

Transformation and other factors of the biological Mass Spectrometry pairwise peak-list Comparison Process

Witold E Wolski^{*1,7}, Maciej Lalowski⁴, Peter Martus⁶, Ralf Herwig¹, Patrick Giavalisco⁵, Johan Gobom¹, Albert Sickmann³, Hans Lehrach¹, Knut Reinert²

¹Max Planck Institute for Molecular Genetics, Ihnestr  e 63-73, D-14195 Berlin, Germany

²Institute for Computer Science, Free University Berlin, Takustr. 9, D-14195 Berlin, Germany

³DFG Research Center for Experimental Biomedicine, University of W  rzburg, Versbacherstr. 9, D-97078 W  rzburg, Germany

⁴Max Delbr  ck Center for Molecular Medicine, Robert-Roessle-Str. 10, D-13125 Berlin-Buch, Germany

⁵Boyce Thompson Institute for Plant Research, Tower Road, Ithaca 14850, NY, USA

⁶Institute for Medical Informatics, Biometry and Epidemiology; Charite University Medicine Berlin, Hindenburgdamm 30 (HBD 30), 12200 Berlin

⁷present address: School of Mathematics and Statistics, Merz Court, University of Newcastle upon Tyne, NE1 7RU, UK

Email: Witold E Wolski^{*} - witek96@users.sourceforge.net; Maciej Lalowski - m.lalowski@mdc-berlin.de; Peter Martus - peter.martus@charite.de; Ralf Herwig - herwig@molgen.mpg.de; Patrick Giavalisco - npg5@cornell.edu; Johan Gobom - gobom@molgen.mpg.de; Albert Sickmann - albert.sickmann@virchow.uni-wuerzburg.de; Hans Lehrach - lehrach@molgen.mpg.de; Knut Reinert - reinert@inf.fu-berlin.de;

^{*}Corresponding author

Abstract

Background: Biological Mass Spectrometry is used to analyse peptides and proteins. A mass spectrum generates a list of the measured mass to charge ratios and intensities of ionised peptides called a peak-list. This information is further used to classify the underlying amino acid sequence by comparing spectra directly. Development of suitable methods of peak-list analysis can be advantageous for many applications.

Methods: The pairwise peak-list comparison is a multistage process consists of: matching peaks embedded in two peak-lists, normalisation, scaling of peak intensities, and of dissimilarity measures. In order to study the criteria influencing the peak-list comparison, we analysed two large datasets consisting of Tandem Mass Spectrometry (MS/MS) and Peptide Mass Fingerprinting (PMF) data. For this reason, we employed analysis of variance type methods using the PMF data as a learning sample, while the MS/MS data served as a validation sample.

Results: Using pairwise peak-list measures we were able to reproduce the assignments made by the database search engines. For both MS/MS and PMF data, the dot-product intensity-based measure computed on log-transformed intensities reached the highest performance. Employing the Fowlkes-Mallows statistics allowed the best performance of binary measures. The results obtained for the PMF dataset were confirmed by MS/MS data.

Conclusions: The results presented here provide an extensive analysis of different factors underlying distance measurements of MS and MS/MS-based protein data and can therefore serve as a reference point for possible applications.

Background

In recent years, mass spectrometry (MS) has emerged as a powerful technique to identify proteins in biological samples [1–4]. For their identification, proteins are usually cleaved into peptides by a protease of known and restricted cleavage specificity, *e.g.* trypsin. The resulting cleavage products can then be analysed by Peptide Mass Fingerprinting (PMF) [5], or subjected to MS/MS fragment ion analysis [6,7]. A PMF is a highly specific set of peptide molecular masses derived from one isolated protein. PMFs can be used to identify the analysed protein in large protein sequence databases by matching the determined peptide molecular masses to values calculated from the amino acid sequences in the database. Similarly, MS/MS spectra can be used for protein identification by searching the determined peptide fragment ion masses against predicted ones, based on amino acid sequence data and fragmentation characteristics of the employed MS instrumentation [8].

Before performing database searches, the MS spectra are processed and the most informative features, namely the monoisotopic peaks, are extracted. This processing consists of several steps, including smoothing, baseline subtraction, peak-extraction and monoisotopic peak determination [9,10]. Spectra pre-processing is usually performed using proprietary software provided by the mass spectrometric instrument vendors. During processing a list of mass over charge (m/z) values of the monoisotopic peaks, and either the area under or the height of those peaks, are obtained. This set of m/z and intensity value pairs is called *peak-list* (PL). In case of PMF data the PL has on average 36 peak/intensity pairs,

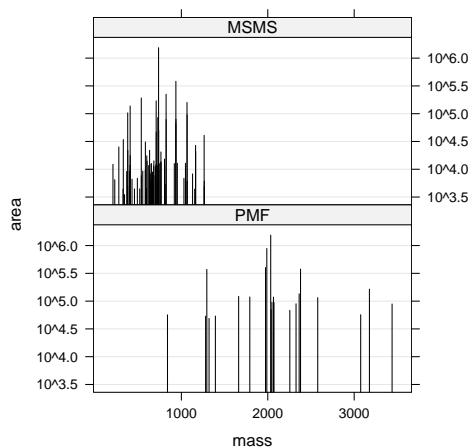


Figure 1: Example of a peak-list stick spectrum for fragment ion MS/MS (top panel) and PMF(bottom panel). X-axis – mass of the peaks, Y-axis – area under the peak.

compared to *e.g.* 100,000 data points of the raw spectra. Figure 1 presents representative PLs of fragment ion MS/MS and PMF, respectively.

The sensitivity and specificity of the peptide identification using database searches can be increased by several methods. This usually includes: calibration [11–13], identification of non-peptide peaks [12, 14, 15], identification and removal of low-quality spectra [16, 17], or validation of the search results using machine-learning algorithms [18, 19]. The sensitivity and specificity of peptide and protein identification can further be increased by the pairwise comparison of the PLs [11, 20–22]. (An additional way to process the data with the aim of increasing sensitivity and specificity of peptide and protein identification is the pairwise comparison of the PLs [11, 20–22].) Yates et al. [20] applied the cross-correlation score, normalised by the auto-correlation of the spectra. They demonstrated that using this measure, the MS spectra could be correctly classified according to their peptide content even if acquired on two different instruments, namely a Triple-Quadrupole Tandem (TSQ) and Quadrupole Ion Trap mass spectrometers (LCQ). They also suggested using pairwise spectra comparison in order to search MS spectra against a reference library (a library of identified spectra), prior to database searching as part of a “subtractive analysis technique”. Tandem mass spectra (unique to an experiment) could be targeted for database searches or *de novo* interpretation. The data size reduction by spectra clustering, using pairwise similarities, became even more important due to high throughput proteomics projects [23, 24], where thousands of protein samples are measured and hundreds of thousands of Tandem MS spectra are obtained. A significant portion of peptides

is analyzed and identified many times even when the instruments control software attempts to prevent the repeated isolation and fragmentation of particular peptides in order to increase the diversity of spectra acquired.

Gentzel et al. [11] used the cross-correlation measure for MS/MS spectra comparison. They computed the similarity score for two parts of the spectra. If these parts exhibited a satisfying similarity score, the spectrums were assumed to be identical. Tabb et al. [22] explored the performance of the normalised dot-product (spectral angle) algorithm to identify duplicated samples. The advantage of the dot-product measure over the cross-correlation algorithm lies in its computational speed. Based on this measure, Tabb et al. [22] and Beer et al. [21] developed software to identify the similar MS/MS spectra. By selection of high-quality group representatives, or applying spectra merging algorithms [11], the quality of the spectra can be increased and analysis time decreased. The saved analysis time can be used to perform more extensive searches in other databases, *i.e.* expressed sequence tag (EST) databases, or to apply computationally demanding, mutation-tolerant search algorithms [25] which depend on partial spectra interpretation [26–28].

Pairwise spectra comparison can also be used “as an informative marker to identify organisms or some other feature of an organism” [20]. For example, Svetnik and Liaw [29] used pairwise spectra comparison to detect novel outliers in large-scale cosmid screening experiments [30]. They used the Pearson and Spearman correlation measures, as well as the Euclidean distance to compute the distances of the spectra, followed by sequential clustering. Serum protein and peptide fingerprints were used in diagnostic medicine to distinguish healthy individuals from those with cancer [31–34].

Our aim was to explore the performance of various distance measures for the classification of spectra. So far the performance of the dot-product measure was compared to the similarity index using MS/MS spectra of structural isomers [35]. Several methods exist to assess the similarity of two random variables. These methods differ in how they account for different aspects of the MS data, which generates alternate results. In this study, using two experimental data sets of PLs (PMF and MS/MS data), we have evaluated several of these measures, namely: the Euclidean and the Manhattan distance, the covariance, the sum of agreeing intensities and the spectral angle. Furthermore, we investigated quantitative measures, *i.e.* Huberts Γ or the relative mutual information measure [36]. In addition, we have studied the effect of parameters such as the weight of the mass measurement reproducibility and the weight of non-matching peaks on the performance of the dissimilarities. We also have evaluated how different transformation and scaling procedures of peak intensities influenced the dissimilarity measure [22, 37, 38].

This article places major emphasis on the fact that peak matching, weighting of measurement accuracy, scaling and normalisation of intensities, and finally the dissimilarity measures form an interdependent *Comparison Process* (CP). The combination of these factors results in 96 approaches for the binary measures and 2688 approaches for the intensity based measures. The comparison of the CPs was performed according to their specificity and sensitivity. To structure these approaches, the significance and strength of them were studied using analysis of variance techniques. The limitations of this approach are discussed in the Conclusions chapter.

Results and Discussion

Amount of duplication in the datasets

The amount of duplicated samples in two datasets (PMF and MS/MS) are presented in Table (1). To identify the PMF spectra, we used the Mascot search algorithm [39], while for the MS/MS spectra the SEQUEST search algorithm [8, 40] was employed. In the PMF dataset, only 176 proteins out of 668 were identified by a single spectrum, while the remaining 492 proteins resulted from 2160 database searches. To evaluate the classification ability of the CPs we used the duplicated spectra, that is spectra assigned to one database identifier (ID) for the PMF data or to the same peptide sequence and the same parent ion charge $z = 2$ for the MS/MS data. We assumed that by the database search a *true* identification of the peptides and proteins was possible. For two PLs X and Y , we defined that they laid *within* a cluster if they were assigned by the database search engine to a protein sequence with the same database ID (PMF), or to a peptide sequence and charge (MS/MS). Similarly two PLs were defined to lay *between* the clusters if their database IDs differed.

