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

Abstract

Mass spectrometry methods of peptide identification involve comparing observed tandem spectra with in silico derived spectrum models. Here we present a proteomics search engine that offers a new variation of the standard approach, with improved results. Our proposed method employs information theory and probabilistic information retrieval on a pre-computed fragmentation database generating a peptide-to-spectrum match (PSM) score modeled on the rarity of the observed fragmentation pattern. Fast and accurate PSM matches are achieved yielding a 10-15% higher rate of peptide identities than current methods. Applications of this search engine are aimed at identifying peptides from large databases consisting of sequence variations, such as homologous proteins expected within an environmental collection of a genetically variable species.

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, and splice variation, in addition to the myriad of biologically potential post translational modifications (PTMs). One key to unlocking the translational potential of proteomics lies in computational methods that accommodate these complexities.

Current applications of tandem MS proteomic search engines compare observed spectra with *in silico* modeled spectra, calculating correlation values by probabilistic methods [1, 2] or by inferred scores based on applications of signal processing and information theory [3]. Published examples of proteomic search algorithms compare each observed spectrum pairwise with all *in silico* spectra, ranking the comparisons by score [4]. From the collection of all rank-one peptide-to-spectrum-matches (PSMs) a false discovery rate (FDR) can be estimated [5-10] 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 [6, 8, 11, 12], and a more recent decoy-free method using a null distribution comprised of alternative target PSMs [13].

The use of FDR estimations is absolutely prudent to research for MS-based proteomics, however there are potential hidden issues associated with randomized decoy sequences employed as the estimated null distribution for FDR calculations [13-16]. Additionally, these effects are likely to exacerbate as the size and complexity of both the target and decoy database increases. This is particularly challenging for researchers attempting to address the inherent protein diversity beyond the typical consensus proteome, ideally including known isoforms and splice variants that arises from genetic and translation variation, or within studies examining multiple variable post-translational modifications. Methods such as Percolator [17] and Nokoi [13] attempt to remedy this by employing machine learning classifiers designed to discriminate between target and decoy by combining PSM scores with additional metrics, although with potential questionable

success[18]. It is therefore prudent to continue the development of alternatives [19] and in particular, continue to improve upon the initial PSM correlation metric.

We present a new proteomic search engine, JAMSE ("Just Another Molecular Search Engine"), that utilizes a precalculated fragmentation database of all non-redundant proteotypic peptides to derive PSMs with a highly discriminative scoring metric. During the pre-computation process individual frequencies and patterns derived from fragment and precursor ions are learned, such that the search strategy is based on both information theory [20] and probabilistic information retrieval [21, 22]. The goal of this approach is to derive a PSM score modeled on the rarity of a fragmentation pattern, as opposed to completeness of the observed spectrum. Indeed, this is not an entirely new concept, as the field of proteomic selected reaction monitoring exploits the rarity in the combination of elution time, precursor mass and a selected fragment mass [23].

For benchmark comparisons, we utilize a decoy-free null distribution for estimating the FDR and introduce the concept of an intrinsically measurable FDR by including protein sequences from an orthogonal species. The pre-computed nature of JAMSE renders a parallel search against a randomized decoy peptide sequences unfeasible, as not all pairwise target peptides are explored. Results, when compared to current approaches in proteomic search engines, indicate a more discriminating PSM correlation metric can be achieved. JAMSE is optimized for Linux, suitable for cloud distribution, and is available at https://www.jamse.app/.

Methods

Data

The yeast MS proteome (*Saccharomyces cerevisiae*) collected by Washburn *et.al.*[24] and made publicly available, was used as a benchmark data set. Briefly, it consists of 64,321 tandem liquid chromatography (LC) MS spectra, derived from a series of orthogonal LC separations coupled to tandem MS experiments designed to deeply probe the yeast proteome.

Databases

Protein sequence databases for JAMSE and Comet searches were 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 was 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 here) 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 here), 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 were built for each of the two protein databases by randomly shuffling each protein sequence.

