Supplementary Note 1 – Assessing spectrum clustering accuracy

209 human datasets (test dataset, Supplementary Table 1) from PRIDE Archive were reprocessed using SpectraST at a 1% peptide FDR (Online Methods). This dataset was used as the basis to analyse clustering accuracy and to compare *spectra-cluster*'s performance with two alternative clustering algorithms, MSCluster[2] and MaRaCluster[3].

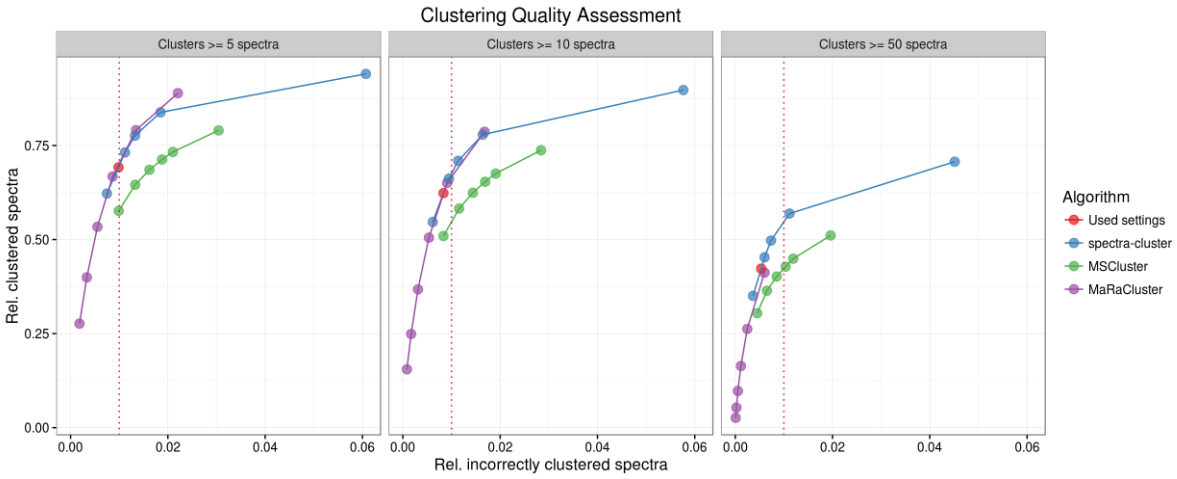

Figure 1.1: The spectra-cluster algorithm leads to less cluster fragmentation as compared to the MaRaCluster algorithm. Clustering sensitivity (y-axis) was assessed based on the number of clustered spectra (shown as relative to the total number of spectra in the test dataset). Clustering specificity (x-axis) was assessed based on the proportion of spectra that were not identified as the most common peptide in a cluster.

As shown in Figure 1.1 both *spectra-cluster* and MaRaCluster outperformed the MSCluster algorithm when clusters larger than 5 and 10 spectra were considered. However, when only taking larger clusters into consideration (containing at least 50 spectra) the *spectra-cluster* algorithm outperformed MaRaCluster. This is in-line with the results presented recently by The and Käll, where MaRaCluster outperformed the MSCluster algorithm but generated smaller clusters[3]. This is a known trade-off in clustering approaches where higher cluster purity leads to smaller clusters and thereby to cluster fragmentation (spectra that should be clustered together are included in different clusters). The results shown in this analysis however show a worse performance of MaRaCluster than shown in the original publication[3]. This is mainly due to the design principles of MaRaCluster: the algorithm was explicitly developed for homogenous datasets. Therefore, a worse performance is to be expected in highly heterogeneous data.

Taking the size of the test dataset into consideration, the chosen peptide FDR of 1% may seem too high. Nevertheless, we argue that any incorrect identification in the reference data will only decrease the measured clustering accuracy. As a proof-of-concept we re-analysed the clustering results of the *spectra-cluster* algorithm using the "Used settings" (as in Figure 1.1) taking only PSMs identified at

0.01% FDR into consideration. As expected, the proportion of clustered spectra remained unchanged while the proportion of incorrectly clustered spectra decreased from 1% (Figure 1.1) to 0.5%.

We also want to highlight that, in terms of scalability, only the MSCluster algorithm and the *spectra-cluster* algorithm were able to process a large repository-scale dataset. In its current implementation MaRaCluster compares all combinations of spectra within the defined precursor tolerance. While generating the results using our test dataset, MaRaCluster generated roughly 172 GB of intermediate data (input size of 22 GB, as uncompressed MGF files). If the amount of data would increase linearly (which is unlikely based on the MaRaCluster algorithm) MaRaCluster would generate roughly 3 TB of intermediate data for analysing the same dataset used in this manuscript. However, the fact that the relationship between the number of analysed spectra and the intermediate data generated is most likely exponential, this fact alone would currently prevent MaRaCluster's use for a repository-sized dataset.

1.1 Influence of Chimeric spectra in the clustering process

To assess the influence of chimeric spectra on the *spectra-cluster*'s accuracy 30% *in-silico* generated chimeric spectra were added to the test dataset. Chimeric spectra were generated per dataset by randomly merging spectra with precursor masses within 2 *m/z* units at different random mixture coefficients of 0.1, 0.2, 0.5, and 1 (both spectra represented at the same abundance). Spectra were generated until the dataset contained 30% of chimeric spectra.

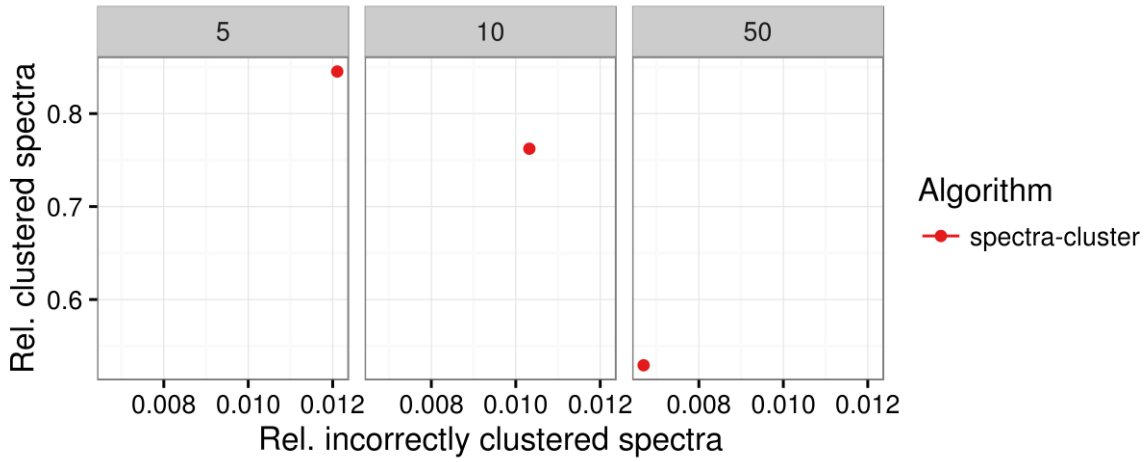

Figure 1.2: Accuracy of the *spectra-cluster* algorithm is not influenced by chimeric spectra. The plots shown only take clusters with at least 5, 10, and 50 spectra, respectively into consideration.

