Visualizing the agreement of peptide assignments between different search engines

Valkenborg Dirk (Orcid ID: 0000-0002-1877-3496)

Visualizing the agreement of peptide assignments between different search engines.

Annelies Agten[1], Joris Van Houtven[1,2,3], Manor Askenazi[4], Tomasz Burzykowski[1], Kris Laukens[5,6], Dirk Valkenborg[1,2,3]

[1]Interuniversity Institute of Biostatistics and Statistical Bioinformatics, Hasselt University, Hasselt, Belgium
[2]UA-VITO Center for Proteomics, University of Antwerp, Antwerp, Belgium
[3]Applied Bio and Molecular Systems, Flemish Institute for Technological Research (VITO), Mol, Belgium
[4]Biomedical Hosting LLC, Arlington, USA.
[5]Adrem Data Lab, Department of Mathematics and Computer Science, University of Antwerp, Antwerp, Belgium
[6]Biomedical Informatics Network Antwerp (biomina), University of Antwerp, Antwerp, Belgium

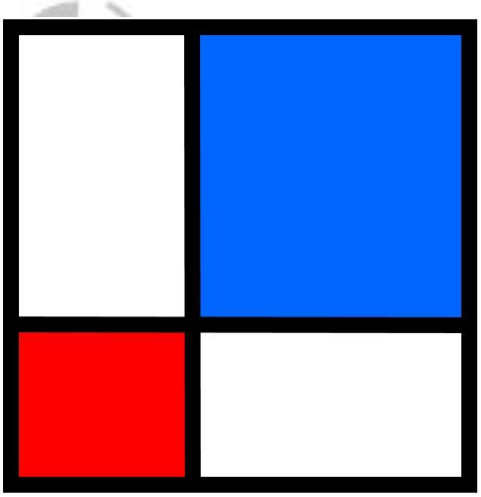
ABSTRACT

There is a trend in the analysis of shotgun proteomics data that aims to combine information from multiple search engines to increase the number of peptide annotations in an experiment. Typically, the degree of search engine complementarity and search engine agreement is visually illustrated by means of Venn diagrams that present the findings of a database search on the level of the non-redundant peptide annotations. We argue this practice to be not fit-for-purpose since the diagrams do not take into account and often conceal the information on complementarity and agreement at the level of the spectrum identification. We promote a new type of visualisation that provides insight on the peptide sequence agreement at the level of the peptide-spectrum-match (PSM) as a measure of consensus between two search engines with nominal outcomes.

We applied the visualizations and percentage sequence agreement to an in-house dataset of our benchmark organism, C. elegans and illustrate that when assessing the agreement between search engine one should disentangle the notion of PSM confidence and PSM identity. The visualizations presented in this manuscript provide a more informative assessment of pairs of search engines and are made available as an R-function in the supplementary materials.

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/jms.4471

TOC



Letter to Al

Dear Al,

I still remember well when we first met. It was at ASMS2011 - I just finalized a review on methods for the computation of isotope distributions and dedicated an entire section on the multinomial expansion to which Yergey and others have contributed a lot. When I noticed your nametag amidst the ASMS crowd, I decided to quickly introduce myself to you. The short introduction became a long conversation and marked the start of a friendship/mentorship and a yearly tradition to meet at ASMS to talk about science, life, and our aspirations.

It was during one of these ASMS dinners that we shared our frustration about the use of Venn diagrams for comparing results between multiple search engines, mainly because they provide a false intuition of search engine complementarity. We argued that rather than increasing the number of peptide identifications by combining search engines results, it would be wiser to focus on sequence agreement when conditioning on the spectrum level. A delineation of this discussion along with a new visualization method to replace the Venn diagrams is presented in this issue as an educational article in your memory.

Can you still recollect our first words? I kept on talking about the manuscript that is also reprinted in this special issue, entitled: “Isotopic distributions in mass spectra of large molecules” and commented on how sharp-witted it was of you to see the link between the multinomial expansion and the calculation of isotope distributions. After my glorification of this article, you kind-heartedly notified that if I would have better studied the author list, I could have noticed that the article was from the hand of James, your brother. This situation still makes me smile when I think about you, but the interesting thing is that back then at ASMS, I started to talk to the wrong Yergey. Call it coincidence, destiny or serendipity, I am glad that we have met that day and it was an honour to have known you and have spent some time with you.

Yours sincerely,

Dirk Valkenborg

INTRODUCTION

Shotgun proteomics has proven to be a powerful technology for identifying proteins in a complex biological sample.[1] To interpret the tremendous amount of fragment ion spectra produced by tandem mass spectrometry (MS/MS) experiments, correlative database search algorithms are often employed. There exists a large variety of correlative search engines to identify peptides using MS/MS[2,3], for which exhaustive overviews are available in the literature.[4-6] These different search algorithms all vary in accuracy, sensitivity and specificity, because they are based on various underlying scoring mechanisms and have different methodological implementations.