Pre-computed peptide fragmentation databases were generated based on the known proteome and utilizing typical 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. The resulting collection of non-redundant peptides was transformed further into a predicted fragmentation database with respect to *in silico* derived patterns, considering both *y-ion* and *b-ion* fragments along with metastable water-loss ions. The fragmentation database served as the main search object wherein observed spectra where queried against.

Search

Database searching was accomplished by querying a set of precursor and fragment mass values, typically filtered by abundance according to similar methods [2], for a given spectrum. The top five PSMs are returned, ranked by a probabilistic score derived from a modified weighted model function [22]. Of important note, JAMSE spectrum searching does not consider all pairwise comparison between the observed spectrum and relevant *in silico* spectra. Rather, much like a modern index search engine, top hits are produced via a modified method of information retrieval [21].

For this study, all searches were limited to a 0.1 Dalton precursor neutral mass tolerance and a 0.5 Dalton fragment measured mass tolerance, with the top 5 hits retained per spectrum. Additionally, the Comet [27] search algorithm, specifically, the metric *xcorr*, was used as a method of comparison. For computational comparison, Comet was run on a single computer utilizing 16 cores and consuming approximately 1.1GB of memory. JAMSE was distributed across two computers utilizing 8 cores each requiring approximately 1.8GB of memory per machine.

FDR Estimations

The estimated FDR (eFDR) was computed either from the target-decoy approach as described by Keller *et.al.*[8], or as the decoy-free approach described by Gonnelli *et.al.* [13]. For the purposes of validating a given eFDR, we here introduce the concept of empirically measured FDR (mFDR). Both eFDR and mFDR methods provide means to define the null distribution and calculate the FDR q-value according to the procedure proposed by Storey *et.al.*[9].

mFDR calculation

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

Search performance evaluation

Evaluation of search performance was based on the total number of PSMs obtained for a given eFDR q-value, with larger numbers considered more desirable. Evaluation of eFDR performance

was based on the comparison to the complimentary mFDR, with a closer match to mFDR considered better.

Results

For the smaller UniProt database, there were more JAMSE matches than Comet matches for a given decoy-free eFDR value (Figure 1a, blue and red solid lines respectively), demonstrating that PSMs generated through JAMSE's probabilistic information retrieval had higher discriminative power than did those from traditional cross-correlation. Although JAMSE found more PSMs, there was very good agreement in the peptides that were identified by both methods; PSMs that differed across the methods often contained the same peptide sequence and differed only in the PTM location (data not shown). For example, at an eFDR (decoy-free) of 0.01, Comet yielded 20,755 PSMs, while JAMSE yielded 23,943 PSMs, a substantial 15.4% gain, and a further look at the overlap in PSM assignments showed 92.6% agreement at FDR 0.01, using the smaller UniProt database (Figure 2, red line). Using the larger UniRef database, the agreement between search algorithms was 89.6% at FDR 0.01, within range of similar comparisons by others [12].

Exploring the target-decoy eFDR for Comet (Figure 1a, orange solid line) there appeared to be an early advantage, over the decoy-free method (red solid line), that diminished quickly at larger cutoffs. It is not known why this occurred, however, we speculate that contributions from highly similar peptides sequences present in the decoy-free method were more common, as the effect was more pronounced in the larger UniRef search results. To note, JAMSE reports similar sequences as alternates with a computed Bayesian posterior probability, akin to estimating the probability of PTM location (data not shown).

When considering the larger UniRef database (Figure 1a, dashed lines) the effect of diminished returns with larger databases was evident (dashed lines lower than solid lines of same color). However, the effect was larger for the decoy-free method (orange lines) than for the target-decoy method (blue and red lines). This difference may have been due to a shift to smaller peptides in the target-decoy's randomized protein sequence database (data not shown), further exacerbated by

the highly homologous nature of the UniRef100 database. The generation of the randomized target-decoy protein sequences tends to produce 10% more peptides of sequence length 6 to 16. Consequently, nearly half of all PSMs for the target-decoy yeast dataset (52%) were between 6 and 16 residues in length (data not shown). This presents a risk in using the target-decoy approach, as FDRs may become inflated due to the proportional increase in short peptide decoy sequences relative to the target database. We theorize that the target-decoy method would benefit from a randomized non-redundant peptide database, however, given the effectiveness of the decoy-free method we did not invest the effort to investigate.