Using data where the identities of the samples were determined by database search algorithms, we were able to examine if the pairwise PL comparison made equal or different assignments to a group (equal or lower sensitivities/specificities for this set of data), than the database search algorithm. We were not able to reveal, if any of these measures had higher sensitivities and specificities than the DB search algorithms used. We could conclude that the amount of duplication in the case of both datasets was significant and that recognising them could significantly reduce the number of searches necessary to identify all proteins.

The number of matches

The intensities of individual peaks may vary considerably between spectra, but the m/z values of fragment ions can be measured with at least the accuracy of a single m/z in majority of the mass spectrometers. If

	Dataset	
	PMF	MS/MS
number of spectra	4532	$\approx 200000^A$
number of identified spectra	2336	26507 ^B
$N = 1$	176	5718
$N \in (1, 5]$	392	1965
$N \in (5, 10]$	66	388
$N \in (10, 25]$	31	354
$N \in (25, 50]$	2	111
$50 < N$	1	43
Identified : proteins (PMF) peptides (MS/MS)	668	8579

Table 1: Number of clusters of given cluster size N . The columns 2 and 3 describe the cluster size in the PMF- and the MS/MS datasets. Number of spectra – number of PLs submitted for database search, identified spectra - spectra assigned to a database ID with an either significant probability based Mowse score (PMF-data) or to a peptide sequence with $Xcorr > 2$, and an ion coverage $> 20\%$ (MS/MS-data) given a parent peptide charge $z = 2$. Identified proteins/peptides - the number of uniquely identified proteins or peptides. ^A – approximate number of spectra derived from ion fragments of peptides with charge $z = 2$. ^B – The number of spectra with charge $z = 2$ of the parent ion ($\approx 53\%$ of all identified spectra).

the primary fragment ions/peptides in a pair of spectra has the same m/z locations, the spectra are judged to result from the same peptide/protein, regardless of their peak intensities.

Table 2, rows 2 and 5, summarises the properties of the PLs like the mass measurement error (MME) and the mass measurement range. Furthermore it provides the *five-number summary* and the mean of the PLs lengths. The probability of i matches given two independent PLs of known length, known mass measurement range and resolution can be modelled using the hyper-geometric distribution [41]. The observed number of matches of *within* and *between* cluster PL pairs are given in Table 2, rows 3,4, and 6,7. In case of the PMF data, we observed for PLs drawn from *between* clusters, a higher number of matches than expected for independent PLs. This difference is due to the fact that proteins analysed from two dimensional (2D) gel electrophoresis samples are not completely separated [42], and because sequence database entries have different database IDs even if the protein sequences exhibit a high fraction of sequence identity (*i.e.* protein families). Nevertheless, the number of matching peaks had a high power to discriminate PLs as being *within* or *between* clusters.

For 75 clusters of various size (2 – 20 samples/cluster) chosen 5 times from the PMF dataset, we have computed the number of matching peaks for all PL pairs. The number of matching peaks was in almost all cases higher, if the PLs compared laid *within* one cluster (magenta histogram, Figure 2 A), than if they laid *between* different clusters (green histogram, Figure 2 A). For example, 95% of *within* cluster PL pairs had more than 4 matches, but only 1% of *between* cluster PL pairs held more than 4 matches. The cases

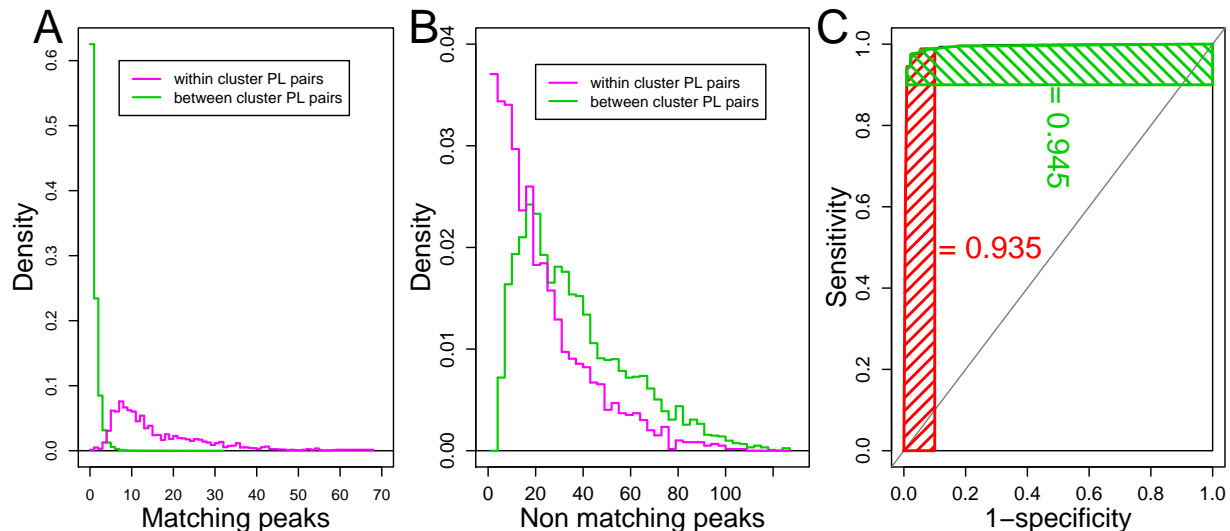


Figure 2: Figure **A** – Histogram of the number (bandwidth = 1) of matching peaks for peak-lists chosen from the same cluster (magenta) and from different clusters (green). Figure **B** – Histogram of the number (bandwidth = 3) of non-matching peaks, if PLs were chosen from the same (magenta) or from different clusters (green). Figure **C** – Receiver Operator Characteristic curve - The sensitivity (TP-rate) is plotted against $FP = 1 - specificity$ using the number of matching peaks as the discriminatory variable. Red dashed area: sensitivity-PAUC – partial area under the ROC curve for FP-rate $\in [0, 0.1]$. Green dashed area: specificity-PAUC – partial area under the ROC curve for sensitivities $\in [0.9, 1]$.

where the number of matches between PL from *within* one cluster equalled zero, could be explained by the fact that the spectra were measured on non-overlapping fragments of the same protein.

The masses of randomly matching peaks will, on average, differ more than the masses of non-random matching peaks. Therefore, weighting of mass measurement accuracy using a triangular function (see Equation 2)¹ was implemented. By this function, the weight of peaks with a small overlap is reduced. Furthermore, in case of the MS/MS PLs, clusters of peaks separated by a mass smaller than the mass measurement accuracy (which is used for searches of matching peaks) were observed. Therefore, if matching two PLs, ambiguous matches (that is, a peak is assigned to more than one peak in the second PL) occurred (see Figure 7, case A). To be able to generate a unambiguous pairwise assignment of peaks we computed the non-crossing matching using standard dynamic programming techniques (cf Methods - Finding the matching peaks).

We conclude that the probability of matches between independent PLs is higher in case of MS/MS than PMF data because of its lower mass measurement accuracy, smaller mass range and larger number of peaks. Hence, the number of matches has a lower discriminating power in case of MS/MS than of PMF data.

¹Equations are provided in the Methods section.

Data	MME [Da]	Mass range [Da]							
		Min.	Max.	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
PMF	± 0.1	713	4050	3	17	30	36	50	124
matching peaks <i>between</i> clusters PLs				0	0	0	0.62	1	32
matching peaks <i>within</i> clusters PLs				0	7	12	15.4	21	68
MS/MS	± 0.5	129	2000	35	97	134	136	170	354
matching peaks <i>between</i> clusters PLs				0	9	15	16	22	94
matching peaks <i>within</i> clusters PLs				8	44	56	57	69	133

Table 2: Peptide (PMF) peak-list and peptide fragment ions (MS/MS) peak-list properties. MME – mass measurement error. The rows 3 and 6 provide a *five-number summary* and the *mean* of the peak-lists lengths (number of peaks in peak-list) in the dataset. Rows 4,5 (PMF) and 7,8 (MS/MS) provide the *five-number summary* and the *mean* of the number of matches observed if comparing *within* and *between* cluster peak-lists pairs. Min. - minimum, 1st Qu. - first quartile, 3rd Qu. - third quartile, Max. - maximum

The number of non-matching peaks

To discriminate PL pairs as being *within* or *between* clusters the number of non-matching peaks can be used. Figure 2 B shows histograms over the number of *non-matching* peaks while comparing PL pairs (in magenta – the number of peaks that did not match if we compared two PLs *within* a cluster; in green – the number of peaks that did not match if we compare PLs pairs *between* two clusters). We observed that the probability of encountering a *within* PL pair increased if the number of non-matching peaks is small. Thus, we were also interested in utilising this feature for better discrimination of *between* and *within* cluster PL pairs. Therefore we evaluated the performance of the following asymmetric binary measures: Gower coefficient (Equation 18) and Fowlkes-Mallows statistics (Equation 19). These measures account for the number of matches and mismatches. If the length of the aligned PLs is defined (see Equation 12), also symmetric binary measures *e.g.* Huberts Γ (Equation 20) and relative mutual information (Equation 23) can be used. Furthermore, we examined if increasing or decreasing the weight of non-matching peaks by the factor of two can increase the performance of the CP.

The evaluation scores

We compared the various *Comparison Processes* (CPs) according to their ability to determine if a PL pair was chosen from *within* one cluster or *between* two clusters. The performance of the CPs was determined using the *Partial Area* (PAUC) under the *Receiver Operating Characteristic* curve (ROC) [43]. The ROC curve (Figure 2 C) was obtained, by drawing the $sensitivity = \frac{TP}{TP+FN}$, where TP - true positives, FN - false negatives against the $1 - specificity = FP_{rate} = \frac{FP}{FP+TN}$, where FP - false positives, TN - true negatives, for the same value of the discriminatory variable, *i.e.* the number of matching peaks. For 4

matches we determined a specificity of 99% and sensitivity of 95%.

Because we were interested in the sensitivity of the CPs only for small values of the *FP*-rate, we computed the PAUC for $1 - \text{specificity} \in [0, 0.1]$ (red-dashed region Figure 2 C), denoted by *sensitivity-PAUC*.