This heterogeneity of principles used in search engines causes a variation in peptide identification, the choice of the search algorithm plays a role in the identification process of peptides and their corresponding proteins.[7] To improve the robustness and confidence in peptide and protein identifications, studies suggest using approaches that combine the results from two or more search engines in terms of a meta-score. Most of these studies have shown that such approaches can greatly enhance mass spectral coverage and specificity, compared to the use of a single search algorithm.[8-19] An overview of the methods for combining multiple search engines is provided by Shteynberg et al.[18]

A popular method to visualize the outcome of multiple search engines is the use of Venn diagrams, displaying the overlap and complementarity for the protein/peptide findings between pairs of search engines. This approach, however, simply employs a comparison of two or more sets of accession numbers or non-redundant peptide sequences, without taking into consideration information from the spectrum level. In this manuscript, we illustrate that the concept of complementarity on the protein/peptide level as defined by the Venn diagram view is flawed. Apart from improving the number of peptide identifications by using multiple search engines, a more interesting metric to consider is the degree of agreement on the peptide annotation at the spectrum level. Ignoring the agreement of peptide identification on the spectrum level becomes an issue when assessing search engine concordance, because the set of possible protein identifications is ultimately restricted by the finite set of proteins present in the database.[20]

In order to construct a Venn diagram, for each search engine a list that contains all confidently assigned, non-redundant peptide annotations is created. In the scenario depicted in Figure 1, there are 3458 non-redundant peptide sequences found in common by Mascot and Sequest. The Mascot search engine contributes 510 additional confident peptide identifications, whilst Sequest adds 1360 additional confident identifications. However, when further investigating these results it may well be that both search engines find the same set of non-redundant peptide sequences, but they do not reach a consensus when conditioning on the spectrum level. This hypothetical scenario is depicted in the table shown in Figure 1 by the coloured cells, where there is agreement in terms of identifying non-redundant peptide sequences, but disagreement in the peptide-to-spectrum-matches (PSMs) when considering the scan numbers. Although in this simplistic example no agreement is reached between the search engines when considering the spectrum level, in contrast, on the peptide set-level, some of these identifications contribute to the intersection area of the Venn diagram. A different type of disagreement is represented by the cells indicated by grey colours. In this case, sequences are only found by one of the search engines. These identifications will contribute to the set of differences or complimentary findings, although, strictly speaking, at the spectrum level both search engines tend to disagree about the peptide annotation for that particular spectrum.

Arguably, the simplistic and hypothetical example presented in Figure 1 is not completely realistic or exhaustive. When multiple search engines are combined, it is possible that spectra will fail to receive a (confident) identification. Missing spectrum annotations would further complicate a concordance analysis and should be taken into account.

Based on the aforementioned considerations, we conclude that the practice of summarizing mass spectrometry-based proteomics experiments by Venn diagrams is prone to misleading interpretation when combining results from multiple search engines. We argue that the agreement and disagreement should be assessed at the level of the spectrum and not on the level of the set of non-redundant peptide identifications, as the latter obscures information about search engine reliability. Therefore, we propose a new visualization that provides this information at first glance. Additionally, we would like to promote the use of different variants of percentage agreement to summarize agreement between a pair of search engines.

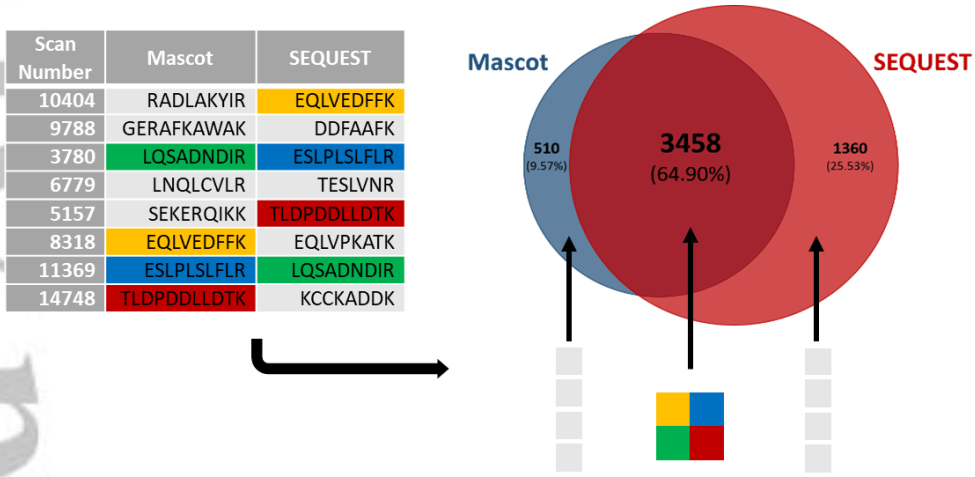


Figure 1: Hypothetical scenario on how disagreement at the level of scan numbers will contribute to the Venn diagram that represents findings at the level of peptide sequence information.

