UniScore, a unified and universal measure for peptide identification by multiple search engines

Tsuyoshi Tabata[1], Akiyasu C Yoshizawa[1], Kosuke Ogata[1], Chih-Hsiang Chang[2], Norie Araki[2], Naoyuki Sugiyama[1,3], Yasushi Ishihama[1,4]

1) Graduate School of Pharmaceutical Sciences, Kyoto University, Kyoto 606–8501, Japan
2) Graduate School of Medical Sciences, Kumamoto University, Kumamoto, 860-8556, Japan
3) Omics Research Center, National Cerebral and Cardiovascular Center, Suita, Osaka 564-8565, Japan
4) Laboratory of Proteomics for Drug Discovery, National Institute of Biomedical Innovation, Health and Nutrition, Ibaraki, Osaka 567-0085, Japan

E-mail ORCID

Tsuyoshi Tabata, yoshi.bata@gmail.com
Akiyasu C Yoshizawa 0000-0002-0870-5502, acyshzw@med.niigata-u.ac.jp
Kosuke Ogata 0000-0002-0634-3990, ogata.kosuke.7w@kyoto-u.ac.jp
Chih-Hsiang Chang, Orcid; 0000-0002-0029-5385, peio0105@gmail.com
Norie Araki, Orcid; 0000-0001-9987-6229, nori@gpo.kumamoto-u.ac.jp
Naoyuki Sugiyama 0000-0001-7633-8495, sugiyama.naoyuki@ncvc.go.jp
Yasushi Ishihama 0000-0001-7714-203X, yishiham@pharm.kyoto-u.ac.jp

Abstract

We propose UniScore as a metric for integrating and standardizing the results of multiple search engines for data-dependent acquisition data in LC/MS/MS-based bottom-up proteomics. UniScore can be calculated from only the matched product ion information by matching the peptide amino acid sequence and product ion spectrum, so it can handle large-scale data with minimal computing resources. When applied to large-scale global proteome data and phosphoproteome data, UniScore produced results that outperformed those obtained using conventional methods. Furthermore, UniScore could also be applied to peptide matching for chimeric spectra without any problems.

Introduction

The identification of amino acid sequences from peptide tandem mass spectra has been carried out for more than 30 years, and in the early days, *de novo* sequencing has been performed by reading the b- and y-ion series that are predominantly observed in low-energy collision-induced dissociation reactions in triple quadrupole and ion trap mass analyzers.[1] In the post-genome era, as protein amino acid sequence databases have become more extensive and very large numbers of spectra have been obtained using LC/MS/MS in data-dependent acquisition (DDA) mode, a number of search engines have been developed to identify peptides from tandem mass spectra, including Sequest[2] and Mascot[3], which are representative of the search engines that match MSMS spectra obtained experimentally against databases. In the early days, the threshold for identification was set based on the statistical scores specific to each search engine, but after the introduction of the target-decoy (TD) approach[4], identification criteria based on false discovery rate (FDR) became common. As public data repositories and open-source search engines (e.g. X! Tandem[5], OMSSA[6], Andromeda/MaxQuant[7,8], Byonic[9], MSGF+[10]) were developed, many large-scale studies were developed and attempts were made to create complete human proteome atlases. The first papers[11, 12] simply merged the results of a large number of different research groups, so many of the hit proteins were redundant, while the decoy hits just piled up, resulting in data quality that lacked reliability[13]. The problems with the target-decoy method was again pointed out as a problem, and an improvement in the FDR control method was

proposed[14]. Meanwhile, even without such large-scale data, attempts have also been made to increase the number of hit proteins by integrating the identification results from multiple search engines for a single raw data set. If the results of each search are simply merged, the rate of overlap between target hits will be higher than the rate of overlap between decoy hits, and this will result in a higher FDR than that set for each search. As a solution to this problem, a number of tools have been reported that merge the results of different search engines while controlling the FDR[15,16,17,18,19].

In another method called peptide sequencing tag (PST)[20], proteins are separated and digested, and the sample solution is analyzed by infusion-nanoESI-MS/MS without LC. Then, it requires a prior interpretation step in which the researcher assigns partial amino acid sequences obtained from MS/MS spectra by manual inspection, and in addition to the obtained stretch of the interpreted sequence, the mass of the precursor peptide, the mass of the first peak in the identified sequence ladder, and the mass of the last peak in the ladder are entered as query data for searching against the protein database. PST is potentially the most powerful search method[21]. However, it has rarely been applied to LC/MS/MS data because it was difficult to fully automate the step of prior interpretation of MS/MS spectra.

Search engines such as Sequest, Mascot, X! Tandem and MaxQuant have been applied not only to high-resolution MS/MS spectra from mass analyzers such as TOF and Orbitrap, but also to low-resolution MS/MS from ion trap. On the other hand, search engines that make use of the advantages of high-resolution MS/MS have also been developed recently. For example, Morpheus[22] uses the specificity obtained through high resolution mass to assign charge states and remove non-monoisotopic peaks in the spectrum preprocessing stage. After such pre-processing, a very simple score is used to identify the product, based on the sum of the number of matching products and the fraction of the spectrum abundance that can be assigned to matching products, using the target-decoy approach. Surprisingly, despite this simple score, it was found to outperform the results of conventional search engines. One concern with such a simplified score is that it lacks the probabilistic interpretation of the conventional scores. However, with the widespread adoption of target-decoy approach[4], PSM only needs to be relatively ordered, and the global and local FDRs can be empirically determined. Furthermore, a Python-based open-source analysis system for large high-resolution MS datasets called

AlphaPept has recently been reported[23]. In this system, X! tandem score, Morpheus score or a newly introduced generic score based on the peptide length, the total number of fragment hits, b-ion hits, and the matched intensity ratio, was used to do peptide-spectrum matching, and a Percolator-like machine learning approach was performed to maximize the peptide identification with FDR estimation. With large-scale public data now easily available, post-processor rescoring tools such as AlphaPept and PeptideProphet[24] are becoming increasingly important[25]. Furthermore, with the addition of universal spectral identifiers[26], it is becoming possible to search across projects[27]. Approaches that make full use of deep learning to generate predicted MS/MS spectra and identify peptides on a large scale using the similarity of experimental data (MSBooster[28]、Oktoberfest[29]、MS2Rescore[30], etc.) are being developed one after another, and the environment for integrating and re-analysing data is gradually being established. However, in order to analyze these large-scale data in an integrated manner, in many cases a large-scale computing environment is required, and there is still a need for the development of methods to facilitate the standardization and integration of large-scale data.