Moreover, we were interested in the specificity of the CPs if high sensitivities are required (*sensitivity* $\in [0.9, 1]$). Hence, we computed the PAUC for the area indicated in Figure (2 C) by the green-dashed region representing the specificities given sensitivities from 0 – 90%, further abbreviated *specificity-PAUC*.

Peak intensities

As well as the mass of the monoisotopic peaks, the intensity of the peaks in spectra was calculated by the feature extraction process. Intensities associated with the masses observed at least twice *within* a cluster (magenta density, Figure 3 A) tend to have higher peak intensities, compared to intensities of peaks whose masses are observed only once *within* a cluster (grey density, Figure 3 A). Intensities I_X and I_Y , of matching peaks in PLs from *within* a cluster, were more strongly correlated ($\text{corr}_{PMF}^{(within)}(I_X, I_Y) = 0.57$, $\text{corr}_{MS/MS}^{(within)}(I_X, I_Y) = 0.61$) (Figure 3 B) than those obtained from *between* clusters ($\text{corr}_{PMF}^{(between)}(I_X, I_Y) = 0.17$, $\text{corr}_{MS/MS}^{(between)}(I_X, I_Y) = 0.04$)² (Figure 3 C). So, the intensity of peaks could be employed for better discrimination of *within* and *between* cluster PL pairs. In line with this observation we studied the performance of the intensity based measures: the covariance (Equation 8), the dot-product (Equation 7), the Manhattan and Euclidean distances (Equation 9), the relative distances Canberra and similarity index (Equation 10), and the sum of agreeing intensities (Equation 11).

Peak intensity transformation

If two peaks match *within* a cluster, the peak intensities are very likely (except from random matches) to be estimates of the number of the ions of the same peptide (PMF) or peptide fragment (MS/MS). The estimates might contain errors; reasons include: random noise, different levels of peptide fragmentation due to variations in *Collision energy*, different signal-to-noise ratios due to varying concentrations of peptide present [22]. The observed error can depend on the observed intensity. Thus, any statistical model would either directly account for the variances or transform the data so that the variances are approximately equal for all peak intensity levels.

The Altman-Bland plots [44] in Figures 3 B and C, show the residues ($\Delta I = I_X - I_Y$) as a function of the

²The correlation was determined for log transformed and root mean square scaled peak intensities.

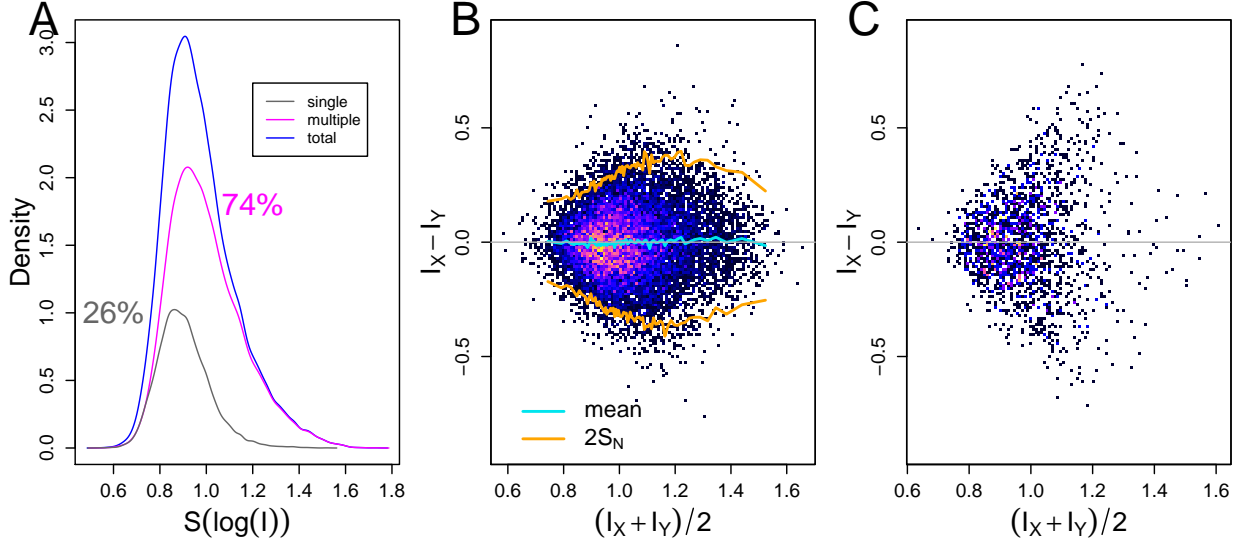


Figure 3: Peak Intensities. Figure **A**) Histogram of intensities: X-axes – Intensity of log transformed root-means-square scaled peak intensities. Y-axis – Frequency. In grey: Histogram of the peak intensities that do **not** match a peak in any other peak-lists (PLs) *within* the same cluster (this mass is observed only once in the cluster). In magenta: Histogram of intensities of peaks that do *match* a peak within any PL *within* cluster (this mass is observed at least twice in the cluster). Figure **B**) Altman Bland plot of intensities of the matching peaks for PLs pairs from *within* a cluster. Figure **C**) Altman Bland plot of intensities of matching peaks for PLs pairs of *between* clusters.

average peak intensity $\bar{I} = (I_X + I_Y)/2$, where I_X and I_Y are the intensities of a matching peak pair (X, Y) . The peak intensities are log-transformed and root mean square scaled (Equation 5). To find out the best variance stabilising transformation, one can examine the *proportionate reduction in variation* R^2 [45], obtained by analysis of the model $|\Delta I| \sim \bar{I} + \bar{I}^2$. This model accounts for the correlation of variance and intensity ($|\Delta I| \sim \bar{I}$), unlike the naive model $\Delta I = E(\Delta I)$ [44]. If the variance is stable, the naive model $|\Delta I| = \bar{I}$ suffices, and the R^2 obtained with the model accounting for the correlation of variance and intensity should be close to zero. In case of the log-transformation and peaks matching *within* a cluster (Figure 3 B), the adjusted R^2 of the model $|\Delta I| \sim \bar{I} + \bar{I}^2$ are 0.04 and 0.02 for PMF and MS/MS respectively. This indicates that the log-transformation gives a better variance stabilisation than the square root transformation ($R^2 = 0.32$ (PMF) respectively $R^2 = 0.16$ (MS/MS)) as suggested by Tabb et al. [22], or the raw data ($R^2 = 0.47$ (PMF) respectively $R^2 = 0.40$ (MS/MS)). Applying the log-transformation extends the scale of peaks with low intensities, while it compresses the scale of peaks with higher intensities. In addition to the raw, root-squared and log-transformed intensities we included the ranking of the intensities [29] among the transformations studied. To elucidate the extent the transformation contributes

to the PAUC score, compared to other factors, we kept the four different transformations in the CP, despite the fact that the best transformation was determined above.

The comparison process

Table 3 summarises the factors of the CP, which can influence the outcome of a pairwise PL comparison and which therefore were analysed in our study. The first step in comparison of PLs is to determine matching and non-matching peaks with given mass measurement accuracy. If one peak is ambiguously assigned to several peaks in the second PL (Figure 7), one can resolve this by computing the non-crossing matching. The next element to be considered is whether the mass measurement accuracy should be modelled [46,47] using Equation 2. Modelling of the mass differences between matching masses did not affect the non-matching peaks. Their influence on the CP was determined by increasing or decreasing their weight by the factor two, using the parameter θ in the dissimilarity equations (cf Methods).

The length of the aligned PLs either equals the sum of the peaks in both PLs minus the number of peaks matching or is user-defined. We set in in Equation 12 for the PMF dataset $N = 250$, and for the MS/MS dataset $N = 400$, which in both cases were approximately twice the length of the longest PL. The *missing* peak pairs, in case of the intensity based measures, were augmented with peaks of zero intensity. Further elements studied, which affected only the intensity-based measures, were the transformation and scaling of peak intensities. The last factor examined, consisted of the distance measures.

In case of the intensity-based measures, there were 2688 sets of factors while in case of the binary measures there were 96 sets. To determine which of these factors was important and how they influenced the scores, we applied analysis of variance techniques (ANOVA) (cf Methods Analysis of Variance). We analysed the PMF scores first and afterwards we examined if the obtained model could be used to explain the properties of the CPs computed on the MS/MS dataset.

ANOVA of the PMF dataset

A large value of sums of squares (SSQ) for a factor in relation to the total SSQ (Tables 4 and 5), indicates that it was important for the correct classification of PLs. Thus, in Table 4 (binary measures) the high value of the SSQ for the factor *measure* reflects the fact that the Fowlkes-Mallows statistics and the Gower coefficient show significantly better values of the PAUC than the Huberts Γ and the relative mutual information (Figure 5). Other important factors if applying binary measures are the weighting of non-matching peaks (θ) and the length of the PLs (Table 4 top panel, column MSQ). In case of the

Factors		Levels				Number		
						Int.	Bin	
1	non crossing matching	yes		no		2		
2	weighting match accuracy	yes		no		2		
3	weight of non-matching peaks	0.5	1	2		3		
4	intensity transformation	I	\sqrt{I}		$\log(I)$	rank(I)	4	0
5	intensity normalisation	tic(I)	$\ I\ $		$S(I)$	Z(I)	4	0
6	alignment length	$M = M_1^X + M_1^Y - M_{11}^{XY}$			$M = const$		2	
7	distance measure	See Methods Section				7	4	
Product of levels for nonzero factors:						2688	96	

Table 3: Factors considered in the comparison process and their levels. Column 1 – Factors: identification of factors, Column 2 – Levels: short summary of the levels (For more details please refer to the Methods section). Column 3 – Number: number of levels. Int. – comparisons considering the intensities; Bin. – binary measures.

intensity based measures, the high value of the mean sum of squares (MSQ) (Table 5 top panel) for the factor scale (intensity scaling procedure) and measure (dissimilarity measure), in comparison with the MSQ of the other factors, indicates that this factors were crucial for the correct classification.

A large value of MSQ or SSQ of an interaction term (denoted by \times Table 4, 5 bottom panel) demonstrates that some combinations of factors were more useful then others. For example, in Table 5 the high value of SSQ for the interaction *measure* \times *scale* reflects the fact that the measure sum of agreeing intensities behaved much better in combination with the vector length scaling (N) or root-mean-square scaling (S) than with the total ion count scaling (T) or with the z -score scaling (Z) (see Figure 5).