As shown in Figure 1.2 the addition of chimeric spectra did not influence clustering accuracy. The slightly higher proportion of incorrectly clustered spectra (1.2 % instead of 1% without chimeric spectra) is due to the fact that only the most abundant peptide was taken into consideration. Chimeric spectra with a mixture coefficient of 1 (both spectra represented at the same abundance)

that were clustered with spectra from the second peptide were therefore counted as incorrect matches.

1.2 Influence of the probabilistic similarity function on clustering accuracy

The *spectra-cluster* Java API is completely modular. Therefore, it is possible to replace, for example, only the similarity function and assess the influence of the different components of the algorithm separately.

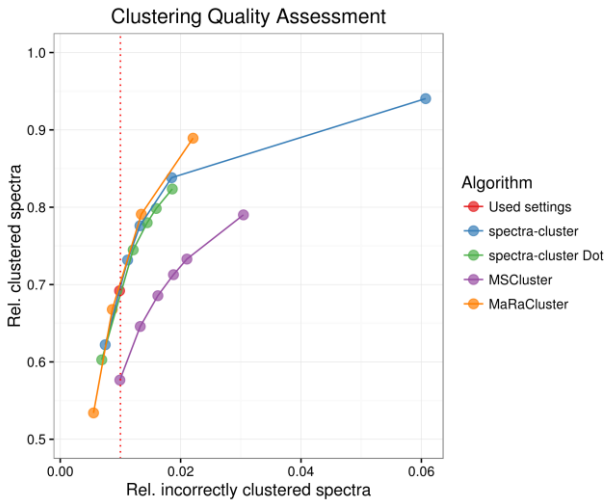

Figure 1.3: Replacing the probabilistic similarity function with the normalised dot-product (spectra-cluster Dot) clustering accuracy was reduced. Nevertheless, this alone does not explain the increased accuracy observed compared to the MSCluster algorithm.

As shown in Figure 1.3 the performance of the *spectra-cluster* algorithm using the normalised dot-product as similarity function performed worse than the probabilistic similarity function. Nevertheless, this effect only plays a small part in the total improvement in clustering accuracy compared to the MSCluster algorithm. Since most features are implemented in a highly similar way as in the MSCluster algorithm (Supplementary Note 7) we are currently unable to identify the primary reason for the observed increase in clustering accuracy in *spectra-cluster*.

1.3 Clustering accuracy in the PRIDE Cluster dataset

As additional validation we analysed the frequencies of the most common and second most common peptide per cluster in the complete PRIDE Cluster dataset. These analyses are based on the originally submitted identification data. The shown estimates therefore represent a worst case scenario since most submissions to PRIDE Archive contain the complete set of identifications before any FDR related filtering was performed.

The average fraction of the most common peptide per cluster is an estimate of cluster purity. Even based on the originally submitted identification data, cluster purity does not decrease with increasing

cluster size as compared to MSCluster[3] (Figure 1.4). Additionally, we observed a marked increase in purity starting from clusters with at least 3 spectra. The slight decrease in cluster purity in very large clusters (250+ spectra) is caused by very large clusters where only 2-5% of the spectra were identified as the most common peptide. These clusters were reanalysed as part of the analysis of incorrect clusters (see the main manuscript). In all of these cases, we were unable to derive reliable identifications for the vast majority of the originally submitted spectra. Therefore, these spectra may represent consistently observed peptides that are prone to (incorrect) identifications but are either not present in the commonly used sequence databases or contain unexpected PTMs.

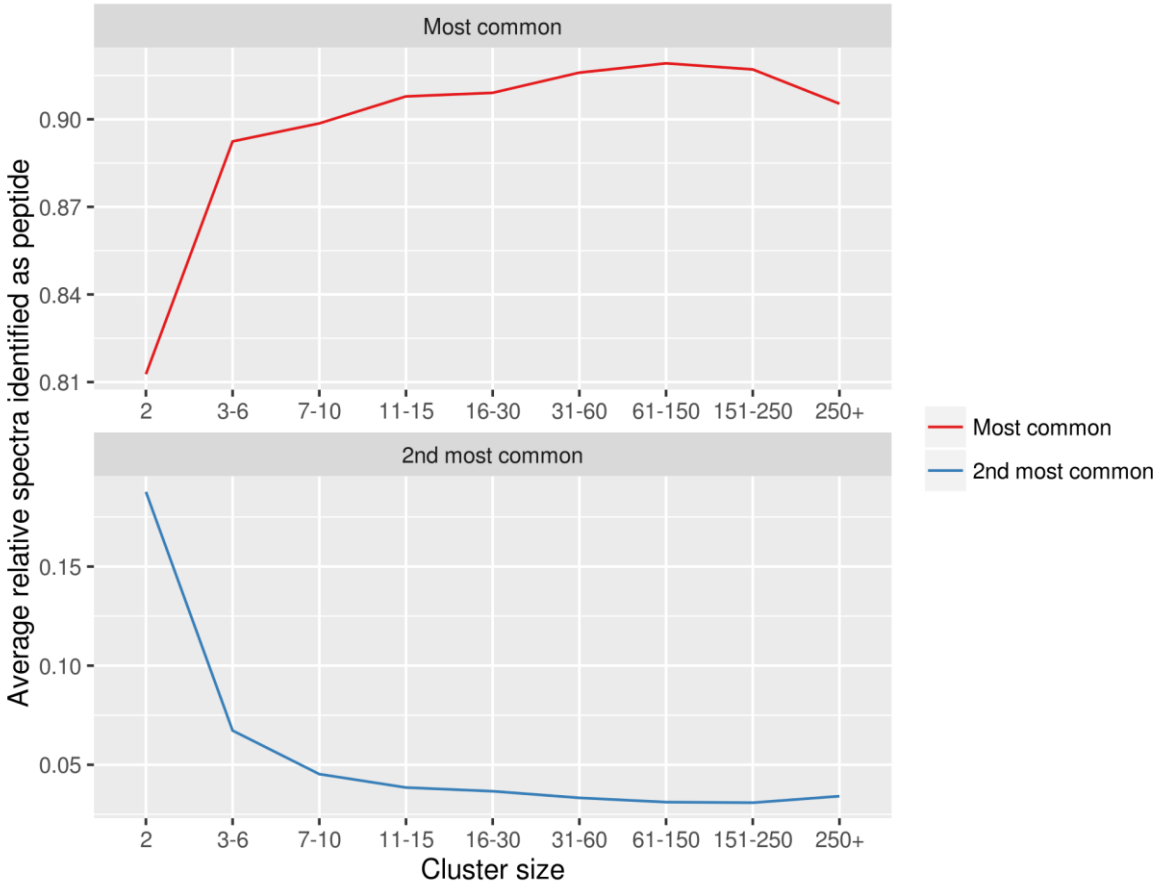

Figure 1.4: The average relative number of spectra identified as the most common (upper panel, red line) and the second most common (lower panel, blue line) show that cluster purity remained stable across the complete PRIDE Cluster dataset. The estimates are based on the originally submitted identification data and therefore represent a worst case scenario.