MATERIALS AND METHODS

PEPTIDE IDENTIFICATION

A spectral dataset of our benchmark organism C. elegans was interpreted by SearchGUI version 3.2.20 as a workflow manager. The 20,581 tandem MS spectra were searched by Comet[21,22] and MS-GF+[23] against a *C. elegans* protein database (ws229 wormbase) downloaded at ftp://ftp.wormbase.org/pub/wormbase/species/c_elegans/assemblies/. Precursor mass tolerance was set at 10 ppm, while fragment mass tolerance was set to 0.5 Da. A maximum of five missed cleavages by trypsin was allowed for. A static modification of 57.021 Da on cysteine was defined to account for carbamidomethylation and a dynamic modification of 15.9949 Da was introduced to account for possible oxidation of methionine. Monoisotopic masses were used for the precursor mass. Furthermore, only first ranked PSMs were considered for analysis, i.e., only one sequence annotation for each fragment ion mass spectrum is retained. The negative logarithm of the MS-GF+ and Comet E-value was adopted as the MS-GF+ and Comet score. For Sequest and Mascot respectively the xcorr and ion score were used. A target-decoy (TD) search strategy was applied on a separate database search with reversed peptide sequences to determine the cut-off values by N_{decoy}/N_{target} for a global PSM-level false-discovery rate (FDR) of 5% for both search engines.[24] A similar workflow using Proteome Discoverer 1.3 was followed to obtain peptide identifications by Mascot[3] and Sequest[2].

Note that we adopted an FDR of 5% for illustrative purposes, while usually an FDR of 1% is deemed more appropriate for proteomic experiments.

PERCENTAGE SEQUENCE AGREEMENT

The main question that was addressed in this research was whether there is an agreement in terms of the peptide-to-spectrum assignments between Comet and MS-GF+ on the one hand and Mascot and Sequest on the other. When analyzing the agreement between two search engines, the concept of sequence agreement in terms of the peptide annotation tends to be conflated with the notion of sequence confidence. In this study we take into account sequence confidence by dichotomizing the confidence score in categories of peptide-spectrum-matches that do or do not pass the false discovery rate (FDR) cutoff. This creates three different scenarios, namely where both, only one or neither one of the search engines confidently annotates the spectrum. To account for the confidence that is assigned to a peptide annotation, sequence agreement - i.e. whether spectra get annotated with the same peptide by both search engines - is discussed conditionally on the three described scenarios. To assess sequence agreement, we calculate the percentage sequence agreement as the proportion of spectra that received the same peptide annotation from both search engines.[25] Although percentage agreement has the advantage of being computationally simple and easy to interpret, it does not allow for the fact that a certain amount of agreement can be expected on the basis of chance alone.[26] This is important to remark because, for example, large proteins that have many peptides are more likely to be found by different engines. Cohen proposed a measure of agreement that corrects for such chance findings.[27] However, application of Cohen's kappa to a mass spectrometry dataset faces the issue of the presence of a large number of distinct categories, i.e., peptide sequences. The use of Cohen's kappa is further elaborated on in detail in the supplementary materials, but we already can mention that the correction for chance findings in the shotgun proteomics setting is negligible. As such, the use of the kappa coefficient boils down to percentage sequence agreement.

MISSING ANNOTATIONS

In general, database-searching engines identify only a proportion of the MS/MS spectra of digested proteins. Due to differences in their implementation, various search engines can handle low-scoring spectra in a vastly distinct way, leading to a discrepancy between the spectra annotated by the different algorithms. In our analysis we ignore missing annotations and filter out spectra that are not identified by both search engines. This filtering step can be perceived as a quality threshold, removing non-confident and most likely conflicting identifications, but also induces some information loss. A more detailed breakdown of the missingness patterns can be found in the supplementary materials.

RESULTS AND DISCUSSION

In total, when FDR filtering is disabled, Sequest provided 20,233 peptide-spectrum-matches. Among them, 12,128 (59%) were also identified by Mascot. Ignoring missing annotations means that 8105 spectra that initially received a Sequest identification are filtered out, or alternatively were set to disagree at a MASCOT score of 0 (no confidence). By contrast, both Comet and MS-GF+ assigned a peptide sequence to most of the available MS/MS spectra when disregarding any confidence threshold. In our experiment, 361 and 362 out of the 20,581 spectra did not receive an identification from MS-GF+ and Comet, respectively. From these non-identified spectra, 328 were missing in both Comet and MS-GF+ results, yielding 20,186 spectra identified by both search engines. The percentage sequence agreement across all of these spectra is equal to 45.06% and 78.89%, for the Comet-MS-GF+ and Sequest-Mascot comparison respectively. It should be remarked that the higher agreement for

the Sequest-Mascot case is due to the missing annotations, which probably reflects an internal quality threshold for reporting.

COMPARISON OF SEARCH ENGINES UNDER A TARGET-DECOY APPROACH

In order to study the influence of confidence (i.e., pass the FDR criteria) on the sequence agreement, the results from the conducted MS/MS experiment are presented in Figure 2. In the scatterplots in the left panels, the identification scores provided by both search engines for each spectrum are plotted on the x- and y-axis. The barplots in the right panels contain a summary of the spectral agreement, related to the corresponding scatterplots. In what follows, different scenarios are proposed for comparing sequence agreement in function of the confidence score, i.e. we discuss search engine concordance in terms of the different quadrants that are indicated in Figure 2.