In this study, we propose a new scoring method for MSMS spectra that does not rely on search engines. This scoring method, which we have named UniScore, is a measure for selecting the most likely candidate peptide sequence from among the candidate peptide sequences, including decoy sequences proposed by various search engines. In addition to the most straightforward 'total number of matched b- and y-ions', we have also adopted the 'number of matched amino acid sequence stretches', which was inspired by the PST method. Here, we will demonstrate the optimization process of the UniScore algorithm, and then apply it to the high-depth global human proteome data and human phosphoproteome data registered in the public repository to demonstrate that the results are significantly improved compared to conventional methods.

Methods

Datasets used for UniScore development were obtained from jPOST repository (https://repository.jpostdb.org/) based on ProteomeXchange identifiers. The LC/MS/MS raw data files and its metadata used in the UniScore application were

obtained from the ProteomeXchange repositories and the original publication. The multiple search engines used and their versions, the versions of the protein databases used for the search, search parameters and other information are listed in Supplementary Table 1. For peak picking of the product ion spectra, we used the peak-picking function of MaxQuant to convert multiply charged ions to singly charged ions, declustered the isotope envelope to pick the monoisotopic peaks, recalibrated the observed *m/z* values using the pre-search results, and generated the mgf files. In some cases, we used ProteoWizard[28,31] with modifications[32] to do the same pre-processing for product ion spectra. These files were input into Comet, X! Tandem, Mascot and MaxQuant, and database searches were performed under the conditions shown in Supplementary Table 1 to obtain the respective identification results. Using the above mgf files, matching of the respective MSMS spectra and peptide candidates was performed with only the top 12 peaks per *m/z* 100 Th, and the matched b- and y-ion list was output for each PSM to calculate UniScores.

The UniScore calculation for each product ion spectrum is as follows; first, the matched b- and y-ions were counted. Based on these matched ions, the total number of amino acids flanked on both sides by these ions (called matched sequence stretch) was counted. Finally, the total number of b-ions, y-ions and matched sequences were simply added together, without weighting, and this was used as the UniScore in this paper. For cases where several different sequences were assigned as candidates in a single spectrum, they were treated as chimeric spectra and the UniScore was calculated for each sequence, except when more than 50% of the annotated peaks were shared.

For the calculation of FDR, PSMs were first ordered by descending UniScore, and for each decoy hit that appeared, the decoy hit occurrence rate was calculated from the number of forward hits and decoy hits up to that score, which was then used as the FDR. FDR at the peptide level was performed by removing redundancy in the sequence only, without considering modifications and charge, and the rest of the procedure was the same as for PSMs. For protein group FDRs, peptides were inferred to protein groups using the reported method[33], and then FDRs were obtained using the same method as above.

Results and Discussion

As test datasets for the development of UniScore, the result files of a single-shot global proteome analysis of HeLa cells using a 2-meter-long silica monolithic column to acquire high-resolution MSMS spectra in DDA mode with an 8-hour gradient using orbitrap were used. The data were acquired in five raw files, split every two hours, so that the UniScore can be evaluated independently for groups of peptides with four different physical properties such as precursor charge, peptide length, hydrophobicity and isoelectric point (Supplementary Figure 1). Analysis using MaxQuant resulted in identifying a total of 34,746 unique peptides and 4,367 protein groups.

When developing UniScore, the parameters that can be read from the MSMS spectrum of a matched peptide are the number of matched product ions (b, y-ion), the number of unmatched b, y-ions, the number of amino acids flanked on both sides by matched product ions (called matched sequence stretch), the peptide length of the precursor ion, the peptide mass, the fraction of the spectrum abundance that can be assigned to matched product ions, etc. The spectrum pre-processing includes the top N peak abundance filtering for each bin, the determination of the charge state of the product ion, and the accumulation algorithm to the monoisotopic ions by isotope declustering.

For this test dataset, we set the initial definition of UniScore to be the sum of the number of matched product ions (b- and y-ions) and the number of matched sequence stretch, and we accepted the MaxQuant-generated peak lists with only the top 6 peaks per 50 Th bin for UniScore peak annotation. Using the four search engines including Mascot, MaxQuant, X! Tandem and Comet, we generated UniScores and compared them with the original results from each engine. The obtained results are shown in Figure 1A-D. In all cases for PSM hit numbers at FDR 1%, the UniScore-based approach succeeded in increasing the number of PSMs. Furthermore, by integrating the results of the four search engines using UniScore, we succeeded in maximizing the number of PSM hits at 1% FDR (Figure 1E). The correlation between the scores of each search engine and UniScore is shown in Figure 2. The coefficient of determination ranged from 0.60 (MQ andromeda score) to 0.90 (Comet Xcorr), but in any case a positive correlation was observed. We also checked the correlation between UniScore and the original score for each of the five fractions, but no change in correlation was observed according to physical properties such as peptide length (Supplementary Figure 2).

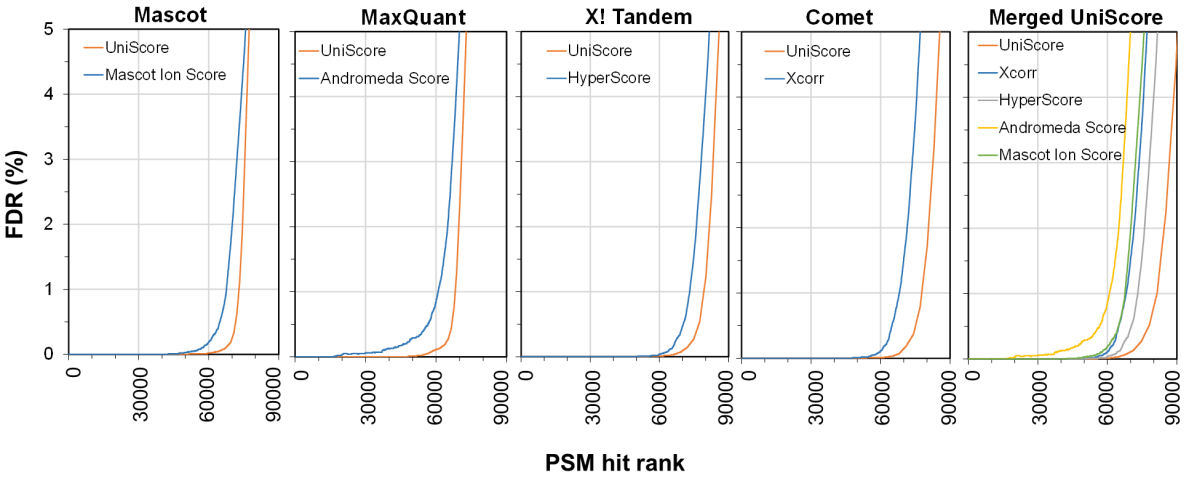


Figure 1: Relationship between PSM hits and FDR by the UniScore-based and the original approaches using four different search engines.