The analysis of the average proportion of spectra identified as the second most peptide should answer the question of whether the clustering algorithm has difficulties to separate certain peptide species or not. This would be observed in the data as a high proportion of clusters where the second most common peptide occurs nearly as frequently as the most common one (Figure 1.5). Similar to the observed increase of cluster purity in larger clusters, the proportion of spectra identified as the

second most common peptide decreased with an increasing cluster size. This is additional evidence that the quality of data in PRIDE Cluster increases in parallel to the size of the dataset.

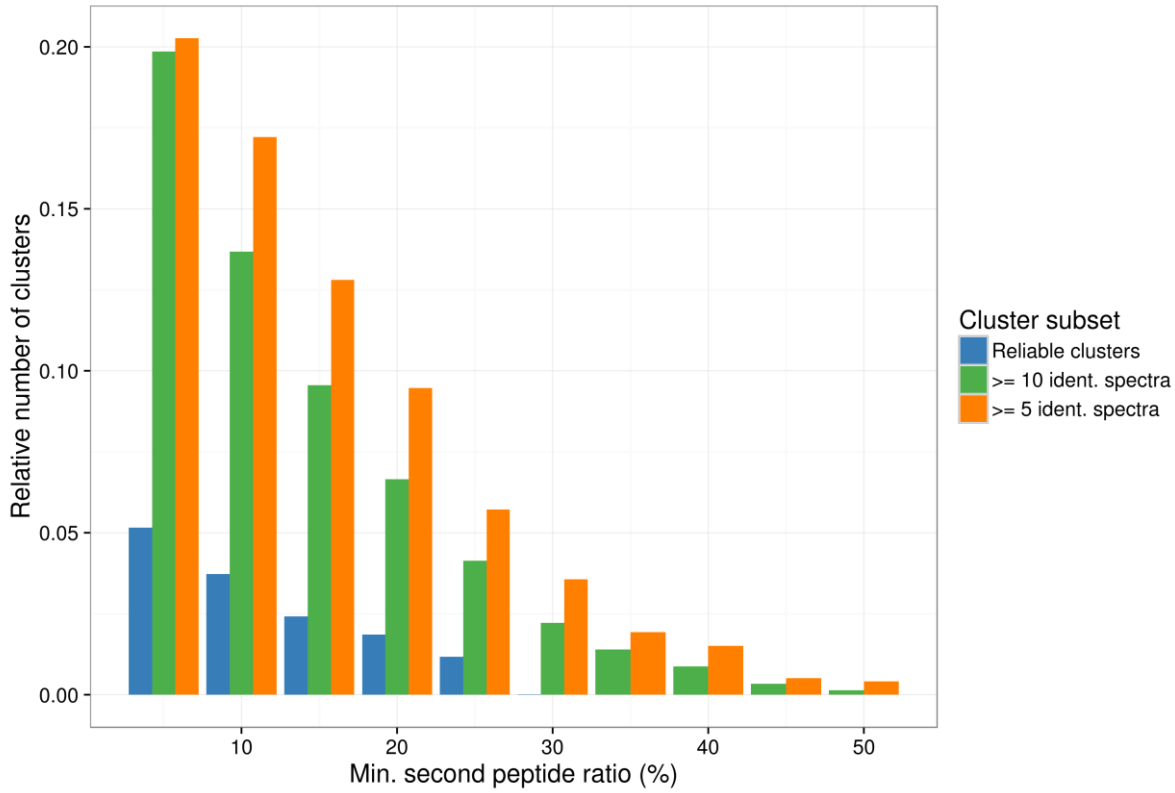

Figure 1.5: The fraction of spectra identified as the second most common peptide was very low across different subsets of clusters (reliable clusters, clusters with >= 5 and >= 10 identified spectra shown). The x-axis represents the N% of spectra identified as the second most common peptide in the cluster.

As shown in Figure 1.5, in only a very small fraction of clusters, a larger proportion of spectra were identified as the second most common peptide. In less than 3% of clusters with at least 5 identified spectra more than 30% of the spectra were identified as the second most common peptide. More than a third of these clusters contained five to ten identified spectra (seen in the difference between the two groups). The larger clusters become, the less frequent this phenomenon occurs.

In reliable clusters more than 15% of the spectra are identified as the second most common peptide only in less than 2.5% of the clusters. In roughly 1% of the reliable clusters a second common peptide (more than 20% of the spectra) is observed. We believe that these cases demonstrate the current limit of the algorithm, where two peptide species cannot be reliably separated. Therefore, we deliberately do not consider these other identifications as incorrect but only define identifications of the most common peptide as reliable.

Supplementary Note 2 – Defining reliable peptide identifications based on the clustering results

The original PRIDE Cluster algorithm was primarily developed to identify reliable identifications in the data submitted to PRIDE Archive. To identify the required parameters that define these reliable identifications we originally performed a machine learning approach on three test datasets[4]. We now repeated this machine learning approach based on the data generated during the analysis of the 209 human test datasets (Supplementary Table 1). Since these represent a representative sample of all data in PRIDE Archive, we are confident that these parameters are suited to define reliable identifications in all of PRIDE Archive.

Every sequence was represented using a vector of properties: the sequence's ratio (the proportion of spectra identified as this sequence within a given cluster), its rank among the sequences identified in the cluster, the total size of the cluster, the number of projects that identified the sequence, the precursor *m/z* range of the cluster and the number of assays in which the sequence was identified. All consensus spectra were identified using SpectraST against the NIST human spectral library at a 1% peptide FDR (see Online Methods). Consensus spectra were deemed reliable if they were identified as the same sequence and the most common sequence identified within the cluster. The machine learning analysis was performed using the Waikato Environment for Knowledge Analysis (WEKA) version 3.6.12. The dataset was randomly split 2:1 in two parts where the larger part was used as training set and the second part as validation dataset. To learn the rules to classify reliable and unreliable sequences the "Conjunctive Rule Learner" was used with the following parameters: 3 folds, minimum total weight 2.0, number of antecedents -1, seed 1. This approach was chosen since it is one of the simplest machine learning algorithms that results in simple, human understandable results. Additionally, we used several tree-learning approaches to test whether these would result in better classification accuracy by choosing different parameters.