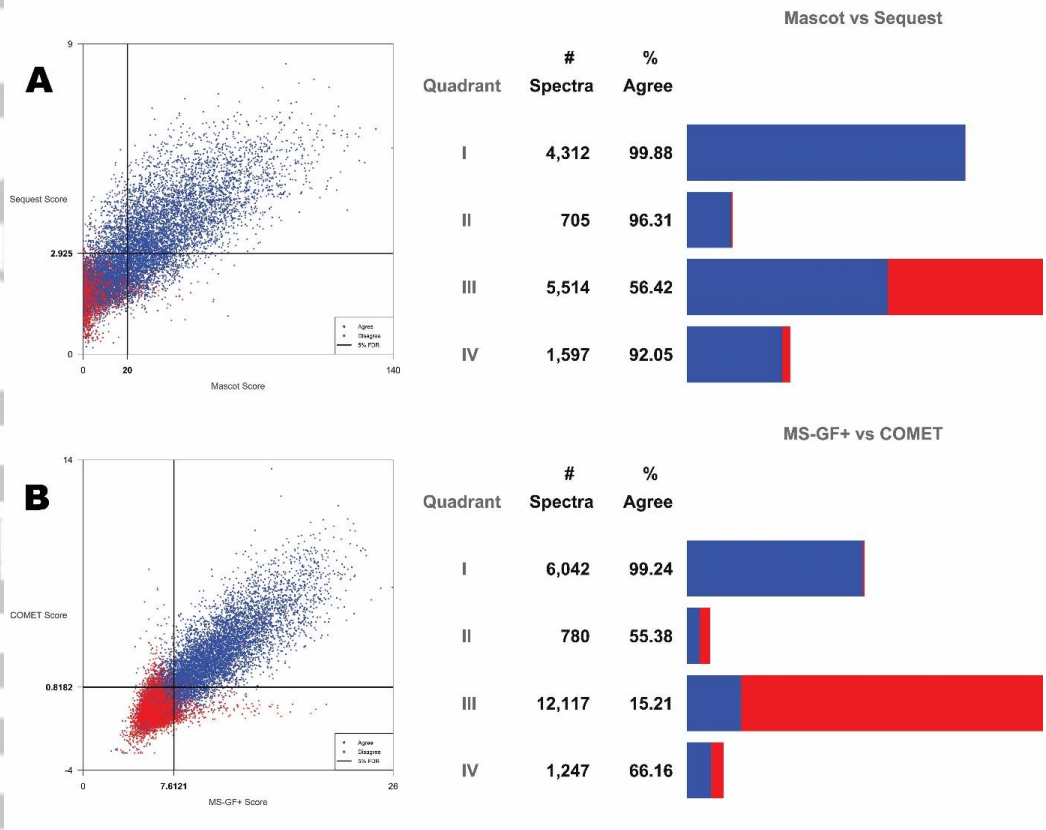


Figure 2: (Left) Scatterplot summarization of the rank 1 search engine results for all spectra between 2 search engines. Every dot in the plot represents a spectrum that is investigated by the search engines. The score of one search engine is denoted on the x-axis, the score of the other search engine is denoted on the y-axis. The colors give a binary indication whether the sequence annotation agrees (blue) or disagrees (red) among the search engines. The horizontal and vertical black lines indicate, in this case, the user-specified 5% FDR cutoff. Quadrant I (top-right) visually represents the percentage sequence agreement of the spectra that receive a confident score from both search engines. Spectra that only comply with an FDR of 5% for one of the two search engines are depicted in quadrants II (top-left) and IV (bottom-right). Spectra that receive an identification with a score below the FDR cut-off for both algorithms are compared in terms of sequence agreement by quadrant III (bottom-left). Spectra that did not receive a score from a search engine have been removed in this plot. **(Right)** Barplot and textual summarization of the percentage sequence agreement analysis from the comparison between two search engines. The quadrants in the barplot correspond to the quadrants indicated in the scatterplot in the left panel. The length of the bars is relative to the number of spectra in the respective quadrant. The colors indicate sequence agreement (blue) or disagreement (red). Spectra that did not receive a score from a search engine have been removed from the analysis. **(A)** Comparison of Mascot and Sequest. **(B)** Comparison of Comet and MS-GF+.

STUDYING THE INFLUENCE OF THE CONFIDENCE RANKING

The obtained results are first compared on the set of confident PSMs in quadrant I, i.e., the peptide identification results that comply with an FDR of 5% for both search engines according to the TD approach described in the materials section. There are 6,042 fragment ion spectra that received a score above the FDR threshold of 5% from both Comet and MS-GF+ (Quadrant I). It is observed that out of these 6,042 spectra, 5,996 (99.24%) agree in sequence annotation between both search engines. Similarly, there are 4,312 fragment ion spectra that received a score above the FDR threshold from both Mascot and Sequest. The percentage sequence agreement amongst these spectra is equal to 99.88%.

It is reassuring that, in general, the dots (i.e., spectra) in quadrant I are overwhelmingly blue (i.e., sequence agreement), but the red dots (I.e., disagreement) are the most acute cases. One red dot in quadrant I would lead to two contributions in a Venn diagram and are therefore the most alarming. These insights are not provided by the area-proportional Venn diagram presentation.