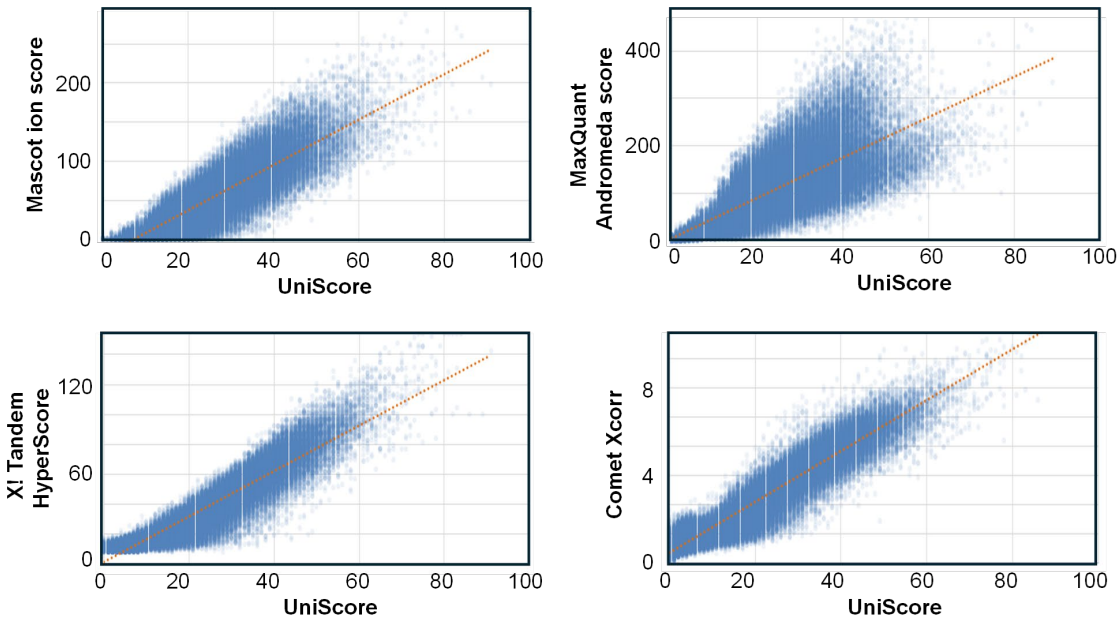


Figure 2: Correlation between UniScore and the scores from four different search engines

Next, we used ProteoWizard with charge conversion function to create a peak list, and used Mascot to investigate the effect of converting product ions to singly-charged ions. As a result, the PSM number increased by only 2.5% with UniScore

and 3.0% with Mascot ion score, so the charge conversion effect was not significant. On the other hand, there was a clear effect of charge conversion on the UniScore value itself, and it was found that the effect of increasing the UniScore was particularly large for peptides with precursor charge numbers greater than 2 (Supplementary Figure 3). When MaxQuant was used instead of ProteoWizard as the peak list generation tool, the increase in PSM was 1.5% and 0.9% for UniScore and Mascot ion score, respectively, and MaxQuant as a peak list generator was slightly better, so MaxQuant was adopted for subsequent experiments. Of the four search engines we investigated, Mascot had the lowest PSM number, and the other three are open source, so we decided to calculate UniScore by merging the results of Comet, X! Tandem, and MaxQuant in subsequent experiments, with a view to future versatility.

We also considered whether to take into account the bin size, the maximum number of peaks in each bin, and the doubly-charged product ions when filtering the signal intensity of product ions. As a result, as expected, it was sufficient to only consider the singly charged ions, and as the bin size increased from 30, 50, to 100 Th, the optimal top N values were 4, 6, and 12, respectively. Among these, bin 100 Th, N=12 gave the largest PSM number (Supplementary Table 2), so we decided to use this condition from here on.

We then optimized the UniScore components and weighting. First, we considered how the number of cleavage sites identified by b- and y-ions should be counted. We considered the case where either b- or y- ions matched, and the case where both b- and y- ions matched, including weighting, and found that the coefficient of [b- or y- ion] and [b- and y- ion] was 0.5:1, that is, the maximum number of PSM was obtained when the total number of b- and y- ions was simply counted (Supplementary Table 3).

In addition, following the PST concept, the additional effect of the number of matched sequence stretches was examined, and the largest PSM was obtained when they were added together at 1:1 compared to each alone. To further emphasize the effect of the matched sequence stretches, we also checked the effect of the square of the number of matched sequence stretches and found a small effect (increase from 79725 to 80498 PSMs), but the increase was less than 1%, so for the sake of score simplicity, we decided to simply add the total number of b- and y-ions and the number of matched sequence stretches to obtain a UniScore. The number of

unmatched b- and y-ions, peptide length and peptide length minus the number of matched sequence stretches were not found to improve the PSM. On the other hand, when all of the potential UniScore components such as b- or y-ions, b- and y-ions, matched sequence stretches, matched sequence stretches squared, number of unmatched b- and y-ions, peptide length minus matched sequence stretches were used as input parameters for the percolator, 81096 PSMs were obtained through machine learning-based scoring. Comparisons with the already reported Morpheus Score and Generic Score were also examined. The results are shown in Table 1: the UniScore performed better than the Morpheus Score, but had fewer PSMs than the Generic Score. However, the use of the Percolator slightly outperformed the number of PSMs from the Generic Score. Surprisingly, the highest PSM was obtained by simply adding the UniScore and Generic Score. As mentioned above, both the Morpheus Score and the Generic Score take into account the fraction of spectral abundance assigned to the matched product ions. Even though the contribution of this term in both scores is not very high, it is a parameter that makes the possibility of chimeric spectra less acceptable. It is well known that multiple peptide ions are present in the isolated window of the precursor ion in the DDA measurements. Therefore, the UniScore + Generic Score, which gives the maximum PSM, was not adopted in this study, but the simple and robust UniScore or UniScore-PCL in combination with Percolator was accepted for the subsequent applications. Compared to the original MaxQuant results, the UniScore approach increased PSM, unique peptides and protein groups by 28.5 %, 14.0 % and 9.5 %, respectively.