All used approaches only selected the sequence's ratio as the most important parameter. As expected, the various tree-based learning approaches led to marginally better classification results. The Conjunctive rule learner, as well as the tree-based learning approaches both selected the sequence's ratio of 0.68 as primary splitting point. The cluster size was never used for the classification by any of the algorithms. The fact that this ratio is comparable to the one found in our original paper (0.62 using a more than ten-fold smaller test dataset)[4] gives us confidence that the employed threshold of 0.7 is suited to identify reliable identifications.

In addition, the fact that the cluster's size was not chosen as a classification parameter, even when the size of the test dataset was more than 10-fold larger, made us change the minimum cluster size

from originally 10 spectra to now 3 spectra. Therefore, reliable identifications are defined as clusters with at least 3 spectra where at least 70% of the spectra were identified as the same peptide (*e.g.* in a cluster with 3 spectra all have to be identified as the same peptide; in a cluster with four spectra one spectrum may be differently identified).

Supplementary Note 3 – The PRIDE Cluster Resource

3.1 Architecture Overview

The PRIDE Cluster architecture consists of two main components (Figure 3.1): (i) the **Clustering Pipeline** enables to extract and cluster spectra in scale using the *spectra-cluster* algorithm. (ii) **Web and Web Services** are responsible for serving all the external requests for viewing and download the clustering results.

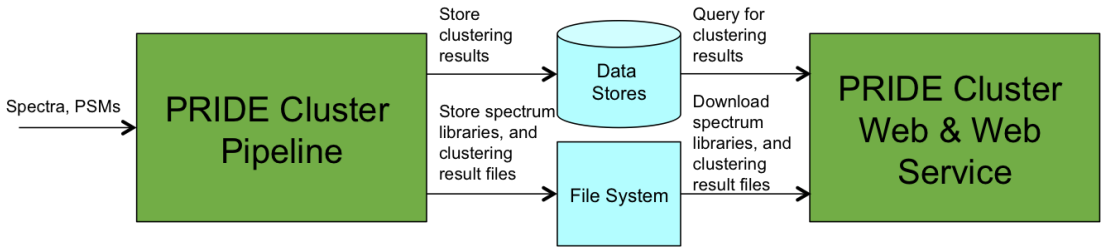

Figure 3.1: The PRIDE Cluster architecture consists of two components: The PRIDE Cluster Pipeline is responsible to create the actual clustering results while the PRIDE Cluster Web & Web Services presents these results to the user.

3.2 Clustering Pipeline

The clustering pipeline consists of three main tasks (Figure 3.2.2): (i) Extract, filter spectra and their peptide identifications from public experiments in PRIDE Archive; (ii) Cluster the extracted spectra using Hadoop "Map Reduce" and the *spectra-cluster* algorithm; (iii) Extract, transform, and load the clustering results into the data store and/or generate spectrum libraries.

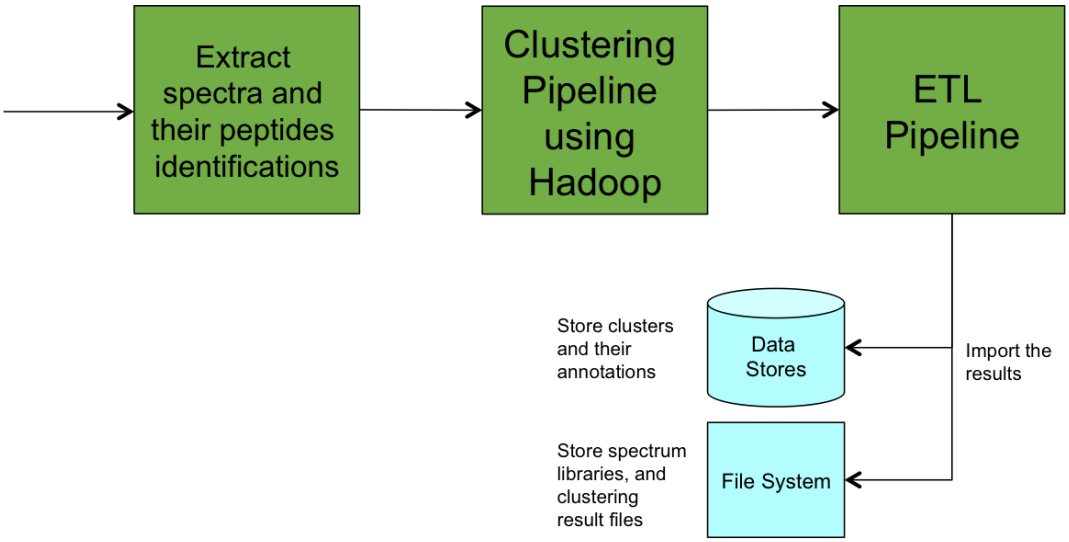

Figure 3.2: The PRIDE Cluster Pipeline is performing the actual clustering of the public data in PRIDE Archive.

The implementation of the clustering pipeline is done using Hadoop "Map Reduce" (http://hadoop.apache.org/docs/r1.2.1/mapred_tutorial.html).

3.3 Web interface & Web Service

The mission of the Web and Web Service is to provide external access to search the clustering results: (i) Search using a peptide sequence; (ii) Filter the results using species and/or PTMs; (iii) Download spectral library files or the clustering result files for further analyses.

The architecture of the Web and Web Service follows the Layered Architecture Pattern (https://en.wikipedia.org/wiki/Multilayered_architecture), where each layer serves only one main responsibility and can only access the layer below (Figure 3.3). An Oracle database (http://www.oracle.com/) is used as the relational store for the clustering results, and one Solr store (http://lucene.apache.org/solr/) is used for storing and enabling the searches. The Spring Framework (http://projects.spring.io/spring-framework/) is the main framework used for implementing both the Web and the Web Service.

The complete source code from both the Web and the Web Service is available on GitHub (https://github.com/PRIDE-Cluster).