Figure 5 B shows the boxplot of the PAUC measure, obtained for the CPs, itemised according the factors dissimilarity measure and intensity scaling method. The Manhattan and Euclidean distances were exhibiting higher variance than the dot product measure and the sum of agreeing intensities. This is due to the weighting of non-matching peaks by θ and due to the factor PL pair length N (see Equation 12), which influenced these distances but neither influenced the outcome of the dot-product nor the sum of agreeing intensities measure (if the intensities were not z -score scaled).

The Boxplot (Figure 4) of the *sensitivity-PAUC* (left) and *specificity-PAUC* (right) score, computed using the dot-product and the sum of agreeing intensities measures (both computed on vector length scaled data), shows how the intensity transformation influenced the classification. As predicted, by the analysis of variance stabilisation, the log transformation of intensities performed best for both measures. We also observed that the variance of the PAUC, with respect to the intensity transformation, computed for the vector length normalised dot-product measure was smaller than the variance of the PAUC computed for

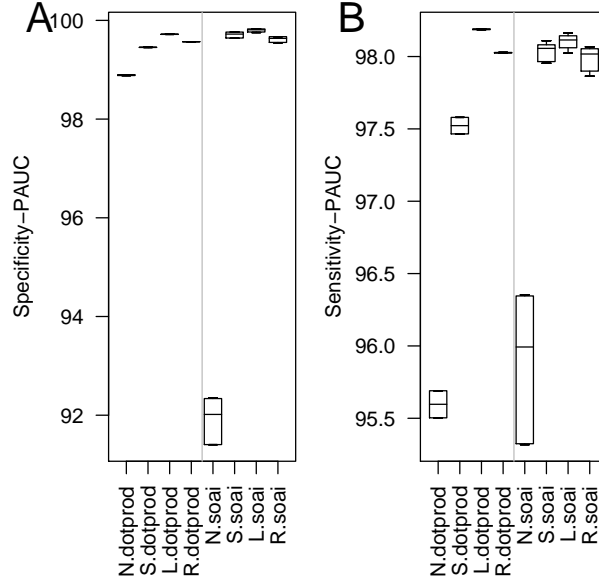


Figure 4: Figure **A**: Boxplot of the specificity-PAUC (specificity given a TP-rate $\in [0.9, 1]$) for the dot-product measure (dotprod) and sum of agreeing intensities (soai). Figure **B** Boxplot of the sensitivity-PAUC (sensitivity given a FP-rate $\in [0, 0.1]$). N – raw intensities, S – square root transformed intensities, L – log transformed intensities, R – intensity ranks.

the sum of agreeing intensities.

The optimum intensity based CP, with respect to high values of PAUCs and a small variance, was the dot-product measure computed on vector length scaled intensities (spectral angle). However, this measure was not able to achieve the maximal PAUC (see Figure 5 B, top panel). For the sum of agreeing intensities, the Manhattan distance (computed on total ion count scaled data), and the Euclidean distance (computed on vector length scaled data) similar or higher values of the PAUC were recorded.

In case of the binary measure based CPs (Figure 5 A) the largest PAUC was measured for the Fowlkes-Mallows statistics (Figure 5 left panel).

The PMF model and the MS/MS dataset

In order to validate the results obtained for the PMF-data, we compared the PAUC obtained by the Fowlkes-Mallows statistics and the vector length normalized dot product with the PAUC measured for other CPs for the MS/MS data (see Figure 6). Figure 6 A revealed, for the binary measures, that the measure Huberts Γ and the relative mutual information performed better, in case of the MS/MS data, than the Fowlkes-Mallows statistics. A possible explanation of the better performance of the symmetric

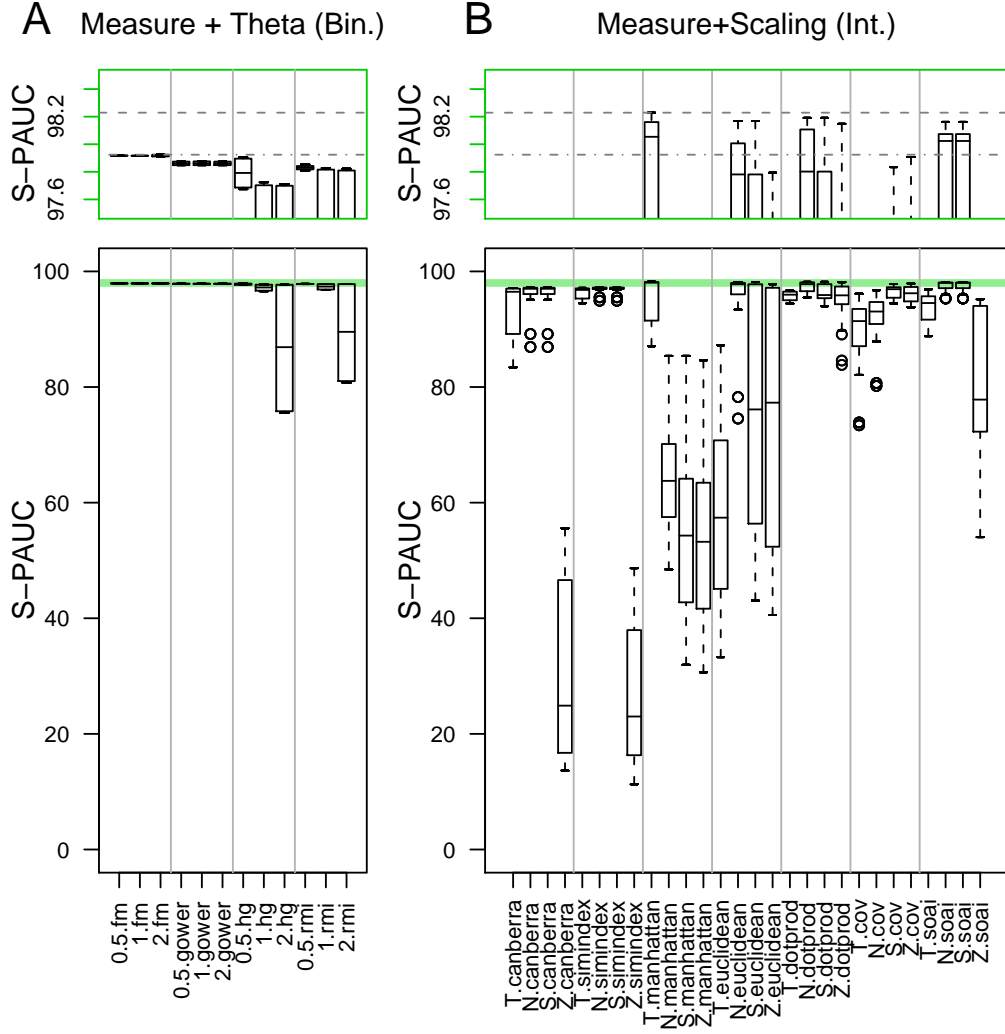


Figure 5: Figure **A**: Boxplot of the sensitivity-PAUC (sensitivity given a FP-rate $\in [0, 0.1]$) itemised according the factors *dissimilarity measure* and θ (weighting of non-matching peaks) for the binary CPs. Figure **B**: Boxplot of the factors *scale* (cf Methods - Scaling) and *measure* of the sensitivity-PAUC (sensitivity given a FP-rate $\in [0, 0.1]$) for intensity based CPs. The **top** panels show a clip (ZOOM) of the bottom boxplot, indicated by the green horizontal line. X-axis labels : fm – Fowlkes-Mallows statistics, gower – Gower coefficients, hg – Huberts Γ , rmi – relative mutual information, canberra – Canberra distance, simindex – similarity index, manhattan – Manhattan distance, euclidean – Euclidean distance, dotprod – dot-product measure, cov – covariance, soai – sum of agreeing intensities. Scaling: T – total ion count, N – vector length, S – root mean square, R – ranks

	specificity-PAUC			sensitivity-PAUC		
	Model with main effects					
Factors	SSQ	df	MSQ	SSQ	df	MSQ
measure	2712	3	904	305	3	102
θ	4621	2	2311	478	2	239
length	2662	1	2662	272	1	272
weight	0	1	0	0	1	0
noncross	0	1	0	0	1	0
residual	16669	87	192	1749	87	20
total	26666	95	300	2805	95	30
	Final model					
Factors	SSQ	df	MSQ	SSQ	df	MSQ
measure	2712	3	904	305	3	102
θ	4621	2	2311	478	2	239
length	2662	1	2662	272	1	272
measure $\times\theta$	4675	6	779	495	6	83
measure \times length	2697	3	899	281	3	94
$\theta\times$ length	4622	2	2311	478	2	239
measure $\times\theta\times$ length	4674	6	779	495	6	83
residual	1	72	0	1	72	0
total	26666	95	300	2805	95	30

Table 4: Influence of factors specifying the Comparison Process (CP) on partial areas under the ROC curve for binary PMF data. For each of the 96 CPs, sensitivity-PAUC (sensitivity given FP-rate $\in [0, 0.1]$) and specificity-PAUC (specificity given sensitivity $\in [0.9, 1]$) (Figure 2 **C**) were determined. A partitioning of sums of squares was performed analogous to analysis of variance. Column 1: identification of factors ; Column 2: raw sum of squares (SSQ) ; Column 3: degrees of freedom (DF, number of factor levels - 1) ; Column 4: Mean Sum of squares (MSQ) = SSQ/(DF) ; MSQ measures the importance of a specific factor for the size of specificity-PAUC (computed for TP-rate $\in [0.9, 1]$) and sensitivity-PAUC (computed for FP-rate $\in [0, 0.1]$). \times denotes interactions between factors. measure – distance measure, noncross – non crossing matching, length – alignment length, θ – weight of non-matching peaks, residual – unexplained SSQ or MSQ, total – column sum of SSQ, df, MSQ.