Table 1: MaxQuant, UniScore and other scores for PSM at FDR 1%.

	MaxQuant Andromeda score	UniScore	UniScore-PCL	Morpheus score	Generic score	UniScore + Generic score
File 1	7,078	8,235	8,853	7,631	8,841	8,938
File 2	19,742	25,229	25,828	23,591	25,910	26,244
File 3	19,232	25,066	25,209	24,747	25,101	25,243
File 4	13,080	17,132	17,136	17,116	17,113	17,206
File 5	2,897	4,063	4,070	4,067	4,067	4,078
Total PSM	62,029	79,725	81,096	77,152	81,032	81,709

Peptide-Chimeric Spectrum Matching

There is a possibility that the peptide sequences suggested by the three search engines will differ for a single spectrum, and this is due to the simultaneous isolation of different precursor ions, where product ions of different origins coexist in a single MS/MS spectrum (chimera spectrum). The precursor ion fraction in standard DDA is reported to be 0.14 [34], so it is thought that chimeric spectra are obtained with a very high probability. The total number of annotated spectra in this test data set was 68,829, and the number of chimeric spectra was 8,676. The number of PSM per spectrum was 2 for most of the spectra, and the number of spectra with 3 or more PSM was only 0.7% (Figure 3A). We then investigated the distribution of the overlap rate of annotated product ions in chimeric spectra with product ions from other precursors. As a result, more than half of the spectra had no overlap at all, and there were almost no cases where the overlap rate exceeded 50%. This suggests that when the search engines returned the sequence candidates, those with a high overlap rate for the product ions were already removed, and only the best hits per spectrum remained as candidate sequences. These results indicate that the chimera spectrum-derived PSM by UniScore is acceptable without additional criteria.

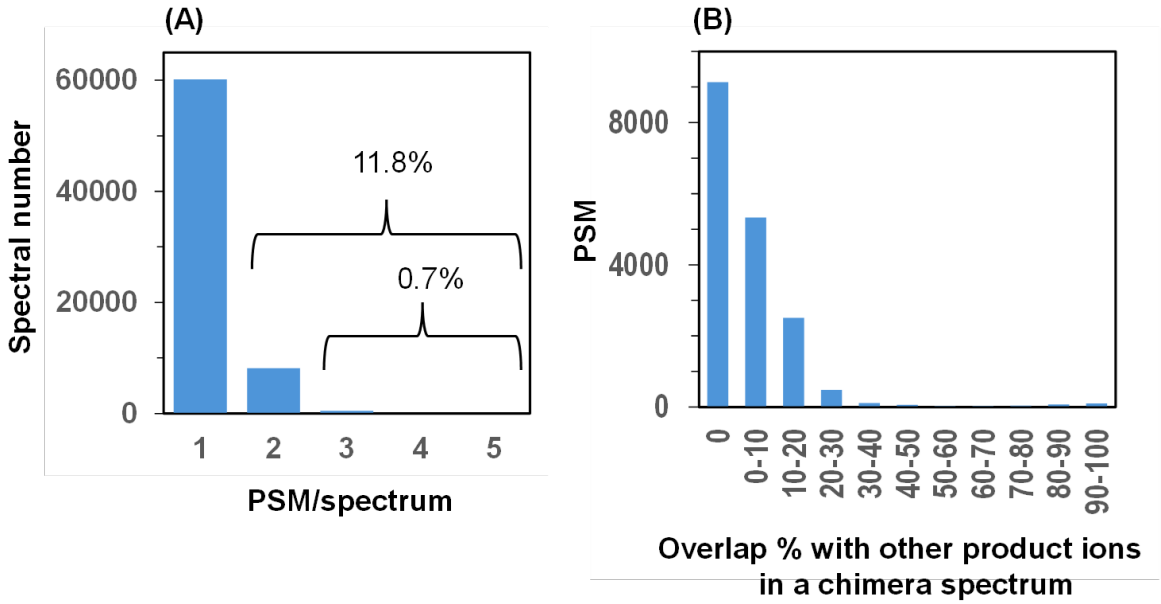


Figure 3: PSM in chimeric spectra. (A) Distribution of PSM number per spectrum, (B) Overlap percentage of product ions with those from other precursor ions in a chimeric spectrum.

Application to human proteome data by LC/MS/MS with extensive pre-fractionation

We selected the HeLa proteome analysis data by Bekker-Jensen et al.[35] as an example of how a combination of high-pH RP fractionation and short gradient can achieve both high efficiency and high depth, and re-analyzed it using UniScore. In the original paper, they analyzed 1 µg of HeLa peptides on the column from each of 46 HpH fractions with 30 min LC-MS/MS gradients (45 min run-to-run), using a total of 34.5 hr resulted in identification of 166,620 unique peptide sequences covering 11,292 protein groups or 10,284 protein-coding genes by MaxQuant. In this study, we re-analyzed the data using MaxQuant, UniScore and UniScore-PCL with the same database and search conditions. Fig. 4A shows a comparison of the PSM values for each of the 46 fractions. In all fractions, the UniScore approach outperformed by about 20% more PSM values than MaxQuant, whereas UniScore-PCL gave slightly better results than UniScore. After merging all the results from the 46 fractions, we calculated the number of identified PSMs, peptides, and protein groups at FDR 1% by UniScore and compared them with MaxQuant (Figure 4B,C,D). As expected, UniScore outperformed MaxQuant, but in this case the difference was not as great as the results from the test dataset used in the development. This is because it has already been shown that increasing the number of human protein identifications by 1000 from 9000 is completely different from increasing it by 1000 from 4000 (Nagaraj et al. 2011). The UniScore-based chimera spectrum content for this dataset was 5.6%, less than half the 11.8% for the test dataset. This is due to the fact that the test dataset was obtained by a single shot, whereas this dataset was heavily fractionated.