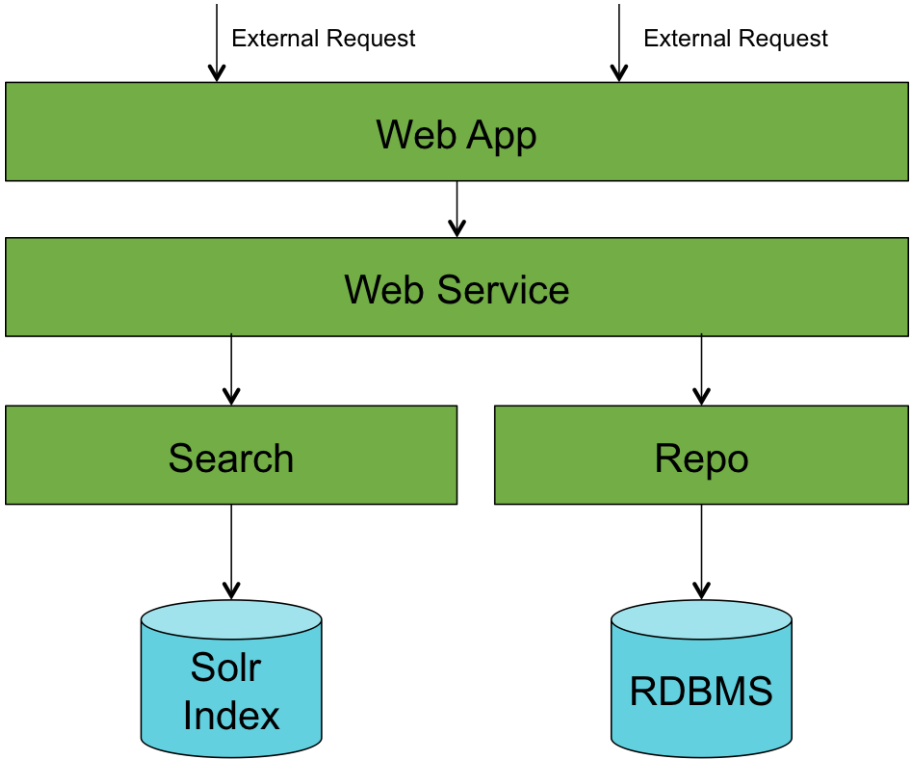

Figure 3.3: PRIDE Cluster Web architecture.

Supplementary Note 4 – Analysing Clustering Results Through the PRIDE Cluster Web Interface

The PRIDE Cluster web interface is available at http://www.ebi.ac.uk/pride/cluster and provides access to the raw clustering results, the pre-compiled sets of interesting clusters (*e.g.* large unidentified clusters, reliable phosphopeptide clusters), the spectral libraries and through the "Peptide Search" functionality, a visualization of the clustering results and the links to the originally submitted data to PRIDE Archive.

The search box on the upper right corner allows the user to search for peptide sequences, filtering by post-translational modifications, as well as species. The peptide search page then displays the results of this search (Figure 4.1). Each cluster is represented as one row in the table with multiple filtering options of the results on the left.

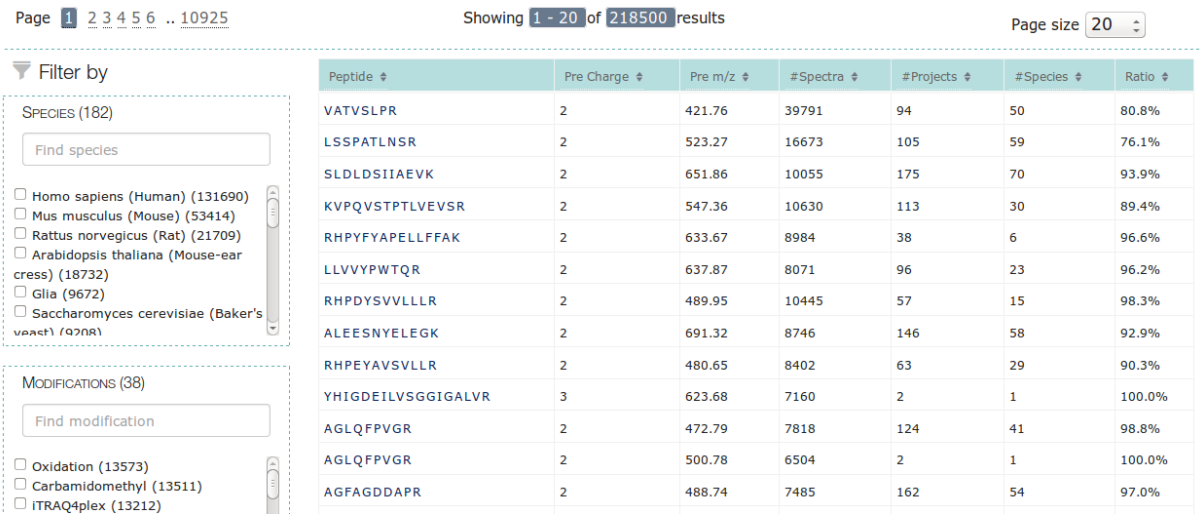

Figure 4.1: The PRIDE Cluster Peptide Search provides a visualisation of the clustering results.

If the user clicks on one of the entries, the cluster's details are displayed (Figure 4.2). As shown in this example, most clusters contain one dominant peptide sequence (in this case 92% of the cluster's spectra were identified as the most common sequence). The other spectra are most commonly identified as other random peptides, where each peptide is often supported by a single spectrum. We believe that these differently identified spectra represent random incorrect identifications.

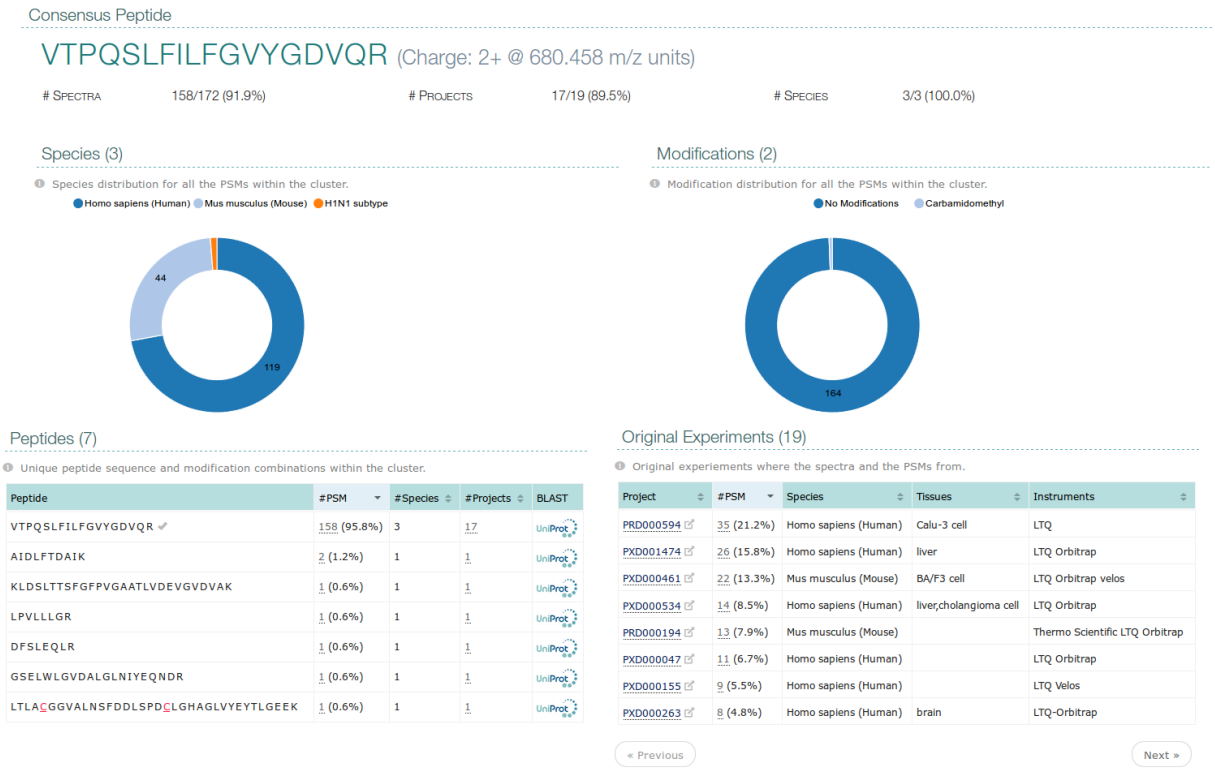

Figure 4.2: The cluster detail view displays a summary of all sequences that are present in the respective cluster.

