveys are large numbers of unidentified spectra which continue to receive attention and research interest [1, 2]. Current use of tandem MS proteomic search engines compare observed spectra with *in-silico* modeled spectra, calculating correlation values by probabilistic methods [3, 4] or by inferred scores based on applications of signal processing and information theory [5]. Published examples of proteomic search algorithms compare each observed spectrum in a pairwise fashion with all *in-silico* spectra, ranking the comparisons by score [6]. From a collection of all rank-one peptide-to-spectrum-matches (PSMs) a false discovery rate (FDR) can be estimated [7-12] using one of two distinct approaches to define a null distribution: the well-referenced target-decoy approach, which utilizes a null distribution derived from comparisons with a database of randomized peptide sequences [8, 10, 13, 14], and a more recent decoy-free method using a null distribution comprised of alternative target PSMs [15].

While the use of FDR estimations is absolutely prudent for MS-based proteomics research, there are potential hidden issues associated with randomized decoy sequences employed as the estimated null distribution for FDR calculations [15-19]. Additionally, these effects are likely to be exacerbated as the size and complexity of both the target and decoy database increases. This is

50 particularly challenging for researchers attempting to unlock previously unidentified
51 fragmentation spectra or address the inherent protein diversity beyond the typical consensus
52 proteome, ideally including known isoforms and splice variants that arises from genetic and
53 translation variation, or within studies examining multiple variable post-translational
54 modifications.

55 The goal is to develop a proteomic search engine for large sequence databases (possibly without
56 limit), that can be distributed in a cloud computing environment while remaining comparable or
57 competitive to existing proteomic mining algorithms. Here we present a new proteomic search
58 engine, JAMSE (Just Another Molecular Search Engine), that utilizes a precalculated
59 fragmentation index of all non-redundant proteotypic peptides to derive PSMs ranked by common
60 and well developed text mining algorithms, thus basing the approach on both information theory
61 [20] and probabilistic information retrieval [21, 22]. The result of this approach is a PSM score
62 modeled on the relevance of indexed fragment ions, as opposed to completeness of the observed
63 spectrum. This is not an entirely new concept for tandem LCMS, as the field of selected reaction
64 monitoring exploits the unique combinations of elution time, precursor mass and a selected
65 fragment mass [23].

66 For benchmark comparisons, a decoy-free null distribution is utilized for estimating the FDR and
67 introduce the concept of an intrinsically measurable FDR by including protein sequences from an
68 orthogonal species. The pre-computed nature of JAMSE renders a parallel search against a
69 randomized decoy peptide sequences unfeasible, as not all pairwise target peptides are explored.
70 Results, when compared to current approaches in proteomic search engines, indicate a more
71 discriminating PSM correlation metric can be achieved. JAMSE is optimized for Linux, suitable

for cloud distribution, although the efforts herein are designed to test the theory, demonstrations can be made available through https://www.jamse.app/.

Methods

Data

The yeast MS proteome (*Saccharomyces cerevisiae*) collected by Ishburn *et.al.*[24] and made publicly available, is used as a benchmark data set. Briefly, it consists of 8 sets of tandem liquid chromatography (LC) MS injections, derived from a series of orthogonal LC separations coupled to tandem MS experiments designed to comprehensively probe the yeast proteome.

Databases

Protein sequence databases for JAMSE and Comet searches are built from subsets of the UniProtKB repository [25] containing known proteins for *Saccharomyces sp.*, along with those for *Arabidopsis sp.* The inclusion of *Arabidopsis sp.* proteins is used as an intrinsic decoy to evaluate the competing methods of FDR estimation – target-decoy and decoy-free.

Enumerated in Table 1 are two distinct databases. The first, built from the UniProt consensus repository (referred to as simply UniProt herein) consisted of 7,980 *Saccharomyces sp.* proteins and 15,878 *Arabidopsis sp.* proteins, totaling 23,858 proteins with 11.5 million unique peptides. The second, built from combined UniRef100 [26] and TrEMBL [25] databases (designated UniRef herein), consisted of 58,829 *Saccharomyces sp*. and 120,863 *Arabidopsis sp.* proteins, totaling 398,119 proteins with 36.8 million unique peptides. Complimentary decoy databases are built for each of the two protein databases by randomly shuffling each protein sequence.

Indexing and Search

The JAMSE approach to peptide sequence identification does not consider relative peak heights, only the presence of a given peak above a threshold, and therefor is not a challenger to methods attempting to apply machine learning to spectral reconstruction [27]. Rather, JAMSE exploits patterns observed in biological polymers [28] that form the fundamental characteristics of molecular mass defect [29]. Similar to the center of mass defect distributions estimated by Mitra, et.al. [30], JAMSE employs a simple algorithm to convert a precise real number m/z value to an integer value that can be indexed by any common indexing engine, essentially binning a given mass value to the nearest mass defect distribution, which we will call a fragment mass index (FMI). We therefore treat a collection of peptides as a corpus and corresponding FMIs as indexable words, allowing us to leverage ready-built text-mining search engines and common predefined ranking algorithms [21, 22, 31-33].

Pre-computed peptides and corresponding FMIs are based on the known proteome, utilizing common proteotypic rules of digestion, considering up to two missed cleavages, while generating all possible combinations of the PTM acetylation applied to the N-terminal peptide and Lysine residues. FMIs are generated with respect to *in-silico* derived fragmentation patterns, considering both *y-ion* and *b-ion* fragments along with metastable water-loss ions.