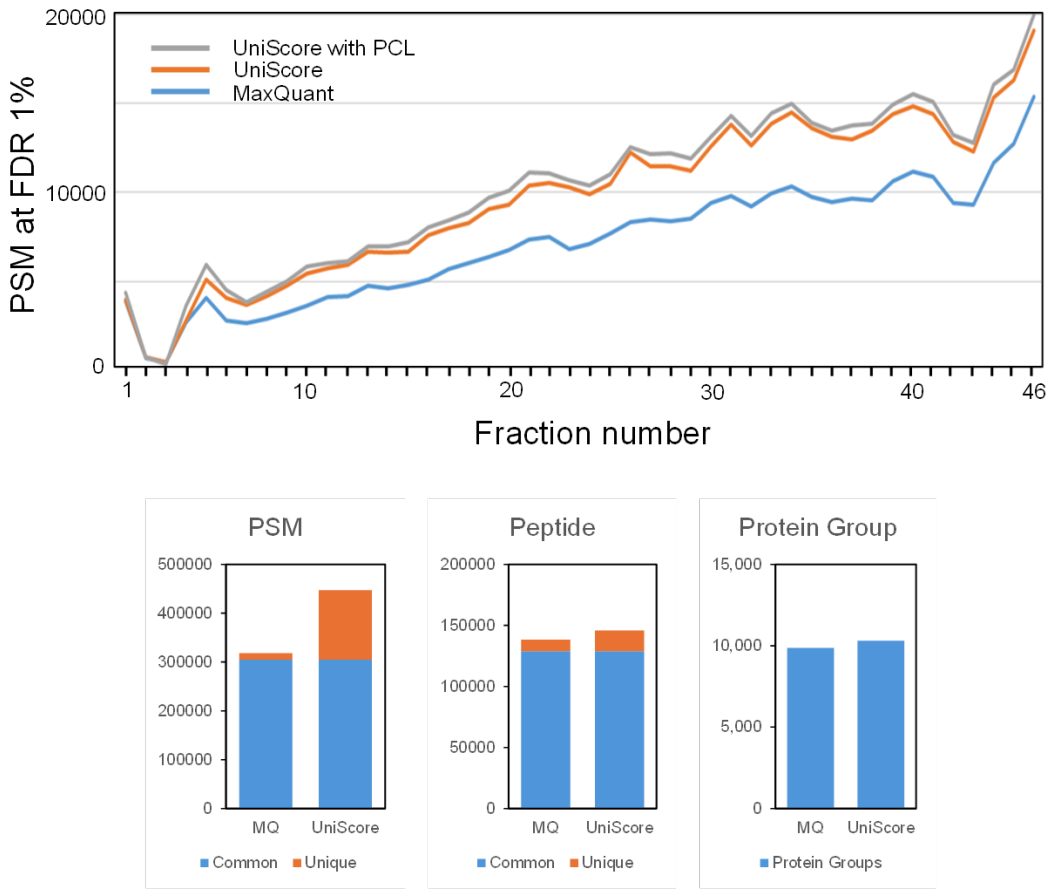


Figure 4: Deep HeLa proteome data analyzed by MaxQuant and UniScore.

Application to human phosphoproteome data by LC/MS/MS following phosphopeptide enrichment

As an application example to post-translational modification proteome data, we examined SARS-COV-2 infected phosphorylation proteome data[36] using UniScore. The samples were Caco-2 cells infected with SARS-CoV-2, and after TMT labeling, phosphopeptide enrichment was performed using Fe-IMAC, and high-pH RP fractionation was performed using StageTip, followed by LC/MS/MS analysis. Proteome Discoverer implemented with Sequest HT was used for data analysis, and the data were controlled at 1% FDR at the peptide level using Percolator with the TD method. Phosphorylation site localization was determined using PhosphoRS implemented within Proteome Discoverer. As we did not have this commercial tool for determining the phosphosite localization, we just compared the sequences and number of phosphate groups without considering the localization of phosphosites.

The results are shown in Figure 5. The UniScore approach significantly improved the number of identified phosphopeptides by 73.2%. When we investigated what kind of phosphorylated peptides were specifically identified by UniScore, we found that longer phosphopeptides were being identified (Figure 5B). One concern with UniScore is that, due to its definition formula, it tends to give higher scores to longer peptides. However, as already shown in Supplementary Figure 2, the bias towards long peptides is almost the same as the Mascot ion score, which is scored based on the probability of random matching. In other words, both UniScore and Mascot ion score tend to be high for long peptides, but it should be noted that this is not a statistical bias because the database search space for precursor ions with higher molecular mass is also narrowed. In any case, UniScore was shown to be effective, particularly for phosphopeptides, which tend to have a long average peptide length.