The user can view all PSMs that correspond to a given peptide (as displayed in the “Peptides” table) or to a given dataset (as displayed in the “Original Experiments” table) through the respective links. These lists then contain additional links to the originally submitted data and the complete submitted identification details. Finally, the cluster’s consensus spectrum (Figure 4.3) and the distribution of precursor m/z deltas (Figure 4.4) are displayed as additional means to assess a cluster’s reliability.

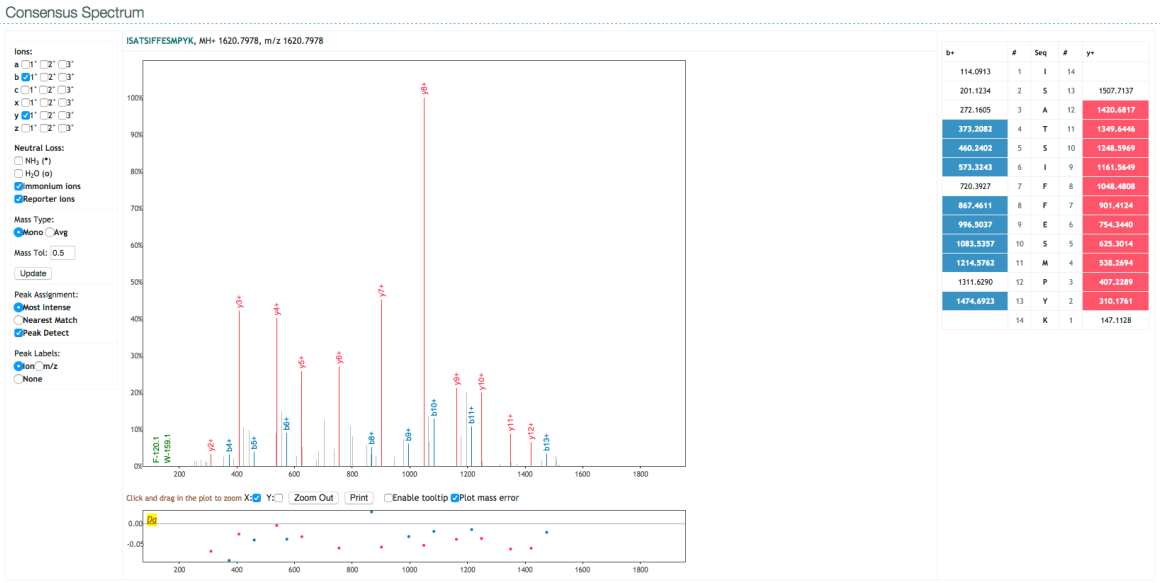

Figure 4.3: Consensus spectra together with the corresponding peak annotation are visualized using the Lorikeet spectrum viewer (http://uwpr.github.io/Lorikeet).

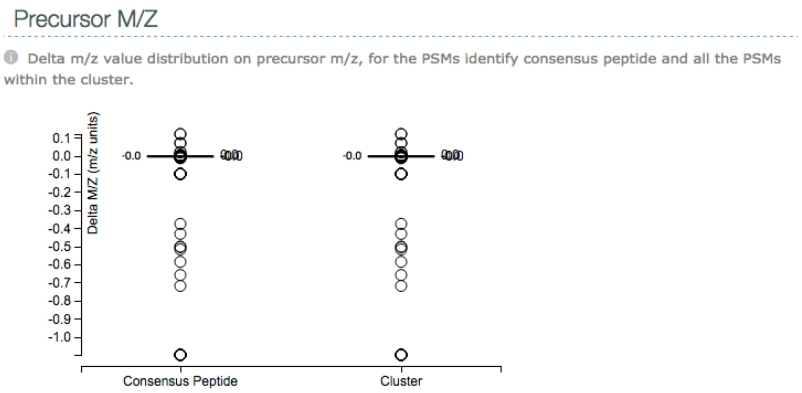

Figure 4.4: The precursor *m/z* delta mass distribution (difference to the theoretical mass of the most common peptide) is visualized as a box plot. The "Consensus Peptide" plot only takes spectra into consideration that were identified as the most common peptide. The "Cluster" plot visualizes the distribution of all spectra within the cluster.

Supplementary Note 5 – Benchmarking the PRIDE Cluster Human Spectral Library

As a benchmark we used a combination of NIST's (National Institute of Standards and Technology) human Orbitrap (November 2014), iontrap (May 2014) and the global proteome machine's (GPM) common Repository of Adventitious Proteins' (cRAP) (downloaded on July 2014). The PRIDE Cluster human spectral library (version 2015-04) contains the consensus spectra of all reliable clusters that contain spectra observed in human datasets in PRIDE Archive. Thereby, it does not only contain spectra coming from peptides with multiple different PTMs, but also automatically the spectra from common contaminants found in human datasets.

First of all, we identified the consensus spectra of both libraries using X!Tandem[5] (version Sledgehammer, 2013.09.01.1) using UniProt's human proteome library (release 2014-07) concatenated with the cRAP sequence database (downloaded on July 2014) and reversed decoy sequences. The precursor tolerance was set to 3 *m/z* units*,* fragment tolerance to 0.4 *m/z* units and the refinement mode disabled. Carbamidomethylation was set as fixed modification and oxidation of M and N-terminal acetylation as variable modifications. The search results were filtered at 1% peptide FDR. Both libraries showed similar accuracy based on this assessment (Table 5.1). The lower fraction of identified spectra from the PRIDE spectral library can be explained by the fact that the library also contains modified peptides whose modifications were not considered as parameters during the search.

	Identified Spectra	**Identical Identifications**
PRIDE human	34%	96%
NIST human	49%	95%

Table 5.1: Identified library spectra using X!Tandem.