	specificity - PAUC			sensitivity - PAUC		
	Model with main effects					
Factors	SSQ	df	MSQ	SSQ	df	MSQ
measure	657	6	110	270	6	45
scale	411	3	137	300	3	100
θ	80	2	40	9	2	5
length	12	1	12	5	1	5
weight	1	1	1	0	1	0
noncross	1	1	1	0	1	0
trans	12	3	4	1	3	0.3
residual	1439	2670	0.5	763	2670	0.3
total	2611	2687	1	1348	2687	0.5
	Final model					
Factors	SSQ	df	MSQ	SSQ	df	MSQ
measure	657	6	110	270	6	45
scale	411	3	137	300	3	100
θ	80	2	40	9	2	5
length	12	1	12	5	1	5
measure×scale	873	18	49	555	18	31
measure× θ	164	12	14	26	12	2
measure×length	47	6	8	48	6	8
residual	366	2639	0.1	135	2639	0.1
total	2611	2687	1	1348	2687	0.5

Table 5: Influence of factors specifying the Comparison Process (CP) on partial areas under the ROC curve for intensity PMF data. For each of the 2688 CPs, sensitivity-PAUC (sensitivity given FP-rate $\in [0, 0.1]$) and specificity-PAUC (specificity given sensitivity $\in [0.9, 1]$) (Figure 2 **C**) were determined. A partitioning of sums of squares was performed analogous to analysis of variance. Column 1: identification of factors ; Column 2: raw sum of squares (SSQ) ; Column 3: degrees of freedom (DF, number of factor levels - 1) ; Column 4: Mean Sum of squares (MSQ) = SSQ/(DF) ; MSQ measures the importance of a specific factor for the size of sensitivity-PAUC (computed for FP-rate $\in [0, 0.1]$) and specificity-PAUC (computed for TP-rate $\in [0.9, 1]$). \times denotes interactions between factors. measure – distance measure, noncross – non crossing matching, length – alignment length, θ – weight of non-matching peaks, trans – peak intensity transformation, residual – unexplained SSQ or MSQ, total – column sum of SSQ, df, MSQ.

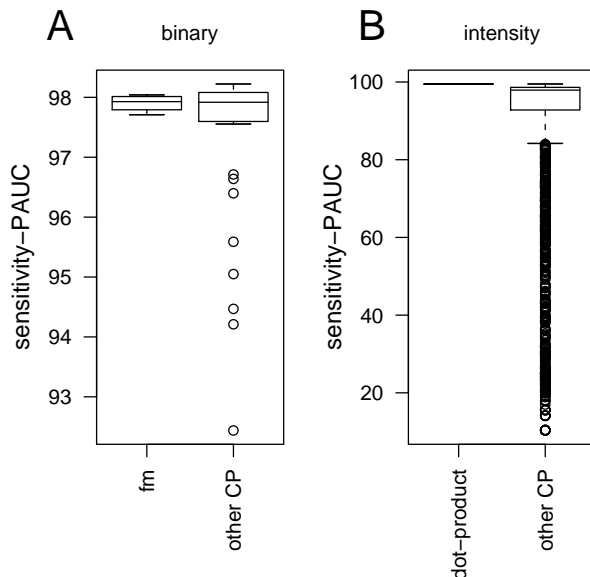


Figure 6: Boxplot **A**: Comparison of the sensitivity-PAUC (computed for FP-rate $\in [0, 0.1]$) computed for the Fowlkes-Mallows (fm) statistics with sensitivity-PAUCs of other binary measures. Boxplot **B**: Comparison of the sensitivity-PAUC (computed for FP-rate $\in [0, 0.1]$) computed for the vector length normalised dot product measure with sensitivity-PAUCs of other similarity measures.

binary measures is, that the MS/MS data have more overlapping distributions of *within* and *between* cluster matches (Table 2). In case of the intensity based CPs, the vector length normalised dot product measure demonstrated a small variance and high performance (Figure 6) reproducing the results for the PMF data. For the sum of agreeing intensities, the Manhattan and the Euclidean distances, if computed with appropriate parameterisation, PAUCs of similar size were measured.

Interestingly, in case of MS/MS data we did not observe for the vector length normalised dot-product any differences due to intensity transformation by taking the square root, logarithm or ranking. They all performed equally well. This was probably due to the fact that we were only able (using a dataset where spectra IDs were assigned by database searches) to determine if a CP performed worse or equal to the database search algorithms, but not the case if it would have performed better.

The PL length

We examined two ways of defining the length of the matched PLs, first by setting $N = 0$ in Equation 12 or second to a user defined value N ($N = 250$ in case of PMF data and $N = 400$ in case of MS/MS data) and augmenting the missing peak-pairs by peaks of zero intensity. The PL length significantly interacts with

the binary measures ($length \times measure$) as well as with the weight of non-matching peaks ($\theta \times length$) as can be seen by the high MSQ values in Table 4 bottom panel. The significant third order interaction $measure \times \theta \times length$ (Table 4 bottom panel: Final model) was observed because of a strong interactions of the factor PL length with the weighting of non-matching peaks (θ) for the symmetric binary measures (Huberts Γ and relative mutual information), while this interaction is not observed for the asymmetric binary measures (Gower coefficient and Fowlkes-Mallows statistics). The best combination for both symmetric binary measures was when $N = 250$ and $\theta = 0.5$ in case of PMF and MS/MS data.

For the intensity based dissimilarity measures, a strong interaction between the factor PL length and measure (Table 5 bottom panel: Final model row $length \times measure$) was observed for the Manhattan and the Euclidean distances. On the other hand, for combinations of measure and scale, which account for the influence of N , there was no such dependence on N , *i.e.* the Manhattan distance (l^p -norm with $p = 1$) computed for total ion count ($l = 1$ -norm) scaled intensities. We conclude that a combinations of measure and scale robust with respect to N should be used.

Differences between binary and intensity based dissimilarities

The dash-dotted line in Figure 5 indicates the maximal sensitivity-PAUC determined for the binary based CPs while the dashed line the maximal sensitivity-PAUC computed for the intensity based CPs. If high sensitivities at a high specificity was required, the intensity based CPs performed better then the binary based CPs. This is because it is very unlikely that samples from different sources will generate spectra where not only the peak masses, but also the peak intensities are similar.

If high specificity at a high sensitivity was required, the order reversed (not shown). The reason is, that if binary coding was used and intensities were discarded, it is unlikely that two spectra from the same sample would generate a high dissimilarity. Large dissimilarities are possible if intensities with large errors are given, as is the case for mass spectra.

Weighting of mass measurement accuracy, computing the non-crossing matching and weighting of non-matching peaks.

The variance explained by the factor non-crossing matching (cf Methods - Finding the matching peaks) and weight of matching peaks (cf Methods - Weighting the missing mass measurement accuracy), was practically zero in case of the PMF data. In case of the MS/MS data the variance explained by the factor weight of matching peaks and non-crossing matching was small, but not zero compared with other factors.

Interactions between these two factors as well as with other factors *i.e.* weighting of non-matching peaks, were observed.

Weighting of mass accuracy decreases the PAUC obtained by a CP in case of measures taking non-matching peaks into account (e.g. Euclidean distance, Huberts Γ). This is because weighting of mass accuracy decreases the weight of matching peak-pairs but does not affect non-matching peaks. For example, in the case of the Euclidean distance, $\sqrt{w(a-b)^2} = \sqrt{wa^2 - w2ab + wb^2}$, weighting of match accuracy decreases the term ab for matching peaks. For non-matching peaks the term ab equals zero. Non-matching peaks contribute to the term a^2 and b^2 and are not influenced by the weighting ($w = 1$) of match accuracy. From Figures 2 A and B we learned that matching peaks have higher discriminating power than non-matching peaks. So, decreasing exclusively the weight of matching peaks decreases the discriminating power of a CP. To compensate for the effect described, we have introduced the weighting of non-matching peaks by θ .

Nevertheless, there are reasons to use both of these procedures if applying the CPs on MS/MS data.

Non-crossing matching corrects for errors of the peak extraction procedure. To decrease the influence of random matches on the dissimilarity, which more frequently occur in MS/MS PLs weighting can be used. We conclude that weighting of mass accuracy and non-crossing matching should be preferred if computing the dissimilarities for MS/MS data.

Conclusions

We reviewed here a large group of dissimilarity measures and examined how they can be adjusted for comparison of PLs. In addition to factors transformation and scaling of peak intensities, we included parameters such as weighting of mass measurement accuracy or computing the non-crossing matching. A new parameter weight of non-matching peaks (θ) was introduced into the computation of distance measures. Finally, we have studied how these factors influence the performance of the CP in relation to the factors scaling and measure. A possible strategy for obtaining the optimum CP would have been to choose the one with the largest related partial area under the curve. However, this method might only have a small generalisation property as it depends strongly on the specific data set chosen. Therefore, we performed a statistical analysis of variance using the factorial structure given in Table 3.

The use of ANOVA techniques in our study is justified if one views the data as fixed and the specific strategies in cluster analysis as factor levels or combinations of factor levels. However, due to multi-modality, our data could not be transformed to approximate normality. Thus we could not calculate

F -ratios and related statistical tests of significance. To test the generalisability of our results we used a second independent data set of MS/MS measurements. The results were similar to those obtained for PMF values, which provides evidence that they might be of general interest.

The best performing measures for high specificities were in case of the PMF dataset the Fowlkes-Mallows statistics with the weight a double weight of non-matching peaks. If high sensitivities are to be obtained the dot product measure computed on log transformed and vector length scaled intensities discriminates efficiently between *within* and *between* cluster pairs. To classify MS/MS data, the use of the dot product measure computed on log transformed and vector length scaled data can be recommended to obtain high sensitivities as well as high specificities. The sensitivity and specificity PAUCs for all CPs are included as datasets into the **BioConductor** [66,67] package **msbase**, which implements all measures discussed and was used to conduct the study.

The CP is a computer model of cluster affiliation, which for a given input of two PLs and various control variables such as weight of non-matching peaks, generates a single output variable further used to classify the PL pair. Computing the dissimilarities for two PLs with a set of input variables (levels of factors) represents a single computer experiment. To conduct the analysis presented here a total number of $\approx 4,000,000,000$ single computer experiments were performed (cf Methods Computation). In order to reduce the number of required computations and explore a wider range of factors and levels it might be beneficial to apply methods to design and analyse computer experiments [48].