The results for the mFDR approach (Figure 1b) were a strong commentary on the effectiveness of the decoy-free over the target-free eFDR methods. For any eFDR method, an mFDR should match, and, given the ability to compute q-values, a plot of mFDR against eFDR should follow the identity line (y = x: grey line). However, the lines depicting the relationship between mFDR and eFDR for the target-decoy approach (orange lines) indicated an overestimation of the eFDR. As above, this may have been due to the bias in peptide sequence lengths in randomized protein decoy databases and other effectes noted in literature [18]. On the other hand, the decoy-free approach for both Comet and JAMSE (Figure 1b, red and blue respectively) demonstrated a closer agreement between eFDR and mFDR (Table 2), with no observable difference given database size, unlike that observed for the target-decoy method.

Conclusions

Search results using JAMSE demonstrate that a highly discriminant PSM score can be constructed using a model derived from information theory and probabilistic information retrieval. Current methods of scoring PSMs do not consider the rarity of given fragment ions, and instead compute a score or probability based on the distribution of fragment completeness for all pair-wise comparisons [2-4, 27-31]. It is reasonable to assume that rare sequences may be identifiable with few fragments while other peptides with a shared or common fragmentation pattern require a near complete representation; JAMSE exploits this relative uniqueness of fragmentation patterns. Additionally, it should be noted that not all experimental objectives require complete sequencing of peptides; many seek a balance of PSM accuracy and counts depending on experimental goals [32-34].

While there are obvious benefits and needs to estimating the FDR, there are risks associated with certain methods as previously highlighted [16, 18, 33], with some researchers proposing new methods to overcome the limitations of target-decoy FDR estimations [35]. The research presented here sides with Gonnelli *et.al.* [13], concluding that a separate decoy database is unwarranted, and the use of a decoy-free eFDR more accurately estimates an empirically derived mFDR. While the work by Gonnelli *et.al.* was aimed at a machine learning application to improve PSM rates, they did not explicitly set out to prove the validity of their proposed FDR method. It is the use of mFDR computation proposed here that validates the utility of a decoy-free FDR null distribution, and furthermore exposes that the target-decoy method generally overestimates the eFDR, subsequently under-reporting the final number of PSMs.

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 in clinical research and other genetically diverse sets.

Although not specifically demonstrated here, JAMSE directly solves the efficiency inherent in searching a very large and homologous redundant sequence space [4, 36, 37], since JAMSE's precomputation can be distributed among many computer cores in parallel. Additionally, precalculated databases like JAMSE's also introduce a path towards real-time identifications [38], since search times of less than 10 milliseconds per spectrum were 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.