Database searching is accomplished by querying a set of observed precursor and fragment mass values, typically filtered by abundance according to similar methods [4], for a given spectrum and converted to respective FMI values. All PSMs returned are ranked by a probabilistic score derived from a common text-mining ranking function [32] as described in Equation 1. The PSM scoring model used in JAMSE is simply the sum of the inverse document frequencies (sum_idf) across all

114 terms, which can be interpreted as the probability that a given peptide (e.g. item/document in a
115 corpus) *D* contains the term *t* (or FMI, as it relates to here to tandem LCMS).

117 $$sum_idf = \sum_{i=0}^{n} -logP(t_i|D)$$

116 *eq. 1*

118 It should be noted that from this equation it is evident that rarer fragments contribute a higher score
119 value and taken together with the overall distribution of fragment masses indicates that larger mass
120 fragments are rarer and, therefore, contribute greater to a given PSM score. Of important note,
121 JAMSE spectrum searching does not consider all pairwise comparisons between the observed
122 spectrum and relevant *in-silico* spectra. Rather, much like a modern index search engine, top hits
123 are produced via this modified method of information retrieval [21].

124 For this study and for comparative search parameters, all searches are limited to a 0.1 Dalton
125 precursor neutral mass tolerance and a 0.5 Dalton fragment measured mass tolerance (i.e. mass
126 defect width), with the top 5 hits retained per spectrum. Additionally, the Comet [34] search
127 algorithm, specifically, the metric *xcorr*, is used the metric for direct comparison to JAMSE
128 *sum_idf*. For computational clarity, Comet is run on a single computer utilizing 16 cores and
129 consuming approximately 1.1GB of memory. JAMSE is distributed across two computers utilizing
130 8 cores each requiring approximately 1.8GB of memory per machine.

131 ***FDR Estimations***

132 The estimated FDR (eFDR) is computed either from the target-decoy approach as described by
133 Keller *et.al.*[10], or as the decoy-free approach described by Gonnelli *et.al.* [15]. For the purposes

134 of validating a given eFDR, we here introduce the concept of empirically measured FDR (mFDR).
135 Both eFDR and mFDR methods provide means to define the null distribution and calculate the
136 FDR q-value according to the procedure proposed by Storey *et.al.*[11].

137 **mFDR calculation**

138 The use of *Arabidopsis sp.* provided a means to calculate the mFDR. Within the constructed
139 protein databases, there are approximately twice as many *Arabidopsis sp.* proteins, and if one
140 considers false positives a random occurrence, then any *Arabidopsis sp.* peptides passing the
141 desired cutoff represented approximately 2/3 the total number of false positives. Thus, the
142 weighted ratio of approximately 3/2 times the number of observed *Arabidopsis sp.* peptides
143 matched represents the mFDR. It is not necessarily suggested to include orthogonal species while
144 searching as finding an appropriate decoy species may be challenging. To be explicit on the choice
145 for *Arabidopsis sp.*, there are only 5,643 peptide sequences shared between the two organisms
146 given the UniProt consensus repository (approx. 5 per 10,000), and 22,434 given the UniRef100
147 and TrEMBL repositories (approx. 6 per 10,000). Any shared sequences identified are assigned to
148 yeast for the purposes of further analysis.

149 ***Evaluation of Search performance***

150 Three distinct evaluations and comparisons are considered; the more traditional total number of
151 PSMs per FDR, the application of estimated verses measured FDR (eFDR ~ mFDR), and a
152 bootstrapped estimation of the receiver operator characteristics (ROC) with corresponding area
153 under the curve (AUC) for given search metrics. Traditional evaluation of search performance
154 between JAMSE and Comet is based on the total number of PSMs obtained for a given eFDR q-
155 value, with larger numbers considered more desirable. Evaluation of eFDR performance (eg.

156 target-decoy and decoy-free) is based on the comparison to the complimentary mFDR, with a
157 closer match to mFDR considered better. Again, it is should be noted that the decoy-free method
158 of FDR estimation is preferred for JAMSE, as this method reduces the computational complexity
159 of the search algorithm intended to only search sequences of natural occurrence. Bootstrap
160 estimations of PMS discriminator metrics detailed in Table 1 was accomplished by randomly
161 selecting a thousand PSM yielding scans, and filtered to the top peptide by corresponding metric
162 for each scan, wherein a peptide having origin from *Saccharomyces sp*. is considered a true match
163 and a peptide from *Arabidopsis sp.* considered a false match. Using this scheme, repeated 1000
164 times, allowed us to generate ROC curves and estimate the performance as a function of AUC and
165 the true positive rate (TPR) at a false positive rate of 0.05. We choose to include the ROC as it
166 reveals how post-processing validation and machine learning methods might exploit that curve to
167 yield more PSMs at a given FPR.