Because the identities of the samples were determined by database search algorithms, we were not able to reveal if any of these measures had higher sensitivities and specificities than the DB search algorithm. It would be interesting to repeat this experiments with spectra, of which “true” identity was determined, using other methods than comparison of experimental with theoretical spectra (as the sequence database search engines do). If such data were available, a detailed analysis how the measures like the sum of agreeing intensities, the Manhattan and the Euclidean distances perform in comparison with the dot-product measure can be examined. A further direction would be to combine the pairwise PL comparison methods presented here with methods that model other PL properties in addition *i.e.* peptide fragmentation patterns [17].

Methods

Data sets and pre-processing

PMF-data

The PMF data employed in this study (4532 PMF MS spectra) was derived from three different and independent proteome studies. One set contains 1193 PMF MS spectra from bacterial (*Rhodopirellula baltica*) samples (unpublished data). These samples were measured on a Bruker Reflex III reflectron MALDI-TOF MS (Bruker Daltonics, Bremen, Germany). Another set, which contains 1539 PMF spectra from mouse (*Mus musculus*) brain tissue samples (unpublished data) was measured on a Bruker Ultraflex reflectron MALDI-TOF MS (Bruker Daltonics, Bremen, Germany), while the final set, which was measured on an Bruker Autoflex reflectron MALDI-TOF MS (Bruker Daltonics, Bremen, Germany), contains 1800 PMF MS spectra from plant tissue (*Arabidopsis thaliana*) [24, 49]. All PMF MS spectra were derived from tryptic protein digests of individually excised protein spots. For this purpose the whole tissue/cell protein extracts of the former mentioned organisms were separated by two-dimensional (2D) gel electrophoresis [49] and visualised with MS compatible Coomassie brilliant blue G250 [24]. The MALDI-TOF MS analysis was performed using delayed ion extraction and employing the MALDI AnchorChipTM targets (Bruker Daltonics, Bremen, Germany). For positively charged ions in the m/z range of 700 – 4,500 m/z were recorded. Subsequently, the monoisotopic masses of the measured peptides were detected by the SNAP algorithm of the XTOF spectrum analysis software (Bruker Daltonics, Bremen, Germany). The sum of the detected monoisotopic masses constitute the raw peak-list (PL), which were calibrated to a mass accuracy of 0.05Da (or higher) by the in-house developed software `mscalib` [67]. Moreover, the `mscalib` software was used to filter the PLs for irregular peaks that did not follow the general peptide mass rule [12, 51]. Additional background peaks (peaks occurring in more than 8% of the spectra [15]) were removed from the PLs. The obtained processed PLs were then used for the protein database searches with the `Mascot` search software (Version 1.8.1) [39] employing a mass accuracy of $\pm 0.1Da$, setting methionine oxidation as a variable and carbamidomethylation of cysteine residues as fixed modification, and allowing a maximum of 1 missed proteolytic cleavage site. Samples with multiple content/identification were removed from the data. Multiple content of samples was determined by removal of all peaks, matching the highest significant hit in the first search and re-submission of the remaining peaks to a new database search.

MS/MS data

To evaluate the distances for the MS/MS data, 70 clusters (spectra assigned to one ID) were randomly chosen (5 replicates obtained) from a large data-set of identified yeast spectra [23]. The protein extraction, sample preparation, measurement and identification was performed as described by Wagner et al. [52]. The analysed MS/MS spectra were recorded on an ESI Ion Trap mass spectrometer (LCQ DECA, Thermo Electron) with the following instrument settings: spray voltage: 1.5 kV; data dependent scanning with one full MS spectrum is followed by four independent MS/MS spectra of the four most intensive ions; minimum signal intensity for a peptide to be selected for fragmentation set to 10^6 ion counts. These selected and fragmented ions were then excluded from further fragmentation events for 1 minute to prevent repeated MS/MS spectra of identical peptides. The collision energy for the peptide fragmentation was automatically set by the instrument, which was controlled by the Xcalibur software (Version 1.2, Thermo Electron). The post acquisition processing was performed with the Bioworks browser package (Thermo Electron). The resulting PLs were automatically stored and assigned to peptide sequences in the yeast protein database [53], by using the **Sequest** database search algorithm (Version 27) [8, 54]. The search parameters employed for the database searches were as follows: a) none or one of the four proteases, as defined by Sickmann et al. [23]; b) mass type: mono isotopic (parent ion and fragment ion) c) amino acid modifications: carbamidomethylated cysteine residues +57Da and oxidation on methionine residues +16Da, while missed cleavage sites (maximum allowed): 1 missed. We considered spectra identified if they had an $X_{corr} > 2$ and an ion coverage of 20%.

Finding the matching peaks

We considered two peaks x and y from different PLs X, Y to *match* with an accuracy a if $|x - y| < a$ (absolute error) or $\frac{|x-y|}{(x+y)/2} \cdot 10^6 < a$ (relative error in part per million (ppm)). Cases where more than one peak in Y match a peak x (Figure 7, case A), were resolved by computing a non-crossing matching of the PLs. A non-crossing matching a *maximum trace* [55, 56], which can be computed in time $O(n \log n)$, where n is the number of peak matches. In our case, we resorted to a simple maximum similarity alignment, which could be banded to improve its $O(n^2)$ time complexity. The optimal trace of two sorted mass lists of matching peaks was found by dynamic programming. Let *qual* be a measure of the goodness of the match of two peaks *i.e.*

$$qual_{\text{abs}} = \max\{0, a - |x - y|\},$$

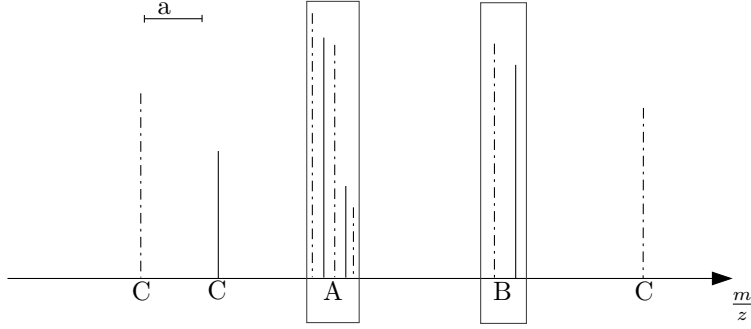


Figure 7: Stick spectrum of two peak-lists X (black lines) and Y (black dot dashed lines). Upper left corner – accuracy of the mass measurement a . A – ambiguous match of five peaks. B – unambiguous match of two peaks. C – peaks not matching.

where x and y are masses of peaks from two distinct PLs. Then our goal was to maximise the overall quality of all matched peaks. We recorded in the matrix $M_{i,j}$ the best possible assignment of the first i peaks in the first list and the first j peaks in the second list. Hence $M_{i,0} = M_{j,0} = 0$ for all i, j , and it is easy to see that the following recurrence can be used to derive the overall best assignment:

$$M_{i,j} = \max \begin{cases} M_{i-1,j-1} + qual \\ M_{i-1,j} \\ M_{i,j-1}. \end{cases} \quad (1)$$

In this way we could find a non-crossing matching, which minimised the overall errors and unambiguously assigned a peak x to a peak y .

Weighting the missing mass measurement accuracy

For computation of the dissimilarities we used the weighting of the mass measurement accuracy [46,47] and the alignment of PL by linear regression [12,57]. To model the accuracy of a given match either we weighted the peak intensities in the matching pairs (intensity based measures) or calculate the weight of the match (binary measure) by a triangular function w :

$$w_i^{(xy)} = \begin{cases} 1 - \frac{|(x-y)|}{a} & \text{if } |(x-y)| < a \\ 0 & \text{if } |(x-y)| \geq a. \end{cases} \quad (2)$$

where a is the maximum displacement evaluated, and x and y are peak masses. If the mass difference $|x - y|$ of two matching peaks increases then the significance of the match is reduced. Before computing the weights, we minimised the overall error of the matching masses by adjusting the two PLs using linear regression.

Non matching peak pairs

Peaks detected in one sample not occurring in the other one (Figure 7, case C), were included in the computation of the dissimilarities. The second PL was augmented with a peak of zero intensity, at the mass of the not-matching peak. In addition, the significance of such a peak pair (and peak intensities) could be weighted with the factor θ . In this study we examined three values of θ , namely 0.5, 1 and 2.

Intensity transformation

The purpose of transformation is to stabilise the variance of the data prior to the statistical analysis. We transformed the peak intensities by taking the square root, as suggested by Tabb et al. [22] or the logarithm. Furthermore, using a non-parametric approach, we replaced the intensities by their ranks within the spectrum [29].

Scaling

The purpose of scaling is to allow the comparison of PLs with different intensity values *i.e.* due to different scale of the detector used or due to different amount of sample. Since intensities in different PLs could have different intensity ranges, we used standard scaling procedures to account for this bias.

- Total ion current count normalisation [37, 38] is defined as :

$$I'_i = \frac{I_i}{\sum_{i=1}^N I_i} , \quad (3)$$

where I_i is the intensity of the peak i in the PL of length N . Here, the intensities are divided by the sum of all intensities, so that after scaling the sum of the intensities in each PL equals one ($\sum_i^n I' = 1$). The total ion count is better known as the l_1 - norm since $I_i > 0 \forall i$

- Vector length normalisation is defined as

$$I'_i = \frac{I_i}{\sqrt{\sum_{i=1}^N I_i^2}} . \quad (4)$$

Here, the peak intensities are divided by the $l = 2$ -norm of the intensity vector, which causes that the Euclidean length of the vector equals one ($\sum_i^n I^2 = 1$).

- Root mean square normalisation is defined as

$$I'_i = \frac{I_i}{\sqrt{\frac{1}{N-1} \sum_{i=1}^N I_i^2}} . \quad (5)$$

Here, the intensities are divided by their root-mean-square [58].

- z -score normalisation is defined as

$$I'_i = \frac{I_i - \bar{I}}{S_N(I)} , \quad (6)$$

where I_i, i, N defined as above, \bar{I} denotes the average intensity of a PL and

$S_N = \sqrt{\frac{1}{N-1} \sum_{i=1}^N (I_i - \bar{I})^2}$. Here, two scaling steps are performed, centring and subsequent division by the standard deviation. This causes that for each scaled PL the average of intensities is zero and the standard deviation equals one.

The scaling is preferred if intensities and variance in an arbitrary sample are much higher than in the other samples, which will determine the outcome of the PL comparisons. Data transformation was applied before the PL matching, whereas data scaling was performed for already matched PLs.

Measures based on peak intensities and intensity ranks.