We additionally searched the spectra available in the PRIDE human spectral library using SpectraST (version 5.0)[6] and the NIST's library (see above). Decoy spectra were created using SpectraST and appended to the library. The results were again filtered at a 1% peptide FDR and deviating results compared to the X!Tandem search were discarded. This search identified 37% of the library's spectra and identified 96% of the spectra as the same peptide.

Finally, we used the above mentioned NIST spectral library and the PRIDE human library to identify the spectra from a state of the art experiment: an HeLa digest recently published by Köcher *et al.* measured on a QExactive mass spectrometer[7] (PRIDE Archive identifier PXD000396, using the run *120312QEx2_RS1_20nl-min_0k1HeLa_14h_01.raw*). The raw file was converted using ProteoWizard's msconvert (216,102 MS/MS spectra)[8]. The NIST search results were filtered at 1% peptide FDR.

Unfortunately, SpectraST was not able to generate decoy spectra for our library since several reported PTMs from the PRIDE library are not supported by SpectraST. Work is under way to resolve this issue and add support for additional PTMs in SpectraST. Therefore, for this comparison the results gathered using the PRIDE spectral library were filtered at the same p-values derived from the NIST library.

Using these settings, 32,351 and 42,265 spectra were identified using NIST's and the PRIDE library, respectively. Nevertheless, it must again be highlighted that the larger number of identifications using the PRIDE library is most likely due to the inadequate filtering of the search results, for the limitation mentioned above. Nevertheless, 19,951 spectra were identified by both libraries of which 98% were identified as the same peptide.

Supplementary Note 6 – Probabilistic Scoring Approach

The probabilistic scoring approach used in the *spectra*-cluster algorithm is an adaption of the probabilistic matching used in Pepitome[9].

After the initial peak filtering (the 70 highest peaks per spectrum are used), a second filtering step is added where only the peaks that explain 50% of the total ion current but at least the 25 highest peaks are used for the matching between the two spectra. If multiple peaks match within the user-defined fragment ion tolerance, the matching pair with the lowest *m/z* difference is selected.

The Hypergeometric Score estimates the probability that the number of matching peaks occurred by random chance. This is modelled by a hypergeometric distribution described by the equation:

$$p(k) = \frac{\binom{s}{k}\binom{N-s}{n-k}}{\binom{N}{n}}$$

Where *k* is the number of matched peaks, *N* the total number of bins (determined by the fragment tolerance), *s* the number of peaks from spectrum 1, *n* the peaks from spectrum 2.

In contrast to Pepitome we only calculate point probabilities to increase performance and make the scoring usable for clustering. Additionally, the similarity between the intensity ranks of the matched peaks is assessed using the Kendall-Tau score[10]. The raw Kendall-Tau score is converted into a probability of obtaining better than the observed intensity correlation by random chance, which is approximated using a normal distribution with $\mu = 0$ and $\sigma 2 = 2(2k + 5)/9k(k - 1)$.

These two probabilities are combined into a single p-value using Fisher's method[11] which is then reported as its negative logarithm.

Supplementary Note 7 – The *spectra-cluster* algorithm

The *spectra-cluster* algorithm is mostly based on the MSCluster algorithm by Frank *et al.*[2].

7.1 Spectrum Filtering

First, precursor ion associated peaks, taking potential neutral losses into consideration are removed. Then, only the 70 highest peaks from the spectrum are retained and the intensities normalised so that the total spectrum intensity (sum of intensities of all peaks) is 1,000.

7.2 Clustering

The *spectra-cluster* algorithm uses a greedy clustering approach where spectra (or clusters) are merged to the first cluster exceeding the set threshold. To alleviate the main disadvantage of this approach (since sub-optimal matches may occur early in the clustering process and thereby lead to sub-optimal results) an iterative approach is used starting with very stringent thresholds that are subsequently decreased to reach the target accuracy. The accuracy is defined as the proportion of correctly classified spectra (*ie.* 99%). The similarity score returned by the used similarity function is converted into this measurement of accuracy using an empirically derived cumulative distribution function (CDF, Supplementary Note 8).

7.2.1 Pre-iterative clustering

In the first clustering round, spectra are binned in 0.2 *m/z* units wide bins. Only spectra that share one of the five highest peaks are compared and clustered at a target accuracy of 99.9%. Additionally, if more than 10,000 clusters are in memory, these are written to disk and the clustering process is continued without these clusters.

7.2.2 Iterative clustering

The following process is repeated 4 times using decreasing accuracies, starting at 99.9% to the final accuracy of 99%. Spectra are binned in 4 *m/z* units wide bins. In the first of these rounds, again, only spectra that share one of the five highest peaks are compared. In subsequent rounds, only spectra (or clusters) that were previously compared and scored among the top 30 matches are compared. If a spectrum or a cluster exceeds the set threshold it is merged with the existing cluster and from that point, only the new consensus spectrum is used for further comparisons.

7.3 Consensus spectrum building

The algorithm used to build consensus spectra is the same than the one used for the previous PRIDE Cluster algorithm[4] and originally described here[2, 12]. The final *m/z* threshold used is set to 0.4 *m/z* units starting from 0.1 *m/z* units, using 0.1 *m/z* units step increases.

7.3.1 Algorithm

For every peak in the consensus spectrum, information about the *m/z* value, intensity and in how many spectra the peak was observed is stored. Since the total number of spectra contributing to the consensus spectrum is known, a peak's probability to be observed in a spectrum can be calculated.

1. Add all peaks from all spectra to the consensus spectrum (*CS*). In case two peaks have an identical *m/z* value, add the intensities and increment how often the peak was observed.
2. Merge identical peaks.
    a. Start at a tolerance of 0.1 *m/z* units- increment by 0.1 *m/z* units until 0.4 *m/z* units are reached.
    b. Merge peaks within the tolerance. Use the weighted average *m/z* (weighted based on the peak's intensities) as the new *m/z.*
3. Adapt peak intensities based on how often they were observed (*Pi*): $I = I * (0.95 + 0.05 * (1 + Pi)^5)$

    where:

    I is the peak's intensity.

    Pi is the probability the peak is detected in a spectrum.

    This formula multiplies the observed intensity by 1 - 2.55.
4. Filter the consensus spectrum:
    a. Keep only the top 5 peaks within every 100 *m/z* units window.

Supplementary Note 8 – Cumulative Distribution Function

To benchmark various combinations of scoring functions and peak picking methods, we generated so-called cumulative distribution functions based on the 209 human datasets used to evaluate clustering accuracy (Supplementary Table 1). We evaluated the hypergeometric score (as point probability and cumulative probability), Kendal-tau statistics, dot-product and the combination of these scores. Additionally, we evaluated the following peak picking algorithms: no peak picking, N highest peaks with 50, 70, and 100 peaks, total ion current filter (percentage of peaks explaining X% of the spectrum's total intensity) for 50% and 70%, and a filter retaining the 6 highest peaks per 100 *m/z* units.