The most interesting quadrants in Figure 2, however, are quadrants II and IV. They indicate spectra that received a confident annotation by only one of the two search engines. For example, there were 1,247 spectra that received a confident ranking by MS-GF+ but that failed to receive a confident score from Comet. Among these spectra, which are represented by the bottom-right quadrant in the scatterplot in Figure 2B, 825 (66.16%) agreed in identification for both search engines. Analogously, there were 780 spectra that received a confident ranking by Comet but that scored below the FDR cutoff for MS-GF+. Among these spectra, which are represented by the top-left quadrant in the scatterplot of Figure 2B, 432 (55.38%) agreed in identification for both search engines.

Our motivation in focusing on the details of quadrants II and IV can best be summed up by (a slightly paraphrased) Tolstoy: “Happy PSMs are all alike; every unhappy PSM is unhappy in its own way”. The original version of this quote, often referred to as the Anna Karenina Principle, suggests that failures are multi-faceted and diverse, whereas success is uniform. In our case, the implication is that two search engines can fail to agree because they produce different peptide identifications, but they can also fail to agree for other reasons, namely because they do not achieve a sufficient degree of confidence on the underlying peptide identification.[28] In contrast, when focusing on quadrants II and IV in Figure 2A for the comparison between Mascot and Sequest, the parts that are considered as a complementary contribution from the multiple search strategy, it can be seen that the sequence annotations are in high agreement but one of the search engines lacks courage to call it a confident finding.

For completion, quadrant III in Figure 2A and 2B constitutes the spectra that failed to receive a confident annotation from either search engine. There are 12,117 spectra that scored below the 5% FDR cut-off for both MS-GF+ and Comet. The percentage sequence agreement on this set of spectra equals 15.21%. Similarly, for the comparison between Sequest and Mascot, 5,514 spectra were scored below the FDR threshold, of which the percentage sequence agreement in terms of sequence annotation equals 56.42%.

AGREEMENT BY PEPTIDE SUMMERIZATION AT THE SEQUENCE LEVEL

In what preceded we have presented the results regarding percentage sequence agreement on the spectrum level for each of the quadrants in Figure 2 separately. We argue that such a representation is much more informative than the sequence set analysis by the Venn diagrams in Figure 3. For example, the search engine agreement presented in Figure 3 is based on the set of peptide annotations that receive a confident score by at least one the search engines, i.e. the set of spectra comprised of quadrants I, II and IV. In what follows, we contrast the agreement at the level of the set of non-redundant peptide annotations (Figure 3) to the percentage sequence agreement on the spectrum level (Figure 2).

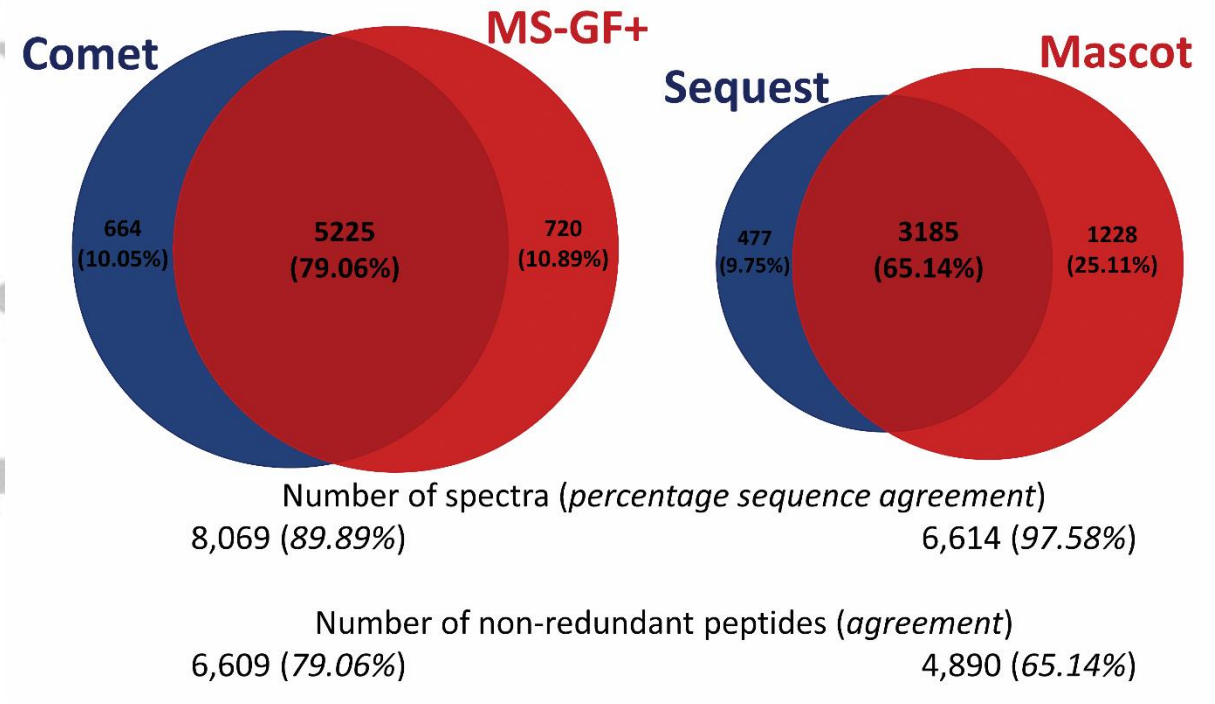


Figure 3: Venn diagrams that represent the agreement on the set of non-redundant peptide annotations comprised of quadrants I, II and IV. Notice the discrepancy between the agreement on the set of peptides and the percentage sequence agreement on the spectrum level.