168

169 Results and Discussion

170 Search results using JAMSE demonstrate that a highly discriminant PSM score can be constructed
171 using a model derived from information theory and probabilistic information retrieval. As detailed
172 in Table 2, JAMSE yielded an additional 3.8% in number of PSMs at FDR 0.05. Exploring the
173 complexity of PSM matching, as detailed in Figure 1, a select set of metrics was evaluated from
174 the search results to compare discriminatory power. For each of the metrics a simple ROC
175 estimation was performed, and this same procedure was applied to the Comet search results and
176 offers a unique comparison. For instance, Figure 1a details the discriminatory power of the FMI
177 dot-product (psm_dp) which yields a considerably higher AUC and TPR (at FDR 0.05) than

178 Comet's corresponding metric (Figure 2, num_matched_ions). Breaking out the dot-product
179 components by fragment *y* and *b* ions, shows that for this particular experiment, and likely a
180 product of the instrument and fragmentation process, *y* fragment ions impart a great contribution
181 over their *b* ion counter parts (Figure 1a). It is perhaps surprising to discover that mass accuracy
182 alone has little discriminatory power, (Figure 1b and Figure 2 massdiff), although upon
183 rationalization, as researchers explore increasingly larger databases, the probability of overlapping
184 *in-silico* spectra increases regardless of mass resolution, albeit for high-resolution measurements
185 that probability is lower.

186 The initial goal in developing JAMSE was to explore information recall as an orthogonal means
187 of peptide sequence identification, was to achieve comparable performance to current proteomic
188 database mining algorithms. Figure 1c shows the bootstrapped ROC for the *sum_idf* ranking
189 function with both a considerably higher overall AUC and TPR (at FDR 0.05) than Comet's two
190 corresponding metrics (Figure 2, xcorr and expect). It should be noted that current methods of
191 scoring PSMs do not consider the rarity of given fragment ions, and instead compute a score or
192 probability based on the distribution of observed completeness of fragmentation for all pair-wise
193 comparisons [4-6, 34-38]. It is reasonable to assume that some sequences may be identifiable with
194 few fragments forming a rare combination, while other peptides with a shared or common
195 fragmentation pattern require a near complete representation; JAMSE exploits this relative
196 uniqueness of fragmentation patterns. Additionally, it should be noted that not all experimental
197 objectives require complete sequencing of peptides; many research objectives seek a balance
198 between PSM accuracy and overall yield depending on experimental goals [39-41].

199 The use of the decoy free method for FDR estimation comes from the need to avoid additional
200 computational resources required to search a reverse sequence database; as the size of the database

201 grows the chance of having iso-spectral overlap increases. Additionally, there is uncertainty in the
202 randomized and shuffled decoy creation process to generate natural sequences, after all only a
203 small fraction of all possible sequence combinations from observation and predicted transitions
204 have been catalogued.

205 Considering the UniProt database (Table 3), JAMSE yielded more PSMs than Comet for a given
206 decoy-free eFDR value (Figure 3a, blue and red solid lines respectively), demonstrating that PSMs
207 generated through JAMSE's probabilistic information retrieval had higher discriminative power
208 than did those from traditional cross-correlation. Although JAMSE found more PSMs, there is
209 very good agreement in the peptides that are identified by both methods; PSMs that differed across
210 the methods often contained the same peptide sequence and differed only in the PTM location
211 (data not shown). For example, at an eFDR (decoy-free) of 0.01, Comet yielded 20,755 PSMs,
212 while JAMSE yielded 23,943 PSMs, a substantial 15.4% gain, and a further look at the overlap in
213 PSM assignments showed 92.6% agreement at FDR 0.01, using the smaller UniProt database
214 (Figure 4, red line). Using the larger UniRef database, the agreement between search algorithms
215 is 89.6% at FDR 0.01, within range of similar comparisons by others [14].

216 Exploring the target-decoy eFDR for Comet (Figure 3a, orange solid line) there appears to be an
217 early advantage, over the decoy-free method (red solid line), that diminished quickly at larger
218 cutoffs. It is not known why this occurred, however, it is speculated that contributions from highly
219 similar peptides sequences present in the decoy-free method are more common, as the effect is
220 more pronounced in the larger UniRef search results. To note, JAMSE reports similar sequences
221 as alternates with a computed Bayesian posterior probability, akin to estimating the probability of
222 PTM location (data not shown).

223 When considering the larger UniRef database (Figure 3a, dashed lines) the effect of diminished
224 returns with larger databases is evident (dashed lines lower than solid lines of same color).
225 However, the effect is larger for the decoy-free method (orange lines) than for the target-decoy
226 method (blue and red lines). This difference may have been due to a shift to smaller peptides in
227 the target-decoy's randomized protein sequence database (data not shown) further exacerbated by
228 the highly homologous nature of the UniRef100 database. The generation of the randomized
229 target-decoy protein sequences tends to produce 10% more peptides of sequence length 6 to 16.
230 Consequently, nearly half of all PSMs for the target-decoy yeast dataset (52%) are between 6 and
231 16 residues in length (data not shown). This presents a risk in using the target-decoy approach, as
232 FDRs may become inflated due to the proportional increase in short peptide decoy sequences
233 relative to the target database [42].