Spectra were randomly selected from a single dataset and compared using the evaluated scoring function. Only incorrect matches were evaluated (defined by a differently identified peptide sequence and a precursor *m/z* difference of at least 4 *m/z* units) and the corresponding similarity score recorded. A maximum of 1 billion comparisons were performed per dataset leading to a total of >100 billion comparisons per score and peak picking combination.

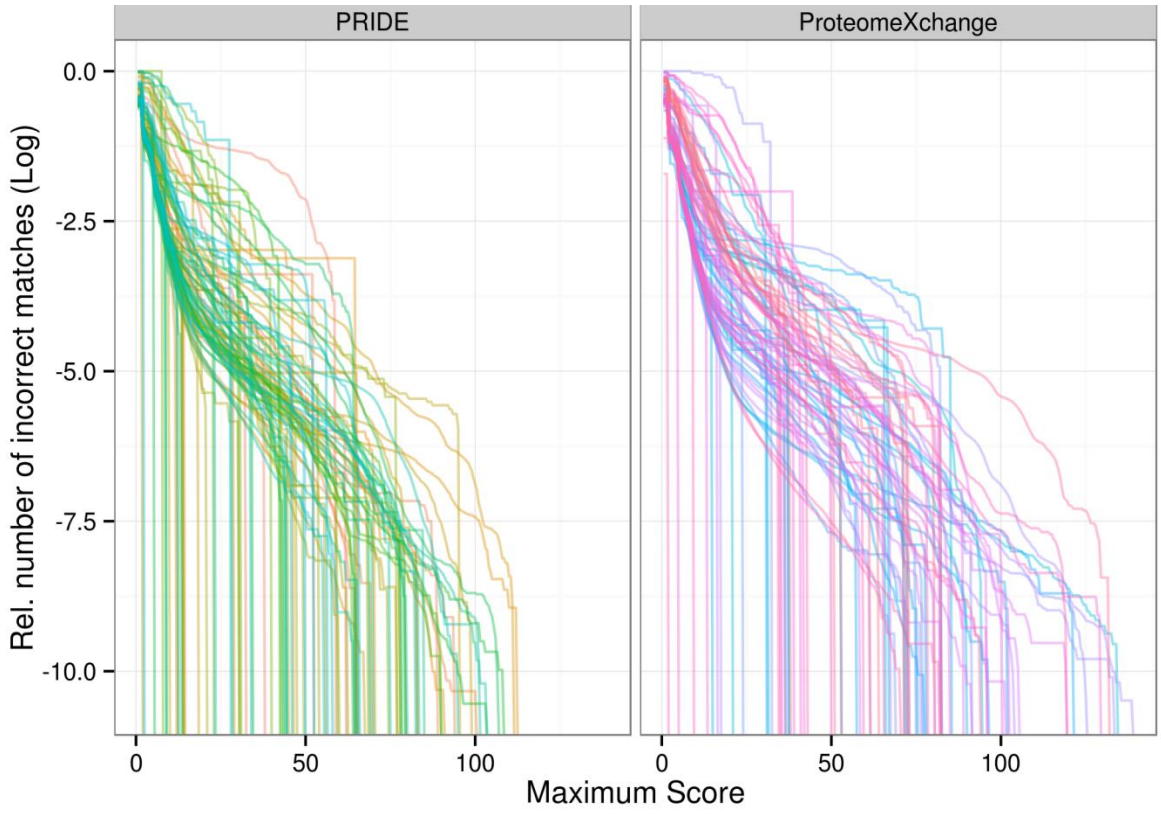

Figure 8.1: The derived cumulative distribution function (CDF) was highly dependent on the original dataset. The plot depicts the CDFs for the entire 209 test datasets, split by datasets submitted to PRIDE (datasets submitted to PRIDE before mid 2012) or to ProteomeXchange (more recent dataset submissions, since mid 2012, once the ProteomeXchange data workflow was implemented). One line represents the results from one single dataset.

Figure 8.1 shows the results for the new probabilistic score using the cumulative probability split by all the evaluated test datasets. As reported previously, the score shows different behaviour in different datasets. This seemed to be independent of the time when the data was originally submitted to PRIDE Archive (PRIDE represents dataset submissions up to mid 2012, ProteomeXchange includes dataset submissions since mid 2012). This behaviour was observed for all evaluated scores and score combinations.

During the spectral clustering process, a single spectrum is compared to potentially thousands of other spectra. Therefore, we need to correct the acquired scores for this multiple testing. This was based on the average Cumulative Distribution Function (CDF) observed for the score. To account for the large variability of the observed distributions, we added a conservative error margin and set the minimum number of observations to 5,000. This means, that even if a spectrum is compared to less than 5,000 other spectra, the probability still is adapted as if 5,000 comparisons were performed. This approach gave us comparable results across most datasets. When setting the specified accuracy to 99%, the proportion of incorrectly clustered spectra found was exactly 1%, as expected (see Figure 1, main manuscript).

References

1. Omenn, G.S. et al. Metrics for the Human Proteome Project 2015: Progress on the Human Proteome and Guidelines for High-Confidence Protein Identification. *Journal of proteome research* **14**, 3452-3460 (2015).
2. Frank, A.M. et al. Spectral archives: extending spectral libraries to analyze both identified and unidentified spectra. *Nature methods* **8**, 587-591 (2011).
3. The, M. & Kall, L. MaRaCluster: A Fragment Rarity Metric for Clustering Fragment Spectra in Shotgun Proteomics. *Journal of proteome research* **15**, 713-720 (2016).
4. Griss, J., Foster, J.M., Hermjakob, H. & Vizcaino, J.A. PRIDE Cluster: building a consensus of proteomics data. *Nature methods* **10**, 95-96 (2013).
5. Craig, R. & Beavis, R.C. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* **20**, 1466-1467 (2004).
6. Lam, H. et al. Development and validation of a spectral library searching method for peptide identification from MS/MS. *Proteomics* **7**, 655-667 (2007).
7. Kocher, T. et al. Development and performance evaluation of an ultralow flow nanoliquid chromatography-tandem mass spectrometry set-up. *Proteomics* **14**, 1999-2007 (2014).
8. Chambers, M.C. et al. A cross-platform toolkit for mass spectrometry and proteomics. *Nature biotechnology* **30**, 918-920 (2012).
9. Dasari, S. et al. Pepitome: evaluating improved spectral library search for identification complementarity and quality assessment. *Journal of proteome research* **11**, 1686-1695 (2012).
10. Kendall, M.G. A new measure of rank correlation. *Biometrika* **30**, 81−93 (1938).
11. Mosteller, F. & Fisher, R.A. Questions and Answers. *American statistician* **2**, 30 (1948).
12. Frank, A.M. et al. Clustering millions of tandem mass spectra. *Journal of proteome research* **7**, 113-122 (2008).