Figures

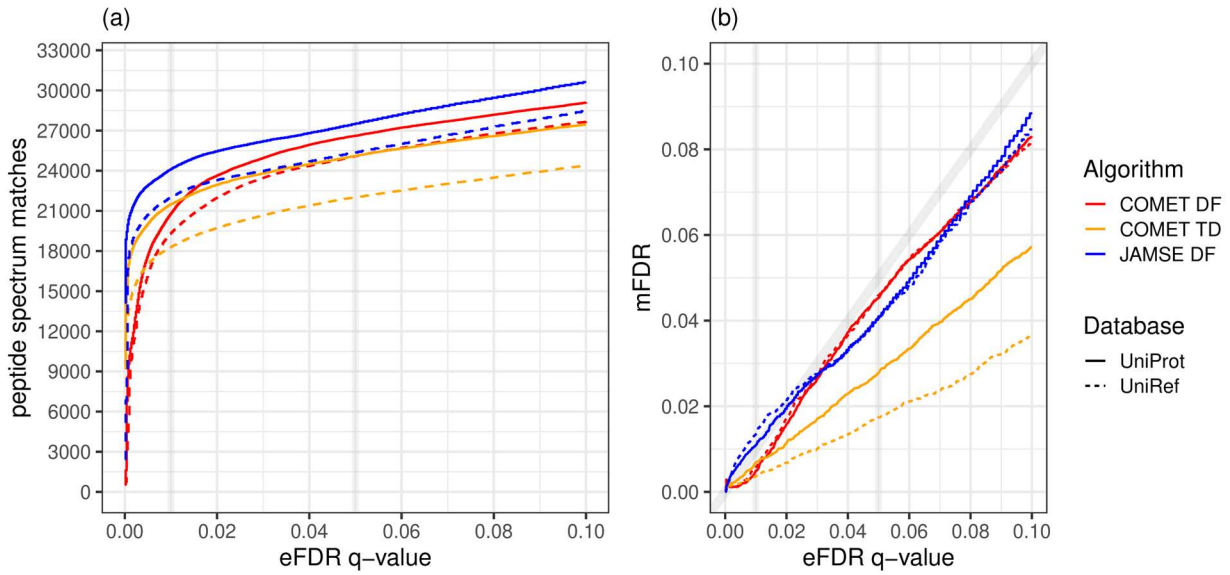

Figure 1. (a) Plot of PSMs recalled as a function of the FDR (based on Comet xcorr and JAMSE psm_score) given the databased used for searching, either the conservative UniProt (solid line) consensus sequences or the combined UniRef and TrEMBL (dashed line) protein sequences. (b) Comparison between eFDR 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.

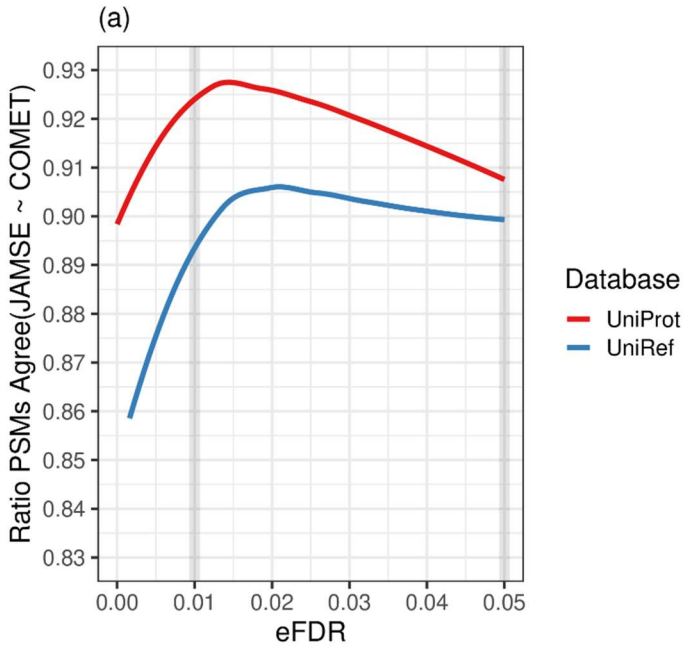

Figure 2. Plot of the fractional agreement between Comet and JAMSE as a function of the eFDR (decoy-free method) given the databased used for searching, either the conservative UniProt (red line) consensus sequences or the combined UniRef100 and TrEmbl (blue line)

Tables

Table 1. Accounting of proteins and peptides in two distinct databases computationally constructed from UniProt repositories [21]. Peptides were generated by modeling tryptic cleavage, 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	

Table 2. Table comparing the eFDR and associated mFDR results for both Comet and JAMSE using the target-decoy and decoy-free methods. The column designated dFDR denotes the 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

References

1. Pappin, D.J.C., et al., Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis, 1999. **20**(18): p. 3551--67.

2. Geer, L.Y., et al., *Open Mass Spectrometry Search Algorithm.* Journal of Proteome Research, 2004. **3**(5): p. 958--964.