234 The results for the mFDR approach (Figure 3b) are a strong commentary on the effectiveness of
235 the decoy-free over the target-free eFDR methods. For any eFDR method, an mFDR should match,
236 and, given the ability to compute q-values, a plot of mFDR against eFDR should follow the identity
237 line (y = x: grey line). However, the lines depicting the relationship between mFDR and eFDR for
238 the target-decoy approach (orange lines) indicated an overestimation of the eFDR. As above, this
239 may have been due to the bias in peptide sequence lengths in randomized protein decoy databases
240 and other effects noted in literature [43, 44]. On the other hand, the decoy-free approach for both
241 Comet and JAMSE (Figure 3b, red and blue respectively) demonstrated a closer agreement
242 between eFDR and mFDR (Table 2) with no observable difference given database size, unlike that
243 observed for the target-decoy method.

Conclusions

This work shows that probabilistic models in information recall can competitively and accurately identify peptides regardless of the size of the database. JAMSE provides a significant improvement for researchers attempting to address peptide sequencing in increasingly complex samples, such as the human proteome collected during clinical research and other genetically diverse sets. Peptide identification strategies have demonstrated weaknesses and therefor it is prudent to continue the development of alternatives [44] and in particular, continue to improve upon the initial PSM correlation metric, of which may enhance the post-search validation and machine learning methods. Providing a platform to expose all known sequence variants to direct comparison can augment these methods. Essentially providing an orthogonal method to unlocking the full translational potential of proteomics.

Despite JAMSE's ability to leverage larger databases, we realize there is also a fundamental limitation to controlling the error rate, especially when larger collections of independent hypothesis are utilized [11, 45, 46]. A veritable Catch-22, increase the search space to identify rarer peptides and decrease the overall yield post FDR analysis. While there are more sophisticated methods for peptide validation, the work described herein is seen as a compliment to post-identification machine learning validation [15, 47] and methods aimed at spectral reconstruction [27]. In addition, this work is not intended to challenge precision mass measurements, rather, given the effective conversion to unit resolution via the association to fragment mass defect distributions (e.g. FMI), this method may improve the yield and accuracy from lower resolution instruments typically utilized in high-throughput clinical research.

Lastly it is hypothesized, that although not specifically demonstrated here, that JAMSE directly solves the efficiency inherent in searching a very large and homologous redundant sequence space [6, 48, 49], since JAMSE's underlying technology leverages common indexing and search methods which can be distributed among many computer cores, or a compute cloud in parallel. Additionally, precalculated databases like JAMSE's also introduce a path towards real-time identifications [50], since search times of less than 10 milliseconds per spectrum are easily achieved given the computational resources and database sizes used here. Future implementations could be used in tandem LCMS data-driven acquisition methods that incorporate peptide sequence identification.