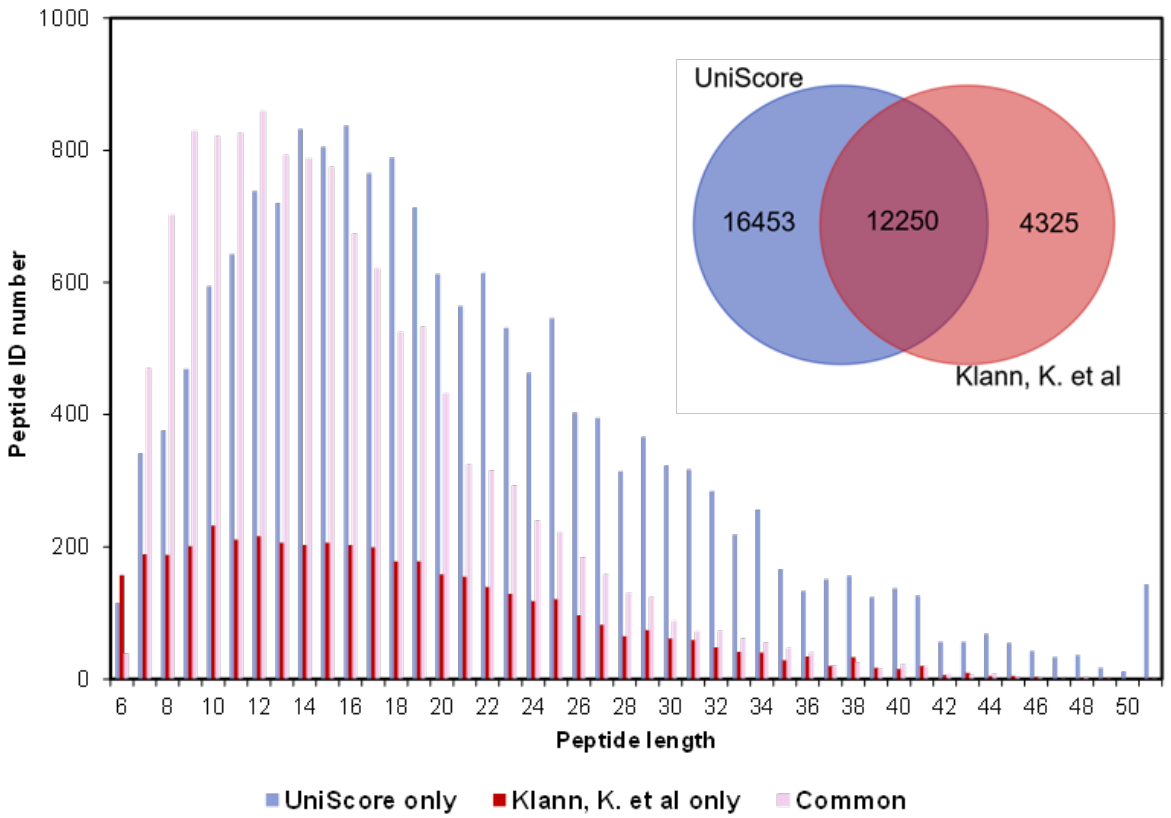


Figure 5: UniScore re-analysis of SARS-CoV-2 infected Caco-2 phosphoproteome data

Conclusions

The UniScore method can be computed with only the product ion information matching the sequence. UniScore can be performed as long as the product ion information matching the sequence is available, making it an easy tool that requires significantly less computational resources than various existing rescoring tools. The combination of multiple search engines to output candidate sequences without bias has not yet been fully optimized. Nevertheless, the large gain as a constant compared to existing search engines indicates the potential of this method. In matching against chimera spectra, the number of PSMs per spectrum is also well-balanced without being extremely large. In principle, matching with the predicted MSMS spectrum using DL should be more powerful for maximizing the identification parameter, but since the MSMS spectrum itself is dependent on experimental conditions, the prediction accuracy cannot be avoided from deteriorating depending on the CID energy and the individuality of the model. The UniScore, on the other hand, is a robust measure that is virtually unaffected by experimental conditions and can be used more widely. Future development is to confirm its robustness (repository scale analysis) by applying it to larger and more diverse data sets (various PTMs, coexistence of Orbitrap and Q-TOF data, etc.). This UniScore concept can also be extended to DIA data in which prioritized selection of product ions would be one of the key steps. Further development is on-going in our laboratory.

DATA AVAILABILITY

The UniScore re-analysis results are available through the jPOST repository (https://jpostdb.org/).

ACKNOWLEDGEMENTS

We thank all members of the jPOST project team. This work was supported by the Life Science Database Integration Project (Database Integration Coordination Program) funded by the Office of the NBDC Program of the Japan Science and Technology Agency [18063028, JPMJND2304].

Conflict of interest statement. None declared.

References

1. Hunt, D. F. *et al.* Characterization of peptides bound to the class I MHC molecule HLA-A2.1 by mass spectrometry. *Science* **255**, 1261–1263 (1992).
2. Eng, J. K., McCormack, A. L. & Yates, J. R. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **5**, 976–989 (1994).
3. Perkins, D. N., Pappin, D. J., Creasy, D. M. & Cottrell, J. S. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20**, 3551–3567 (1999).
4. Elias, J. E. & Gygi, S. P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods* **4**, 207–214 (2007).
5. Craig, R. & Beavis, R. C. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* **20**, 1466–1467 (2004).
6. Geer, L. Y. *et al.* Open mass spectrometry search algorithm. *J. Proteome Res.* **3**, 958–964 (2004).
7. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26**, 1367–1372 (2008).
8. Cox, J. *et al.* Andromeda: a peptide search engine integrated into the MaxQuant environment. *J. Proteome Res.* **10**, 1794–1805 (2011).
9. Bern, M., Kil, Y. J. & Becker, C. Byonic: advanced peptide and protein identification software. *Curr. Protoc. Bioinformatics* **Chapter 13**, 13.20.1–13.20.14 (2012).
10. Kim, S. & Pevzner, P. A. MS-GF+ makes progress towards a universal database search tool for proteomics. *Nat. Commun.* **5**, 5277 (2014).
11. Wilhelm, M. *et al.* Mass-spectrometry-based draft of the human proteome. *Nature* **509**, 582–587 (2014).
12. Kim, M.-S. *et al.* A draft map of the human proteome. *Nature* **509**, 575–581 (2014).
13. Ezkurdia, I., Vázquez, J., Valencia, A. & Tress, M. Analyzing the first drafts of the human proteome. *J. Proteome Res.* **13**, 3854–3855 (2014).
14. Savitski, M. M., Wilhelm, M., Hahne, H., Kuster, B. & Bantscheff, M. A Scalable Approach for Protein False Discovery Rate Estimation in Large Proteomic Data Sets. *Mol. Cell. Proteomics* **14**, 2394–2404 (2015).
15. Shteynberg, D., Nesvizhskii, A. I., Moritz, R. L. & Deutsch, E. W. Combining results of multiple search engines in proteomics. *Mol. Cell. Proteomics* **12**, 2383–2393 (2013).
16. Vaudel, M. *et al.* PeptideShaker enables reanalysis of MS-derived proteomics data sets. *Nat. Biotechnol.* **33**, 22–24 (2015).
17. Guruceaga, E. *et al.* Enhanced Missing Proteins Detection in NCI60 Cell Lines Using an Integrative Search Engine Approach. *J. Proteome Res.* **16**, 4374–4390 (2017).
18. Deutsch, E. W. *et al.* Tiered Human Integrated Sequence Search Databases for Shotgun Proteomics. *J. Proteome Res.* **15**, 4091–4100 (2016).
19. Chen, Y. E. *et al.* APIR: Aggregating universal proteomics database search algorithms for peptide identification with FDR control. *Genomics Proteomics Bioinformatics* **22**, (2024).
20. Mann, M. & Wilm, M. Error-tolerant identification of peptides in sequence databases by peptide sequence tags. *Anal. Chem.* **66**, 4390–4399 (1994).
21. Frank, A., Tanner, S., Bafna, V. & Pevzner, P. Peptide sequence tags for fast database search in mass-spectrometry. *J. Proteome Res.* **4**, 1287–1295 (2005).
22. Wenger, C. D. & Coon, J. J. A proteomics search algorithm specifically designed for high-resolution tandem mass spectra. *J. Proteome Res.* **12**, 1377–1386 (2013).
23. Strauss, M. T. *et al.* AlphaPept: a modern and open framework for MS-based proteomics. *Nat. Commun.* **15**, 2168 (2024).
24. Deutsch, E. W. *et al.* Trans-Proteomic Pipeline: Robust Mass Spectrometry-Based Proteomics Data Analysis Suite. *J. Proteome Res.* **22**, 615–624 (2023).