3. Eng, J.K., A.L. McCormack, and 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, 1994. **5**(11): p. 976-989.

4. Verheggen, K., et al., Anatomy and evolution of database search engines-a central component of mass spectrometry based proteomic workflows. Mass Spectrometry Reviews, 2017(April 2016).

5. Strimmer, K., *A unified approach to false discovery rate estimation.* BMC bioinformatics, 2008. **9**(1): p. 303.

6. Kall, L., et al., Assigning significance to peptides identified by tandem mass spectrometry using decoy databases. J Proteome Res, 2008. **7**(1): p. 29-34.

7. Alves, G., A.Y. Ogurtsov, and Y.K. Yu, *Assigning statistical significance to proteotypic peptides via database searches.* J Proteomics, 2011. **74**(2): p. 199-211.

8. Keller, A., et al., Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. Analytical Chemistry, 2002. **74**(20): p. 5383--5392.

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

10. Kall, L., et al., Posterior error probabilities and false discovery rates: two sides of the same coin. J Proteome Res, 2008. **7**(1): p. 40-4.

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

12. Balgley, B.M., et al., Comparative Evaluation of Tandem MS Search Algorithms Using a Target-Decoy Search Strategy. Molecular \& Cellular Proteomics, 2007. **6**(9): p. 1599--1608.

13. Gonnelli, G., et al., *A decoy-free approach to the identification of peptides.* Journal of proteome research, 2015. **14**(4): p. 1792-1798.

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

15. Noble, W.S., Mass spectrometrists should search only for peptides they care about. Nature Methods, 2015. **12**(7): p. 605--608.

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

17. Käll, L., et al., Semi-supervised learning for peptide identification from shotgun proteomics datasets. Nature methods, 2007. **4**(11): p. 923.

18. Danilova, Y., et al., Bias in false discovery rate estimation in mass-spectrometry-based peptide identification. Journal of proteome research, 2019. **18**(5): p. 2354-2358.

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

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

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

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

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

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

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

26. Suzek, B.E., et al., *UniRef: comprehensive and non-redundant UniProt reference clusters.* Bioinformatics, 2007. **23**(10): p. 1282-1288.

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

28. Kim, S. and P.A. Pevzner, MS-GF+ makes progress towards a universal database search tool for proteomics. Nature Communications, 2014. **5**(1): p. 5277.

29. Dorfer, V., et al., MS Amanda, a universal identification algorithm optimized for high accuracy tandem mass spectra. Journal of proteome research, 2014. **13**(8): p. 3679-3684.

30. Perkins, D.N., et al., Probability-based protein identification by searching sequence databases using mass spectrometry data. ELECTROPHORESIS: An International Journal, 1999. **20**(18): p. 3551-3567.

31. Craig, R. and R.C. Beavis, *TANDEM: Matching proteins with tandem mass spectra.* Bioinformatics, 2004. **20**(9): p. 1466--1467.

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

33. Cooper, B., *The problem with peptide presumption and low Mascot scoring.* Journal of proteome research, 2011. **10**(3): p. 1432-1435.

34. Carr, S.A., et al., Targeted Peptide Measurements in Biology and Medicine: Best Practices for Mass Spectrometry-based Assay Development Using a Fit-for-Purpose Approach. Molecular \& Cellular Proteomics, 2014. **13**(3): p. 907--917.

35. Coute, Y., C. Bruley, and T. Burger, Beyond target-decoy competition: stable validation of peptide and protein identifications in mass spectrometry-based discovery proteomics. bioRxiv, 2019: p. 765057.

36. Cottrell, J.S., *Protein identification using MS/MS data.* Journal of Proteomics, 2011. **74**(10): p. 1842--1851.

37. Di Silvestre, D., et al., *Bioinformatics tools for mass spectrometry-based proteomics analysis.* Reviews in Pharmaceutical and Biomedical Analysis-BENTHAM SCIENCE PUBLISHERS, 2010. **3**: p. 30-52.

38. Schweppe, D.K., et al., Full-featured, real-time database searching platform enables fast and accurate multiplexed quantitative proteomics. bioRxiv, 2019: p. 668533.