Acknowledgments

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Figures

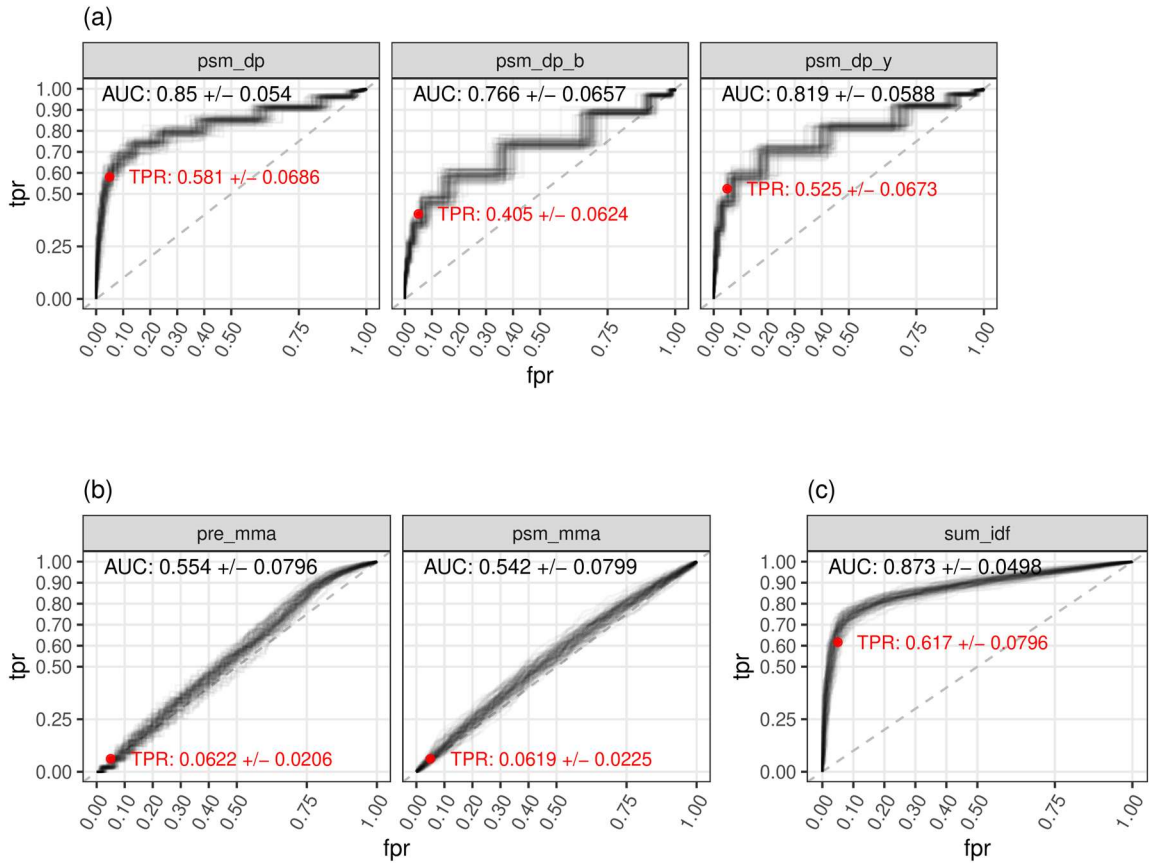

Figure 1. ROC plots for the discriminatory power of 100 rounds of a thousand randomly selected ms2 scans for the (a) dot-product between observed and in-silico, for all predicted FMIs (psm_dp), just the y fragment FMIs (psm_dp_y) and just the b fragment FMIs (psm_dp_b), (b) the absolute value of precursor mass measurement accuracy (pre_mma) and the absolute value of mean of fragment mass accuracy (psm_mma), (c) the ranking function selected for JAMSE, the sum of all term's inverse document frequency (sum_idf).

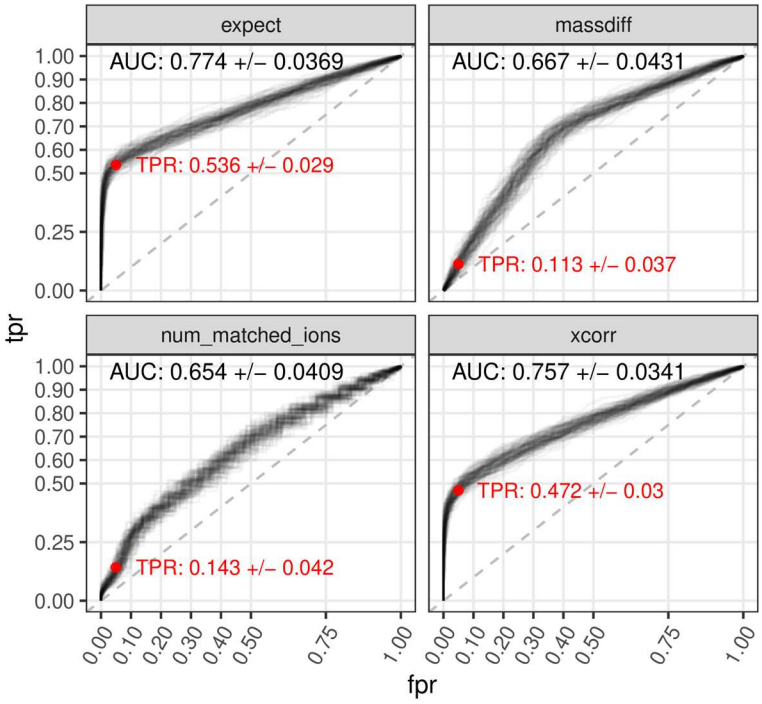

Figure 2. ROC plots for the discriminatory power of 100 rounds of a thousand randomly selected ms2 scans for selected the Comet metrics describing the absolute value of precursor mass measurement accuracy (massdiff), the number of matched fragment ions between observed and in-silico spectra (num_matched_ions), the cross-correlation coefficient (xcorr), and the computed expectation value (expect).

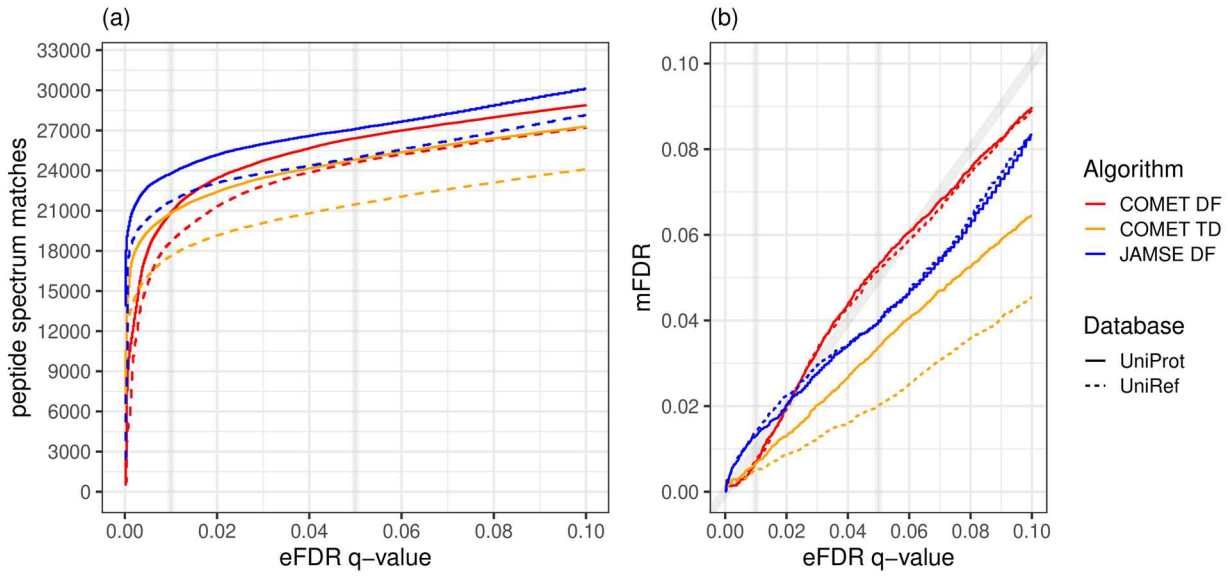

Figure 3. (a) Plot of PSMs recalled as a function of the FDR (based on Comet xcorr and JAMSE sum_idf) given the databased used for searching, either the conservative UniProt (solid line) consensus sequences or the combined UniRef and TrEMBL (dashed line) protein sequences. (b) Comparison between eFDR and mFDR where the solid grey line represents the function x=y, where the method of FDR estimation (eFDR) matches that of the species-specific mFDR.