There are 8,069 spectra that receive a confident annotation by either Comet, MS-GF+ or both. These spectra constitute 6,609 unique peptide annotations. Notice that 79.06% of the peptide annotations are found by both search engines, which contributes to the intersection of the Venn diagram. Additionally, Comet and MS-GF+ account for 10.05% and 10.89% of the peptide annotations that could not be confidently identified by both search engines. Analogously, there are 6,614 spectra confidently identified by Sequest and Mascot. The identifications of these spectra are composed of 4,890 non-redundant peptide annotations, of which 65.14% are found in common by Sequest and Mascot. Sequest and Mascot separately contribute 9.75% and 25.11% additional peptide annotations. The metrics from the Venn diagram can be contrasted with the percentage sequence agreement on the spectral level for quadrants I, II and IV. The percentage sequence agreement for the respective Comet-MS-GF+ and Mascot-Sequest comparison are equal to 89.89% and 97.58% suggesting a large agreement instead of the complementarity suggested by Figure 3. By no means the results presented in the barplots in Figure 2, can be inferred from the Venn diagram mainly due to the presence of redundant peptide annotations. Another effect that complicates the inference of Figure 2 from Figure 3 was already announced in the introduction. For example, a confident sequence disagreement in

quadrant I can result in the irrational effect of a complementarity in the Venn diagram at the level of non-redundant peptide annotations. On the other hand, any findings in quadrant II or IV, whether search engines agree or disagree, does not necessarily lead to a complementarity in the finding at the level of non-redundant peptide annotations as these peptide findings can be covered by another peptide-spectrum-match. As previously discussed, the comparison of search engines conflates two different questions, namely:

1. Is there an agreement in sequence annotation between search engines?
2. Are the search engines (both) confident about the sequence annotation?

It is possible that search engines assign the same sequence annotations to a spectrum, but one search engine is very confident about this assignment, whilst the other lacks courage to make the call. This discrepancy was made clear especially in the comparison between Comet and MS-GF+. On the other hand, there might also be spectra that are confidently identified by both search engines, but that do not agree on their peptide annotation. The analysis in this manuscript demonstrates that Venn diagrams provide only a restricted view on search engine agreement and/or complementarity. Instead we could use the visualizations proposed in Figure 2 in order to make a more complete evaluation about search engine agreement.

CONCLUDING REMARKS

In this manuscript, we propose the concept of percentage sequence agreement on peptide identification at the spectrum level to assess agreement in the context of database search engines instead of summarizing results at the level of peptides or proteins by means of a Venn diagram. When assessing the agreement between search engines, it should be clear where and how confidently they agree. Typically, in database searches this is a combination of both sequence annotation and sequence confidence. Therefore, we propose to visualize the outcome of a database search by the Mondrian-like plots presented in Figure 2 that divide the search engine results into four quadrants depending on the search engine's confidence. The percentage sequence agreement and some simple summary statistics can be computed for each of these quadrants as depicted in the barplot of Figure 2. From these summary statistics, it is trivial to compute an overall percentage sequence agreement for quadrants I, II and IV by means of a weighted average.

Besides, the more in-depth view on search engine concordance for the quadrants individually, provided by Figure 2, these plots gives us the opportunity to associate a notion of quality on the peptide annotation if classical database search engines are used, i.e., no 2nd pass or chimeric spectra search engines. There are four scenarios possible:

1. Green flag (safely passes QC):
    a. The search engines strongly agree on a peptide annotation (blue dots in quadrant I).
    b. The search engines weakly agree on a peptide annotation (blue dots in quadrant II and IV). It means that both search engines recognize the same peptide in the presented fragment ion series but one of the engines is not confident about its finding.
2. Orange flag (warrants further investigation):
    a. The search engines weakly disagree on a peptide annotation (red dots in quadrant II and IV), it means that the search engines assign different peptides to the observed ion fragments, but one of the search engines is not confident about its finding. Therefore, the identification of the uncertain search engine can be discarded.

3. Red flag (does not pass QC): The search engines strongly disagree on a peptide annotation (red dots in quadrant I). Action should be undertaken to untie the conflicting identification. Provided aforementioned assumptions about chimeric spectra and standard search engines, a spectrum should not give rise to two identifications.
4. White flag (unidentified spectra): The findings in quadrant III are not of concern since both search engines are uncertain about the peptide annotation. Nevertheless, in case they agree on a peptide annotation, this could hint towards the presence of explanatory peptide ions in the spectrum.