25. Kalhor, M., Lapin, J., Picciani, M. & Wilhelm, M. Rescoring Peptide Spectrum Matches: Boosting Proteomics Performance by Integrating Peptide Property Predictors Into Peptide Identification. *Mol. Cell. Proteomics* **23**, (2024).
26. Deutsch, E. W. *et al.* Universal Spectrum Identifier for mass spectra. *Nat. Methods* **18**, 768–770 (2021).
27. Wen, B. & Zhang, B. PepQuery2 democratizes public MS proteomics data for rapid peptide searching. *Nat. Commun.* **14**, 2213 (2023).
28. Yang, K. L. *et al.* MSBooster: improving peptide identification rates using deep learning-based features. *Nat. Commun.* **14**, 4539 (2023).
29. Picciani, M. *et al.* Oktoberfest: Open-source spectral library generation and rescoring pipeline based on Prosit. *Proteomics* **24**, e2300112 (2024).
30. Buur, L. M. *et al.* MSRescore 3.0 Is a Modular, Flexible, and User-Friendly Platform to Boost Peptide Identifications, as Showcased with MS Amanda 3.0. *J. Proteome Res.* **23**, 3200–3207 (2024).
31. Chambers, M. C. *et al.* A cross-platform toolkit for mass spectrometry and proteomics. *Nat. Biotechnol.* **30**, 918–920 (2012).
32. Tabata, T., Yoshizawa, A. C. & Ishihama, Y. jPOST Tools (I): Utilities for Peak List Processing. *JPDM* **2**, 4 (2020).
33. Nesvizhskii, A. I. & Aebersold, R. Interpretation of shotgun proteomic data: the protein inference problem. *Mol. Cell. Proteomics* **4**, 1419–1440 (2005).
34. Michalski, A., Cox, J. & Mann, M. More than 100,000 detectable peptide species elute in single shotgun proteomics runs but the majority is inaccessible to data-dependent LC-MS/MS. *J. Proteome Res.* **10**, 1785–1793 (2011).
35. Bekker-Jensen, D. B. *et al.* An Optimized Shotgun Strategy for the Rapid Generation of Comprehensive Human Proteomes. *Cell Syst* **4**, 587–599.e4 (2017).
36. Klann, K. *et al.* Growth Factor Receptor Signaling Inhibition Prevents SARS-CoV-2 Replication. *Mol. Cell* **80**, 164–174.e4 (2020).

Supplementary Table 1

Data analysis parameters for HeLa global proteome dataset (PXD005159/JPST000200)

Search engines and other tools with version	
Mascot	2.7.0
MaxQuant:	1.6.17.0（Download date: 2021/01/26）
Comet	2019.01 rev. 5
X!Tandem	2015.04.01.1
Percolator	v3-04
ProteoWizard	3.0.21021.e2078b75f 64-bit（Download date:2021/01/25)
Search parameters	
Database	UniProt human (total 42,362 seq x 2) consisting of sprot_human_20200318.fasta isoforms from UniProt release 2020_01 MaxQuant contaminants.fasta decoy based on reversed seq, concatenated
Input peak list	mgf files generated by MaxQuant
Fixed modifications	Carbamidomethyl (Cys)
Variable modifications	Oxidation (Met)
MS1 tolerance	Mascot, X! Tandem, Comet: 10 ppm MaxQuant 1st search: 20 ppm, 2nd search: 4.5 ppm
MS2 tolerance	Mascot, MaxQuant, X! Tandem: 20 ppm Comet: high resolution
Enzyme	Trypsin
Missed cleavage	1
Peak matching for UniScore	Top 12 peaks within every 100 Th bin

Supplementary Table 2: Effect of product ion abundance filtering on UniScore-PSM at FDR 1%

Bin size (Th)	30			50							100								
Top N/bin	3	4	5	4	5	6	6	7	8	10	6	8	10	11	12	12	13	14	15
Product ion charge	1	1	1	1	1	1	1,2	1	1	1	1	1	1	1	1	1,2	1	1	1
File 1	7,901	8,253	7,972	8,199	8,084	8,266	7,909	7,911	7,995	8,080	7,654	7,765	8,114	8,228	8,296	7,961	8,347	8,383	8,404
File 2	23,257	23,769	23,445	22,995	23,530	23,767	23,823	23,898	23,465	23,576	22,048	23,189	23,669	23,768	23,840	23,353	23,350	23,411	23,446
File 3	21,898	21,862	21,887	21,853	21,926	21,943	21,874	21,879	21,889	21,897	21,739	21,882	21,926	21,935	21,947	21,875	21,873	21,879	21,885
File 4	14,400	14,404	14,408	14,390	14,399	14,405	14,410	14,407	14,408	14,409	14,383	14,397	14,401	14,405	14,405	14,409	14,406	14,406	14,408
File 5	3,290	3,290	3,290	3,290	3,290	3,291	3,292	3,291	3,291	3,291	3,288	3,288	3,289	3,290	3,291	3,292	3,291	3,291	3,291
Total	70,746	71,578	71,002	70,727	71,229	71,672	71,308	71,386	71,048	71,253	69,112	70,521	71,399	71,626	71,779	70,890	71,267	71,370	71,434

Search parameters as listed in Supple Table 1 except the parameters in top N in bin X Th.