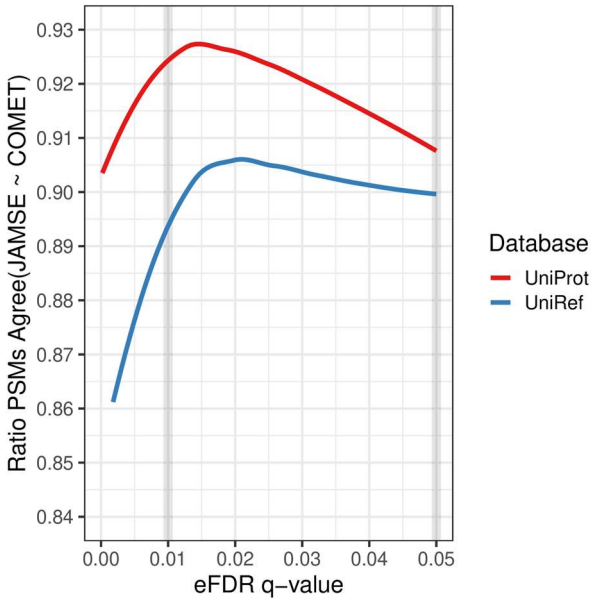

Figure 4. Plot of the fractional agreement between Comet and JAMSE as a function of the eFDR (decoy-free method) given the databased used for searching, either the conservative UniProt (red line) consensus sequences or the combined UniRef100 and TrEmbl (blue line).

309 # Tables

310 *Table 1. Accounting of proteins and peptides in two distinct databases computationally*
311 *constructed from UniProt repositories [21]. Peptides are generated by modeling tryptic cleavage,*
312 *with the variable PTM of acetylation on K and N-term.*

Designation	**Source**	**Organism**	**Proteins**	**Unique Peptides w/ PTMs**	
				Individual	Union
UniProt	UniProt (Jan 2019)	Saccharomyces sp.	7,980	3,701,494	11,493,070
		Arabidopsis sp.	15,878	7,813,095	
UniRef	UniRef100, TrEMBL (Jan 2019)	Saccharomyces sp.	58,829	11,352,240	36,757,734
		Arabidopsis sp.	120,863	25,493,700	

313

314

315 *Table 2. Table detailing the proteomic search metrics evaluated for both JAMSE and Comet. Each*

316 *of these metrics are evaluated by ROC in Figure 1 and Figure 2.*

Source	Metric	Description
Comet	xcorr	cross correlation coefficient
	expect	calculated expectation value
	massdiff	the absolute value of the difference between the in-silico precursor mass and the observed precursor mass
	num_matched_ions	the total number of fragment ions that align given the mass tolerance
JAMSE	pre_mma	the absolute value of the difference between the in-silico precursor mass and the observed precursor mass
	psm_mma	the mean absolute value of the difference between the in-silico fragment masses and the observed fragment masses
	psm_dp	the total number of fragment ions that align given the mass tolerance
	psm_dp_y	the total number of y series fragment ions that align given the mass tolerance
	psm_dp_b	the total number of b series fragment ions that align given the mass tolerance
	sum_idf	inverse document frequency summed across each term

317

318 *Table 3. Table comparing the eFDR and associated mFDR results for both Comet and JAMSE*
319 *using the target-decoy and decoy-free methods. The column designated dFDR denotes the*
320 *difference between mFDR and eFDR.*

Database	eFDR	Method	Algorithm/Method	PSMs	mFDR	dFDR
UniProt	0.01	target-decoy	Comet	21,404	0.007	-0.003
		decoy-free	Comet	20,755	0.005	-0.005
			JAMSE	23,943	0.011	0.001
	0.05	target-decoy	Comet	24,639	0.028	-0.022
		decoy-free	Comet	25,796	0.045	-0.005
			JAMSE	26,768	0.041	-0.008
UniRef (UniRef100 & TrEmbl)	0.01	target-decoy	Comet	18,265	0.004	-0.006
		decoy-free	Comet	19,279	0.006	-0.004
			JAMSE	21,816	0.013	0.003
	0.05	target-decoy	Comet	21,750	0.017	-0.033
		decoy-free	Comet	24,293	0.046	-0.004
			JAMSE	24,658	0.041	-0.009

321

322

References

[1] J. Griss, Y. Perez-Riverol, S. Lewis, D.L. Tabb, J.A. Dianes, N. Del-Toro, M. Rurik, M. Walzer, O. Kohlbacher, H. Hermjakob, R. Wang, J.A. Vizcano, Recognizing millions of consistently unidentified spectra across hundreds of shotgun proteomics datasets, Nature Methods 13(8) (2016) 651--656.