For the example of MS-GF+ and Comet, the aforementioned numbers can be easily obtained from Figure 2 as follows:

- #spectra(Green) = 0.9924 * 6043 + 0.5538 * 780 + 0.6616 * 1247 = 5997 + 432 + 825 = 7254 spectra
- #spectra(Orange) = (1-0.5538) * 780 + (1-0.6616 * 1247) = 348 + 422 = 770 spectra
- #spectra(Red) = (1-0.9924) * 6042 = 46 spectra

We want to emphasise that the minimalistic QC approach presented here should not be over-interpreted. Sequence agreement and spectral confidence depend on the selected search engines and since we only consider two search engines at the time, a severe selection bias in the QC results can occur. This selection bias can be remediated by a sensitivity analysis on the spectra for multiple pairwise search engine comparisons and summarizing the QC results for a more robust interpretation. Another disadvantage of the binary approach is that we lose the fine-grained information present in the score distributions of the search engines. A possible solution is to revert to probabilistic models that compute a meta-score for the quality of the identification on a continuous scale. For example, the methods of Shteynberg et al. and Searle et al. allow for such a meta-analysis approach.[17,18] Such probabilistic models yield two major advantages compared to the simple analysis presented in this manuscript. Firstly, the probabilistic meta-score is able to register the nuances around the FDR threshold, making it slightly less sensitive to the selection of the search engine pairs. Secondly, when extending the visual quality control to more than three search engines, the interpretation of the plots becomes complicated. A meta-score can easily incorporate information from multiple search engines, making it less vulnerable for the selection bias due to the pairwise comparison.

Furthermore, it should be noted that missing annotations, especially when they only occur for one of the search engines, complicate the agreement analysis. In our analysis, Mascot fails to reports a score for a substantial proportion of spectra. However, it can be shown (supplementary materials) that the spectra that do not receive a score by either of the search engines, mainly produce low-scoring identifications by the other search engine and imply some internal quality filter. Nonetheless these missing scores and annotations would land mainly in quadrant III and would thus not substantially influence the agreement analysis when not taken into account. A more thorough approach would be to perform a sensitivity analysis by random imputation of the missing score and agreement. Alternatively, a worst case approach could be considered that sets the missing score to the minimum score and the annotation is assumed to disagree. Latter findings could be contrasted with the complete case as reported in Figure 2.

For illustrative purposes, we have considered only 2 sets of different search engines in this manuscript, namely Sequest and Mascot as the two most employed commercial database search engines to-date and Comet and MS-GF+ as two open-source search algorithms available through SearchGUI. However, our illustrations are not restricted to the cases presented in this paper and different comparisons, e.g. Comet-Mascot or other search engines, e.g. Andromeda can be employed. Moreover, when

comparing multiple search engines, the new visualisation approach is easy to generalize up to three search engines. For area-proportional circular Venn diagrams such an extension is usually impossible[29]. For higher dimension the comparison becomes visually difficult and other metrics have to be computed such as a generalisation of the percentage agreement or Cohen's kappa. The R script to produce the scatterplots and barplots of Figure 2 for any combination of two search engines is provided in the supplementary information. We should also note that in our analysis, we have used both raw match scores (for the Mascot-Sequest comparison) and probability-based scores (for the Comet-MS-GF+ comparison). Our visualization approach is flexible and can operate on any type of score, however, comparisons should always be made at the same score types as non-monotone transformations could skew the scatter plot data. Preferably, a probabilistic score is used to compare results across different search engines. For example, when investigating Figure 2A, one can observe that the 5% FDR threshold is not well aligned with the region of sequence disagreement (red dots in quadrant III), whereas in Figure 2B the FDR nicely separates the quadrants with dominant sequence agreement and disagreement.

To conclude, the take-home message of this manuscript is that the nature of disagreement among search engines should be verified, i.e. whether disagreement originates from having the same ID, but low confidence versus a true divergence of opinion. This divergence can be visualized with the proposed Figure 2 with the same effort as the Venn diagrams at the peptide level, but the figures give a more informative view. These visualizations also combat the perverse effects of the Venn diagram summary that reward strong disagreement as high complementarity. Instead of striving towards complementarity one should strive for better search engine concordance and progress towards a universal search engine.