Supplementary Table 3: Effect of matched b-, y-ions on UniScore-PSM at FDR 1%

Coefficient for (b and y) sites x 2	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
Coefficient for (b or y) sites	1	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1	0
File 1	1,270	1,616	3,078	5,241	7,442	7,544	6,067	4,439	3,650	3,528	3,143
File 2	8,359	11,315	15,403	19,413	23,195	23,426	20,988	18,296	15,158	14,603	10,723
File 3	14,576	17,022	20,985	22,758	24,254	24,687	23,752	21,676	18,229	16,598	14,582
File 4	14,629	16,280	16,848	17,137	17,161	17,143	16,958	16,301	14,927	13,253	11,023
File 5	3,931	4,021	4,053	4,073	4,069	4,061	4,020	3,824	3,200	2,333	1,890
Total	42,765	50,254	60,367	68,622	76,121	76,861	71,785	64,536	55,164	50,315	41,361

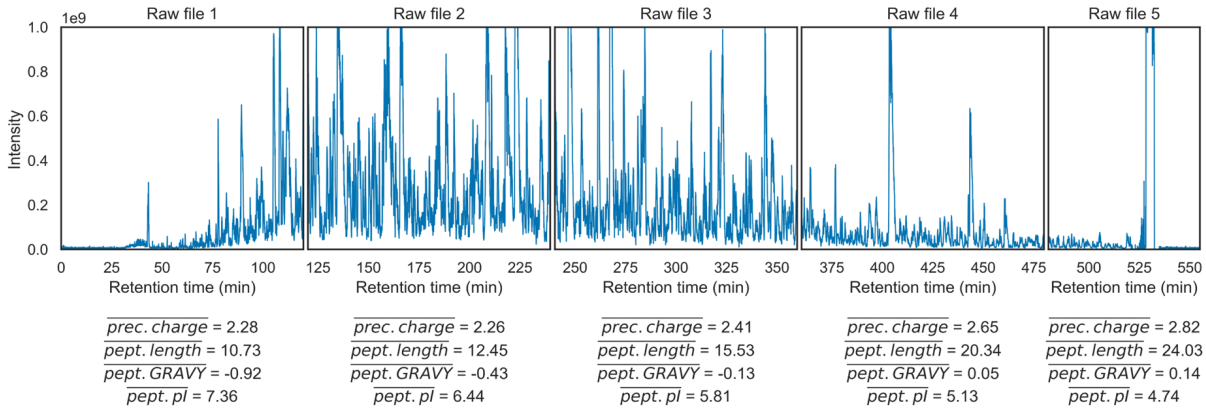
Supplementary Table 4: Effect of matched sequence stretches on UniScore-PSM at FDR 1%

Coefficient for (b, y-ions)	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
Coefficient for (matched sequence stretches)	1	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1	0
File 1	7,449	7,706	7,710	7,959	8,184	8,235	8,211	8,124	8,134	8,134	7,544
File 2	23,659	25,256	25,246	25,147	25,213	25,229	25,015	24,744	24,473	24,464	23,426
File 3	24,533	24,954	24,965	24,991	25,042	25,066	25,035	25,005	24,932	24,907	24,687
File 4	16,865	16,991	17,017	17,058	17,116	17,132	17,156	17,167	17,170	17,170	17,143
File 5	3,945	4,010	4,015	4,022	4,048	4,063	4,070	4,072	4,068	4,068	4,061
Total	76,451	78,917	78,953	79,177	79,603	79,725	79,487	79,112	78,777	78,743	76,861

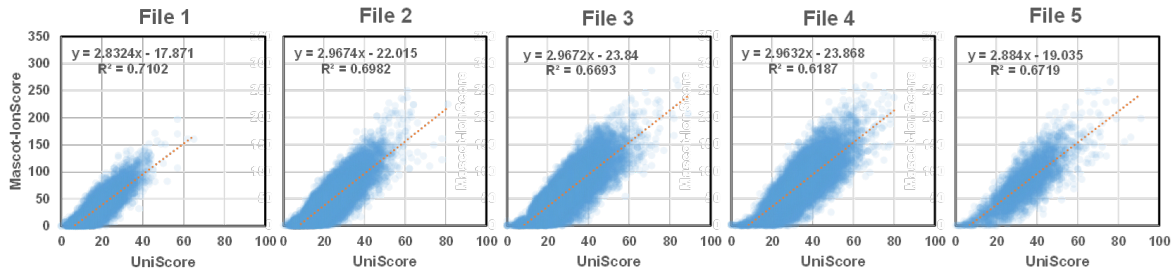
Supplementary Table 5: Effect of matched sequence stretches squared on UniScore-PSM at 1% FDR.

(b,y-ions) + (matched sequence stretches)	1	1	1	1	1	1	1	1	1	1
Coefficient for (matched sequence stretches)^2	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.1
File 1	8,441	8,441	8,444	8,354	8,248	8,246	8,229	8,228	8,209	8,124
File 2	25,759	25,759	25,759	25,759	25,757	25,742	25,711	25,710	25,702	25,631
File 3	25,106	25,106	25,106	25,106	25,106	25,109	25,109	25,107	25,107	25,114
File 4	17,136	17,136	17,136	17,136	17,136	17,136	17,136	17,130	17,128	17,118
File 5	4,053	4,053	4,053	4,053	4,045	4,043	4,043	4,043	4,043	4,049
Total	80,495	80,495	80,498	80,408	80,292	80,276	80,228	80,218	80,189	80,036

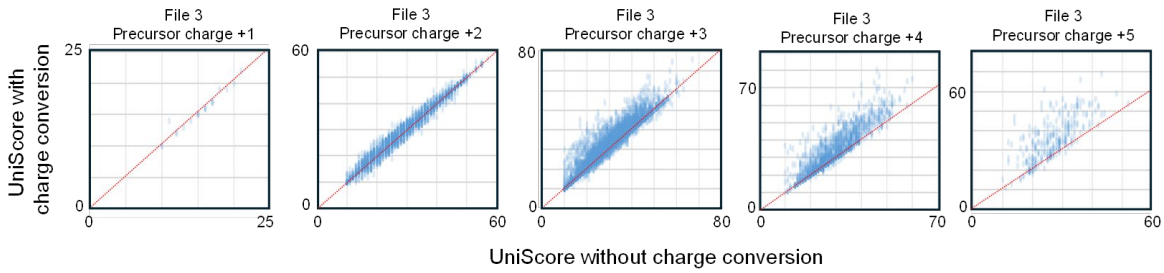
Supplementary Figures



Supplementary Figure 1: Basepeak chromatogram of a HeLa global proteome sample. Data was obtained from PXD005159/JPST000200.



Supplementary Figure 2: Correlation between UniScore and Mascot ion score for peptides with different physical properties in File 1-5.



Supplementary Figure 3: Effect of charge conversion on UniScore of peptides with different precursor charges. The red dot line shows y=x.