[2] J.M. Chick, D. Kolippakkam, D.P. Nusinow, B. Zhai, R. Rad, E.L. Huttlin, S.P. Gygi, A mass-tolerant database search identifies a large proportion of unassigned spectra in shotgun proteomics as modified peptides, Nature Biotechnology 33(7) (2015) 743--749.

[3] D.J.C. Pappin, D.M. Creasy, J.S. Cottrell, D.N. Perkins, Probability-based protein identification by searching sequence databases using mass spectrometry data., Electrophoresis 20(18) (1999) 3551--67.

[4] L.Y. Geer, S.P. Markey, J.A. Kowalak, L. Wagner, M. Xu, D.M. Maynard, X. Yang, W. Shi, S.H. Bryant, Open Mass Spectrometry Search Algorithm, Journal of Proteome Research 3(5) (2004) 958--964.

[5] J.K. Eng, A.L. McCormack, J.R. Yates, An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database, Journal of the American Society for Mass Spectrometry 5(11) (1994) 976-989.

[6] K. Verheggen, H. Rder, F.S. Berven, L. Martens, H. Barsnes, M. Vaudel, Anatomy and evolution of database search engines-a central component of mass spectrometry based proteomic workflows, Mass Spectrometry Reviews (April 2016) (2017).

[7] K. Strimmer, A unified approach to false discovery rate estimation, BMC bioinformatics 9(1) (2008) 303.

[8] L. Kall, J.D. Storey, M.J. MacCoss, W.S. Noble, Assigning significance to peptides identified by tandem mass spectrometry using decoy databases, J Proteome Res 7(1) (2008) 29-34.

[9] G. Alves, A.Y. Ogurtsov, Y.K. Yu, Assigning statistical significance to proteotypic peptides via database searches, J Proteomics 74(2) (2011) 199-211.

[10] A. Keller, A.I. Nesvizhskii, E. Kolker, R. Aebersold, Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search, Analytical Chemistry 74(20) (2002) 5383--5392.

[11] J.D. Storey, A direct approach to false discovery rates, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 64(3) (2002) 479-498.

[12] L. Kall, J.D. Storey, M.J. MacCoss, W.S. Noble, Posterior error probabilities and false discovery rates: two sides of the same coin, J Proteome Res 7(1) (2008) 40-4.

[13] J.E. Elias, S.P. Gygi, Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry, Nature Methods 4(3) (2007) 207--214.

[14] B.M. Balgley, T. Laudeman, L. Yang, T. Song, C.S. Lee, Comparative Evaluation of Tandem MS Search Algorithms Using a Target-Decoy Search Strategy, Molecular \& Cellular Proteomics 6(9) (2007) 1599--1608.

[15] G. Gonnelli, M. Stock, J. Verwaeren, D. Maddelein, B. De Baets, L. Martens, S. Degroeve, A decoy-free approach to the identification of peptides, Journal of proteome research 14(4) (2015) 1792-1798.

[16] W.S. Noble, Mass spectrometrists should search only for peptides they care about, Nature Methods 12(7) (2015) 605--608.

[17] A. Sticker, L. Martens, L. Clement, Mass spectrometrists should search for all peptides, but assess only the ones they care about, Nature methods 14(7) (2017) 643-644.

[18] W.S. Noble, U. Keich, Response to Mass spectrometrists should search for all peptides, but assess only the ones they care about, Nature Methods 14(7) (2017) 644.

[19] R.J. Chalkley, When target–decoy false discovery rate estimations are inaccurate and how to spot instances, Journal of proteome research 12(2) (2013) 1062-1064.

[20] C.E. Shannon, A mathematical theory of communication, Bell system technical journal 27(3) (1948) 379-423.

[21] S.E. Robertson, K.S. Jones, Relevance weighting of search terms, Journal of the American Society for Information science 27(3) (1976) 129-146.

[22] K.S. Jones, S. Walker, S.E. Robertson, A probabilistic model of information retrieval: development and comparative experiments: Part 2, Information processing & management 36(6) (2000) 809-840.

[23] L. Anderson, C.L. Hunter, Quantitative Mass Spectrometric Multiple Reaction Monitoring Assays for Major Plasma Proteins, Molecular \& Cellular Proteomics 5(4) (2006) 573--588.

[24] M.P. Washburn, D. Wolters, J.R. Yates III, Large-scale analysis of the yeast proteome by multidimensional protein identification technology, Nature biotechnology 19(3) (2001) 242.

[25] A. Bairoch, R. Apweiler, The SWISS-PROT protein sequence data bank and its supplement TrEMBL, Nucleic acids research 25(1) (1997) 31-36.

[26] B.E. Suzek, H. Huang, P. McGarvey, R. Mazumder, C.H. Wu, UniRef: comprehensive and non-redundant UniProt reference clusters, Bioinformatics 23(10) (2007) 1282-1288.

[27] S. Gessulat, T. Schmidt, D.P. Zolg, P. Samaras, K. Schnatbaum, J. Zerweck, T. Knaute, J. Rechenberger, B. Delanghe, A. Huhmer, Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning, Nature methods 16(6) (2019) 509.

[28] A.V. Nefedov, I. Mitra, A.R. Brasier, R.G. Sadygov, Examining Troughs in the Mass Distribution of All Theoretically Possible Tryptic Peptides, Journal of Proteome Research 10(9) (2011) 4150--4157.