REFERENCES

1. Aebersold R, Mann M. Mass spectrometry-based proteomics. *Nature.* 2003;422(6928):198-207.
2. Eng JK, McCormack AL, Yates JR. 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):976-989.
3. Perkins DN, Pappin DJC, Creasy DM, Cottrell JS. Probability-based protein identification by searching sequence databases using mass spectrometry data. *ELECTROPHORESIS.* 1999;20(18):3551-3567.
4. Nesvizhskii AI. Protein Identification by Tandem Mass Spectrometry and Sequence Database Searching. In: Matthiesen R, ed. *Mass Spectrometry Data Analysis in Proteomics.* Totowa, NJ: Humana Press; 2007:87-119.
5. Kapp E, Schütz F. Overview of Tandem Mass Spectrometry (MS/MS) Database Search Algorithms. *Current Protocols in Protein Science.* 2007;49(1):25.22.21-25.22.19.
6. Hubbard SJ. Computational Approaches to Peptide Identification via Tandem MS. In: Hubbard SJ, Jones AR, eds. *Proteome Bioinformatics.* Totowa, NJ: Humana Press; 2010:23-42.
7. Sadygov RG, Cociorva D, Yates JR. Large-scale database searching using tandem mass spectra: Looking up the answer in the back of the book. *Nature Methods.* 2004;1(3):195-202.
8. Elias JE, Haas W, Faherty BK, Gygi SP. Comparative evaluation of mass spectrometry platforms used in large-scale proteomics investigations. *Nature Methods.* 2005;2(9):667-675.
9. Alves G, Wu WW, Wang G, Shen R-F, Yu Y-K. Enhancing Peptide Identification Confidence by Combining Search Methods. *Journal of Proteome Research.* 2008;7(8):3102-3113.
10. Chamrad DC, Körting G, Stühler K, Meyer HE, Klose J, Blüggel M. Evaluation of algorithms for protein identification from sequence databases using mass spectrometry data. *PROTEOMICS.* 2004;4(3):619-628.
11. Jones AR, Siepen JA, Hubbard SJ, Paton NW. Improving sensitivity in proteome studies by analysis of false discovery rates for multiple search engines. *PROTEOMICS.* 2009;9(5):1220-1229.
12. Kapp EA, Schütz F, Connolly LM, Chakel JA, Meza JE, Miller CA, Fenyo D, Eng JK, Adkins JN, Omenn GS, Simpson RJ. An evaluation, comparison, and accurate benchmarking of several publicly available MS/MS search algorithms: Sensitivity and specificity analysis. *PROTEOMICS.* 2005;5(13):3475-3490.
13. Kwon T, Choi H, Vogel C, Nesvizhskii AI, Marcotte EM. MSblender: A Probabilistic Approach for Integrating Peptide Identifications from Multiple Database Search Engines. *Journal of Proteome Research.* 2011;10(7):2949-2958.
14. Price TS, Lucitt MB, Wu W, Austin DJ, Pizarro A, Yocum AK, Blair IA, FitzGerald GA, Grosser T. EBP, a Program for Protein Identification Using Multiple Tandem Mass Spectrometry Datasets. *Molecular & Cellular Proteomics.* 2007;6(3):527.
15. Resing KA, Meyer-Arendt K, Mendoza AM, Aveline-Wolf LD, Jonscher KR, Pierce KG, Old WM, Cheung HT, Russell S, Wattawa JL, Goehle GR, Knight RD, Ahn NG. Improving Reproducibility and Sensitivity in Identifying Human Proteins by Shotgun Proteomics. *Analytical Chemistry.* 2004;76(13):3556-3568.
16. Searle BC. Scaffold: A bioinformatic tool for validating MS/MS-based proteomic studies. *PROTEOMICS.* 2010;10(6):1265-1269.
17. Searle BC, Turner M, Nesvizhskii AI. Improving Sensitivity by Probabilistically Combining Results from Multiple MS/MS Search Methodologies. *Journal of Proteome Research.* 2008;7(1):245-253.
18. Shteynberg D, Nesvizhskii AI, Moritz RL, Deutsch EW. Combining Results of Multiple Search Engines in Proteomics. *Molecular & Cellular Proteomics.* 2013;12(9):2383.
19. Yu W, Taylor JA, Davis MT, Bonilla LE, Lee KA, Auger PL, Farnsworth CC, Welcher AA, Patterson SD. Maximizing the sensitivity and reliability of peptide identification in large-scale

proteomic experiments by harnessing multiple search engines. *PROTEOMICS.* 2010;10(6):1172-1189.
20. Serang O, Käll L. Solution to Statistical Challenges in Proteomics Is More Statistics, Not Less. *Journal of Proteome Research.* 2015;14(10):4099-4103.
21. Eng JK, Hoopmann MR, Jahan TA, Egertson JD, Noble WS, MacCoss MJ. A Deeper Look into Comet—Implementation and Features. *Journal of The American Society for Mass Spectrometry.* 2015;26(11):1865-1874.
22. Eng JK, Jahan TA, Hoopmann MR. Comet: An open-source MS/MS sequence database search tool. *PROTEOMICS.* 2013;13(1):22-24.
23. Kim S, Pevzner PA. MS-GF+ makes progress towards a universal database search tool for proteomics. *Nature Communications.* 2014;5:5277.
24. Nesvizhskii AI. A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. *Journal of Proteomics.* 2010;73(11):2092-2123.
25. Berk RA. Generalizability of behavioral observations: a clarification of interobserver agreement and interobserver reliability. *Am J Ment Defic.* 1979;83(5):460-472.
26. Baer DM. REVIEWER'S COMMENT: JUST BECAUSE IT'S RELIABLE DOESN'T MEAN THAT YOU CAN USE IT. *Journal of Applied Behavior Analysis.* 1977;10(1):117-119.
27. Cohen J. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement.* 1960;20(1):37-46.
28. Tolstoy L. *Anna Karenina.* Russia: The Russian Messenger; 1877.
29. Micallef L, Rodgers P. eulerAPE: Drawing Area-Proportional 3-Venn Diagrams Using Ellipses. *Plos One.* 2014;9(7):e101717