In our study, we investigated several pairwise similarity measures to compare two PLs. These measures are either measures of similarity (such as covariance) or measures of distance (such as the l^p metrics). In order to make both classes of measures comparable, we transformed each similarity measure into an appropriate dissimilarity measure. Moreover, we introduced the factor w_i to weight missing mass measurement accuracy (cf Methods - Weighting the missing mass measurement accuracy) and non-matching peaks (cf Methods - Non matching peak pairs).

The dot-product of two vectors is defined

$$DP(I^x, I^y) = \sum_{i=1}^N w_i^{xy} (I_i^x)(I_i^y) , \quad (7)$$

where I^X and I^Y are the intensity vectors of two matched PLs (cf Methods - Finding the matching peaks) of length N , and w_i^{xy} is defined by Equation 2 for matching peaks and equals θ for non-matching peaks. In case of sum-mean-square, total ion count and vector length scaling the product of not matching peak-pairs is zero and therefore this measure is independent of θ . If the intensities of the matched PLs are z -score scaled, the outcome will depend on the value of θ . Furthermore, augmenting the PLs by zero pairs in order to increase their length will increase DP for z -score and root-mean-square scaled data. The most prominent representative of this family is the spectral angle (dot-product of vector length normalised data). It has a geometrical interpretation. It is equal to the *cosine* of the angle enclosed by the two vectors.

Covariance. The covariance is a measure of dependency between random variables I^x and I^y [59] and is defined

$$\text{Cov}(I^x, I^y) = \frac{\sum_{i=1}^n w_i^{xy} (I_i^x - \bar{I}^x)(I_i^y - \bar{I}^y)}{n - 1}, \quad (8)$$

where I^x, I^y, N, w_i^{xy} is defined as above.

The best known representative of this family of measures is the Pearson correlation, which is obtained if we compute the covariance of z -score scaled intensity vectors.

Metric-based measures. The Euclidean and Manhattan distances belong to the family of l^p metrics and can be expressed using equation

$$D(I^x, I^y) = \left(\sum_{i=1}^N w_i^{xy} |I_i^x - I_i^y|^p \right)^{1/p}. \quad (9)$$

In case of the Euclidean distance $p = 2$, and for the Manhattan distance $p = 1$. The Euclidean distance penalises large intensity differences stronger than the Manhattan distance. The outcome of this measure will change due to different sample-wise scaling of the intensities. In case of the z -score scaling the outcome will depend on the user defined PL length N (Equation 12).

Similarity index and Canberra distance. The Similarity index [35] and Canberra Distance [60] measure the relative distance and can be expressed by equation

$$D(I^x, I^y) = \left(\frac{\sum_{i=1}^N w_i^{xy} \left| \frac{I_i^x - I_i^y}{I_i^x + I_i^y} \right|^p}{\sum_{i=1}^N w_i^{xy}} \right)^{1/p}. \quad (10)$$

Setting $p = 2$ yields the similarity index, while $p = 1$ results in the Canberra distance. Similarly, as in case of the l^p metrics, the similarity index with $p = 2$ will be more influenced by large intensity differences than the Canberra distance.

If the term $x + y$ in the denominator equals zero due to $x = -y$, with $x \neq 0 \vee y \neq 0$, infinity $+\infty$ is returned [58].

Sum of agreeing intensities. The sum of agreeing intensities is defined by equation

$$SOAI(I^x, I^y) = 1 - \frac{\sum_{i=1}^n w_i^{xy} \max\{(I_i^x + I_i^y)/2 - |I_i^x - I_i^y|, 0\}}{\sum_{i=1}^n w_i^{xy} (I_i^x + I_i^y)/2}. \quad (11)$$

It shares the property with the similarity index and the Canberra distance, where each pair of matching peaks will contribute to the final score a proportion in the range of $[0, 1/n]$. The sum of agreeing intensities however, puts more emphasis on the agreement of peak intensities. Peak pairs whose intensity differences are larger than their average intensity get the weight zero.

		object X		
		$x=0$	$x=1$	
object Y	$y=0$	M_{00}^{XY}	$\theta \cdot M_{10}^{XY}$	M_0^Y
	$y=1$	$\theta \cdot M_{01}^{XY}$	$M_{11}^{XY} = \sum_{i=0}^n w_i^{xy}$	M_1^Y
		M_0^X	M_1^X	M

Table 6: Modified contingency table. $M = \max\{N, c + \theta \cdot (M_{01}^{XY} + M_{10}^{XY}) + M_{11}^{XY}\}$ with N defined by the user and $c = 1$ in case of Hubert’s Gamma or $c = 0$ otherwise.

Binary measures

In contrast to the previous measures taking into account the intensities of the PLs, we investigated measures that only use qualitative information in the sense that they evaluate the number of matching and mismatching peaks of both PLs. Essentially these measures are numerical functions in the contingency Table 6 derived from both PLs. To include the weighting of missing accuracy by the w (see Equation 2) and the weighting of non-matching peaks by θ , we introduced a generalized version of the contingency table. All binary measures introduced below can be computed on the entries of the contingency Table 6. Peaks present in list X , but not in list Y , are denoted by M_{10}^{XY} , likewise present in Y , but not in X by M_{01}^{XY} . We multiplied the mismatches by θ to assign a variable weight. Therefore, M_{10}^{XY} , as well as M_{01}^{XY} was replaced by $\theta \cdot M_{10}^{XY}$ and $\theta \cdot M_{01}^{XY}$, respectively. To include the weighting of missing mass accuracy in computing the dissimilarities one can set $M_{11}^{XY} = \sum_{i=0}^n w_i^{xy}$, with w_i^{xy} is defined in Equation 2.

Our data are asymmetric in the sense that we can only evaluate existing peaks and do not count the absence of peaks in both PLs at a mass. Measures that utilize only this information are the Gower coefficient (18) or Fowlkes-Mallows statistics (19). Additionally, we were interested in the performance of measures that take into account the marginal M and hence the entry M_{00}^{XY} is required (Hubert’s Γ (20) or the relative mutual information (23)).

Since the PLs can have different length and the maximal PL length is undefined, we defined the entry M length of a matched PLs pairs as follows

$$M = \max\{N, c + \theta \cdot (M_{01}^{XY} + M_{10}^{XY}) + M_{11}^{XY}\}, \quad (12)$$

where N is an arbitrary user defined constant, and $c = 1$ in case of the Huberts Γ and $c = 0$ otherwise. By this definition, due to the use of the maximum function we avoided the case that M_{00}^{XY} becomes less than zero (see equation 13 for definition of M_{00}^{XY}). In this study we used two different values of N . We set $N = 0$ and the second value equal to twice the length of the longest PL in each dataset.

Given all entries of the modified contingency Table (6), the marginals could be computed by equations (13 – 17).

$$M_{00}^{XY} = M - (\theta \cdot M_{01}^{XY} + \theta \cdot M_{10}^{XY} + \sum w_i^{xy}) , \quad (13)$$

$$M_1^X = \theta \cdot M_{10}^{XY} + \sum w_i^{xy} , \quad (14)$$

$$M_1^Y = \theta \cdot M_{01}^{XY} + \sum w_i^{xy} , \quad (15)$$

$$M_0^X = \theta \cdot M_{01}^{XY} + M_{00}^{XY} , \text{ and} \quad (16)$$

$$M_0^Y = \theta \cdot M_{10}^{XY} + M_{00}^{XY} . \quad (17)$$

Jaccard/Gower Coefficient. The matching peak count is the dot-product of the two PLs and counts the number of matching peaks (M_{11}^{XY}). Since PLs have different numbers of non-zero elements, this dot product must be normalised by the total counts. The Jaccard coefficient is a normalised version of the matching peak count, whose distance version is given by

$$G(X, Y) = \frac{M_{01}^{XY} + M_{10}^{XY}}{M_{01}^{XY} + M_{10}^{XY} + M_{11}^{XY}} . \quad (18)$$

A generalised version of the Jaccard coefficient in which M_{01}^{XY} and M_{10}^{XY} is weighted by a constant θ was introduced by Gower et al. [61].

Fowlkes-Mallows statistics. The Fowlkes-Mallows statistics [62] (introduced in the context of clustering validation by use of contingency tables) are the matching peak counts normalised by the geometric mean of the PLs lengths. The equation of the distance-like version is given by

$$FM(X, Y) = \frac{M_{11}^{XY}}{\sqrt{M_1^X \cdot M_1^Y}} . \quad (19)$$

Huberts Γ . Using binary signals we can transform the formula of the correlation coefficient such that it uses the values of the contingency table to obtain

$$HG(X, Y) = \frac{M \cdot M_{11}^{XY} - M_1^X \cdot M_1^Y}{\sqrt{M_0^X \cdot M_1^X \cdot M_0^Y \cdot M_1^Y}} . \quad (20)$$

We observed that the nominator was maximised if all signals were expressed equally. To avoid the fact that the denominator becomes zero (which is the case if M_0^X or $M_0^Y = 0$ and occurs if one PL is included in the other) we set $c = 1$ in equation (12).

Relative mutual information. We were additionally interested in the performance of information theoretic concepts. Given the two PLs, X and Y , the amount of information about PL X inherent in PL Y (and vice versa) is given by the *mutual information* (**H**) [63]

$$H(X; Y) = \sum_{i=0}^1 \sum_{j=0}^1 \frac{M_{ij}^{XY}}{M} \log_2 \left(\frac{M_{ij}^{XY} \cdot M}{M_i^X \cdot M_j^Y} \right). \quad (21)$$

To be able to use the mutual information as a similarity measure, so it could distinguish positive from negative correlation, we introduced the following scaling term [64]

$$\Delta = \begin{cases} -1 & \text{if } M_{11}^{XY} < (M_1^Y \cdot M_1^X)/M \\ 0 & \text{if } M_{11}^{XY} = (M_1^Y \cdot M_1^X)/M \\ 1 & \text{otherwise.} \end{cases}$$

Furthermore, we adjusted it for the information inherent in the individual PLs. The adjustment was done by the entropy of the individual PLs, which for a PL X is given by

$$H(X) = - \sum_{i=0}^1 \frac{M_i^X}{M} \log_2 \frac{M_i^X}{M}. \quad (22)$$

Thus, we defined the relative mutual information:

$$RH(X, Y) = \Delta \frac{2H(X; Y)}{H(X) + H(Y)}. \quad (23)$$

The $-RH$ is small if both PLs are similar and high if they differ.