[29] E. Kendrick, A Mass Scale Based on CH2= 14.0000 for High Resolution Mass Spectrometry of Organic Compounds, Analytical Chemistry 35(13) (1963) 2146-2154.

[30] I. Mitra, A.V. Nefedov, A.R. Brasier, R.G. Sadygov, Improved mass defect model for theoretical tryptic peptides, Anal Chem 84(6) (2012) 3026-32.

[31] S. Robertson, Understanding inverse document frequency: on theoretical arguments for IDF, Journal of documentation (2004).

[32] A. Aizawa, An information-theoretic perspective of tf–idf measures, Information Processing & Management 39(1) (2003) 45-65.

[33] S. Robertson, H. Zaragoza, The probabilistic relevance framework: BM25 and beyond, Now Publishers Inc2009.

[34] J.K. Eng, T.A. Jahan, M.R. Hoopmann, Comet: an open-source MS/MS sequence database search tool, Proteomics 13(1) (2013) 22-24.

[35] S. Kim, P.A. Pevzner, MS-GF+ makes progress towards a universal database search tool for proteomics, Nature Communications 5(1) (2014) 5277.

[36] V. Dorfer, P. Pichler, T. Stranzl, J. Stadlmann, T. Taus, S. Winkler, K. Mechtler, MS Amanda, a universal identification algorithm optimized for high accuracy tandem mass spectra, Journal of proteome research 13(8) (2014) 3679-3684.

[37] D.N. Perkins, D.J. Pappin, D.M. Creasy, J.S. Cottrell, Probability-based protein identification by searching sequence databases using mass spectrometry data, ELECTROPHORESIS: An International Journal 20(18) (1999) 3551-3567.

[38] R. Craig, R.C. Beavis, TANDEM: Matching proteins with tandem mass spectra, Bioinformatics 20(9) (2004) 1466--1467.

[39] J.S. Cottrell, D.M. Creasy, Response to: the problem with peptide presumption and low Mascot scoring, Journal of proteome research 10(11) (2011) 5272-5273.

[40] B. Cooper, The problem with peptide presumption and low Mascot scoring, Journal of proteome research 10(3) (2011) 1432-1435.

[41] S.A. Carr, S.E. Abbatiello, B.L. Ackermann, C. Borchers, B. Domon, E.W. Deutsch, R.P. Grant, A.N. Hoofnagle, R. Httenhain, J.M. Koomen, D.C. Liebler, T. Liu, B. MacLean, D.R. Mani, E. Mansfield, H. Neubert, A.G. Paulovich, L. Reiter, O. Vitek, R. Aebersold, L. Anderson, R. Bethem, J. Blonder, E. Boja, J. Botelho, M. Boyne, R.A. Bradshaw, A.L. Burlingame, D. Chan, H. Keshishian, E. Kuhn, C. Kinsinger, J.S.H. Lee, S.-W. Lee, R. Moritz, J. Oses-Prieto, N. Rifai, J. Ritchie, H. Rodriguez, P.R. Srinivas, R.R.a. Townsend, Targeted Peptide Measurements in Biology and Medicine: Best Practices for Mass Spectrometry-based Assay Development Using a Fit-for-Purpose Approach, Molecular \& Cellular Proteomics 13(3) (2014) 907--917.

[42] N. Gupta, N. Bandeira, U. Keich, P.A. Pevzner, Target-decoy approach and false discovery rate: when things may go wrong, Journal of the American Society for Mass Spectrometry 22(7) (2011) 1111-1120.

[43] Y. Danilova, A. Voronkova, P. Sulimov, A. Kertész-Farkas, Bias in false discovery rate estimation in mass-spectrometry-based peptide identification, Journal of proteome research 18(5) (2019) 2354-2358.

[44] Y. Couté, C. Bruley, T. Burger, Beyond target-decoy competition: stable validation of peptide and protein identifications in mass spectrometry-based discovery proteomics, bioRxiv (2019) 765057.

[45] Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: a practical and powerful approach to multiple testing, Journal of the Royal statistical society: series B (Methodological) 57(1) (1995) 289-300.

[46] Y. Benjamini, Discovering the false discovery rate, Journal of the Royal Statistical Society: series B (statistical methodology) 72(4) (2010) 405-416.

[47] L. Käll, J.D. Canterbury, J. Weston, W.S. Noble, M.J. MacCoss, Semi-supervised learning for peptide identification from shotgun proteomics datasets, Nature methods 4(11) (2007) 923.

[48] J.S. Cottrell, Protein identification using MS/MS data, Journal of Proteomics 74(10) (2011) 1842--1851.

[49] D. Di Silvestre, S. Daminelli, P. Brunetti, P. Mauri, Bioinformatics tools for mass spectrometry-based proteomics analysis, Reviews in Pharmaceutical and Biomedical Analysis-BENTHAM SCIENCE PUBLISHERS 3 (2010) 30-52.

[50] D.K. Schweppe, J.K. Eng, D. Bailey, R. Rad, Q. Yu, J. Navarrete-Perea, E.L. Huttlin, B.K. Erickson, J.A. Paulo, S.P. Gygi, Full-featured, real-time database searching platform enables fast and accurate multiplexed quantitative proteomics, bioRxiv (2019) 668533.