Since the inequalities

$$H(X; Y) \geq 0 \text{ and } H(X; Y) \leq \min\{H(X), H(Y)\}$$

holds, this measure is bounded to the interval $[-1, 1]$. The relative mutual information has been introduced before [36], in the context of clustering gene expression data.

Diagnostic evaluation scores

The ROC curve is a graphical plot of the sensitivities versus the 1-specificities determined on the same value of the discriminatory variable. In order to evaluate the CP for their capability to group proteins according to their spectral distance, we used the partial area of interest under ROC curve (PAUC) [43]. The area was calculated using the trapezoidal integration rule.

Maximize sensitivity given specificity

Here, we asked which fraction of the samples would be recognised as being truly the same (sensitivity), if we allowed a given fraction of false assignments (1-specificity). We were interested in the sensitivities of the CP with any 1-specificity in the range of 0 – 10%. We computed the partial area under the curve for this range of 1-specificity values (sensitivity-PAUC).

Maximize specificity given sensitivity

Here, we asked which measure should be used to get as few as possible false assignments given a high sensitivity. A good CP will be the one which maximises the specificity for a given sensitivity. We were interested in the specificities given by any sensitivity in the range of 90 – 100%. Hence, we computed the partial area under the curve given if plotting specificity against the sensitivity (specificity-PAUC).

Analysis of variance

Aim of the statistical analysis was to evaluate different strategies of the CP with respect to the two PAUCs (cf Methods - Diagnostic evaluation scores). Each strategy of the CP was defined by the combination of seven factors as given in Table 3. The whole set of CPs shows a completely balanced factorial structure analogous to those used in analysis of variance (ANOVA) [45] with PAUCs as quantitative measurements. Thus, we presented the results by partitioning sums of squares for the different PAUCs into components attributable to the different factors and their interactions. We only presented raw and mean sums of squares divided by the respective degrees of freedom (number of factor levels - 1), but we did not calculate *F*-ratios or P-values for factors and interactions, because of the mentioned deviation from normality. The ANOVA identifies factors of the CP which do not change the PAUCs considerably and allows proposals for the optimum strategy of the CP. The PMF data set was the learning sample and the MS data set was the validation sample for the choice of the optimum CP.

Computation

All scores presented in the results section were computed for 75 clusters. The clusters were sampled from the datasets without replacement. For each cluster we randomly chose 2 – 20 (PMF-data) 2 – 7 (MS/MS) samples. This procedure was repeated 5 times and the average of the scores was computed. The CPs were computed with a mass measurement error of 1*Da* for the MS/MS data, and of 0.2*Da* for the PMF data. The computation of the CPs was performed using the in-house developed R [65] package `msbase`, which is

available on `BioConductor` [66,67]. The PAUC areas were computed using in-house developed `R` functions. Other `R` packages provide a huge variety of statistical tools for further analysis of the dissimilarities such as clustering algorithms and validation or multidimensional scaling methods [68].

Abbreviation

- PL - peak-list.
- CP - comparison process.
- ROC - receiver operating characteristic curve.
- PAUC - partial area under the curve.
- TP - true positive.
- FP - false positive.
- FN - false negative.
- TN - true negative.

Acknowledgements

We would like to thank Florian Markowetz, Colin Gillespie, Vanessa Hall and Stale Nygard for proofreading the manuscript. We thank the anonymous referees for helpful comments. Further thanks to the `R`-help mailing list [65] for continuous help while implementing the package `msbase` and with other `R` related questions. This project was funded by the National Genome Research Network (NGFN) of the German Ministry for Education and Research (BMBF), and the Max Planck Society.

Authors contributions

HL, JG, KR and RH gave initial input to the research.

WEW implemented the dissimilarities, evaluation framework, and designed the graphical figures.

WEW and PM planned and carried out the analysis.

WEW, ML, PM, KR, PG and JG wrote the manuscript.

AS provided the MS/MS data set.

JG and PG provided the PMF data set.

All authors contributed to the final version of the manuscript and approved it.

References

1. Fenyo D: **Identifying the proteome: software tools.** *Current Opinion in Biotechnology* 2000, **11**:391–395.
2. Griffin TJ, Aebersold R: **Advances in proteome analysis by mass spectrometry.** *J Biol Chem* 2001, **276**:45497–500.
3. Patterson SD: **Data analysis—the Achilles heel of proteomics.** *Nat Biotechnol* 2003, **21**(3):221–2.
4. Aebersold R, Mann M: **Mass spectrometry-based proteomics.** *Nature* 2003, **422**(6928):198–207.
5. Mann M, Hojrup P, Roepstorff P: **Use of mass spectrometric molecular weight information to identify proteins in sequence databases.** *Biol Mass Spectrom* 1993, **22**(6):338–345.
6. Johnson R, Martin S, Biemann K, Stults J, Watson J: **Novel Fragmentation Process of Peptides by Collision-Induced Decomposition in a Tandem Mass Spectrometer: Differentiation of Leucine and Isoleucine.** *Analytical Chemistry* 1987, **59**(21):2621–2625.
7. Smith RD, Loo JA, Edmonds CG, Barinaga CJ, Udseth HR: **New developments in biochemical mass spectrometry: electrospray ionization.** *Anal Chem* 1990, **62**(9):882–99.
8. Sadygow 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.
9. Gras R, Muller M, Gasteiger E, Gay S, Binz PA, Bienvenut W, Hoogland C, Sanchez JC, Bairoch A, Hochstrasser DF, Appel RD: **Improving protein identification from peptide mass fingerprinting through a parameterized multi-level scoring algorithm and an optimized peak detection.** *Electrophoresis* 1999, **20**(18):3535–3550. [(eng)].
10. Strittmatter EF, Rodriguez N, Smith RD: **High mass measurement accuracy determination for proteomics using multivariate regression fitting: application to electrospray ionization time-of-flight mass spectrometry.** *Anal Chem* 2003, **75**(3):460–8.
11. Gentzel M, Kocher T, Ponnusamy S, Wilm M: **Preprocessing of tandem mass spectrometric data to support automatic protein identification.** *Proteomics* 2003, **3**(8):1597–610.
12. Wool A, Smilansky Z: **Precalibration of matrix-assisted laser desorption/ionization-time of flight spectra for peptide mass fingerprinting.** *Proteomics* 2002, **2**(10):1365–1373.
13. Gobom J, Mueller M, Egelhofer V, Theiss D, Lehrach H, Nordhoff E: **A calibration method that simplifies and improves accurate determination of peptide molecular masses by MALDI-TOF MS.** *Anal Chem* 2002, **74**(15):3915–3923. [(eng)].
14. Levander F, Rognvaldsson T, Samuelsson J, James P: **Automated methods for improved protein identification by peptide mass fingerprinting.** *Proteomics* 2004, **4**(9):2594–601.
15. Chamrad DC, Koerting G, Gobom J, Thiele H, Klose J, Meyer HE, Blueggel M: **Interpretation of mass spectrometry data for high-throughput proteomics.** *Anal Bioanal Chem* 2003, **376**(7):1014–22.
16. Moore RE, Young MK, Lee TD: **Method for screening peptide fragment ion mass spectra prior to database searching.** *J Am Soc Mass Spectrom* 2000, **11**(5):422–6.
17. Sun W, Li F, Wang J, Zheng D, Gao Y: **AMASS: Software for Automatically Validating the Quality of MS/MS Spectrum from SEQUEST Results.** *Mol Cell Proteomics* 2004, **3**(12):1194–9.
18. Anderson DC, Li W, Payan DG, Noble WS: **A new algorithm for the evaluation of shotgun peptide sequencing in proteomics: support vector machine classification of peptide MS/MS spectra and SEQUEST scores.** *J Proteome Res* 2003, **2**(2):137–46.
19. Keller A, Nesvizhskii AI, Kolker E, Aebersold R: **Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search.** *Anal Chem* 2002, **74**(20):5383–92.
20. Yates JR, Morgan SF, Gatlin CL, Griffin PR, Eng JK: **Method To Compare Collision-Induced Dissociation Spectra of Peptides: Potential for Library Searching and Subtractive Analysis.** *Anal. Chem.* 1998, **70**:3557–3565.
21. Beer I, Barnea E, Ziv T, Admon A: **Improving large-scale proteomics by clustering of mass spectrometry data.** *Proteomics* 2004, **4**(4):950–60.
22. Tabb DL, MacCoss MJ, Wu CC, Anderson SD, Yates JR: **Similarity among tandem mass spectra from proteomic experiments: detection, significance, and utility.** *Anal Chem* 2003, **75**(10):2470–7.

23. Sickmann A, Reinders J, Wagner Y, Joppich C, Zahedi R, Meyer HE, Schonfisch B, Perschil I, Chacinska A, Guiard B, Rehling P, Pfanner N, Meisinger C: **The proteome of *Saccharomyces cerevisiae* mitochondria.** *Proc Natl Acad Sci U S A* 2003, **100**(23):13207–12.
24. Giavalisco P, Nordhoff E, Kreitler T, Kloeppel KD, Lehrach H, Klose J, Gobom J: **Proteome Analysis of *Arabidopsis Thaliana* by 2-D Electrophoresis and Matrix Assisted Laser Desorption/Ionization Time of Flight Mass Spectrometry.** [To appear in *Proteomics*].
25. Pevzner PA, Dancik V, Tang CL: **Mutation-Tolerant Protein Identification by Mass Spectrometry.** *Journal of Computational Biology* 2000, **7**(6):777–787.
26. Mann M, Wilm M: **Error-tolerant identification of peptides in sequence databases by peptide sequence tags.** *Anal. Chem.* 1994, **66**:4390–4399.
27. Tabb DL, Saraf A, Yates JRr: **GutenTag: high-throughput sequence tagging via an empirically derived fragmentation model.** *Anal Chem* 2003, **75**(23):6415–21.
28. Clauser KR, Baker P, Burlingame AL: **Role of accurate mass measurement (+/- 10 ppm) in protein identification strategies employing MS or MS/MS and database searching.** *Anal Chem* 1999, **71**(14):2871–82.
29. Svetnik V, Liaw AI: **Detecting Novel Samples in Mass Spectral Data: A Clustering Approach.** In *Proceedings of the 33rd Symposium on the Interface*. Edited by Wegman E, Braverman A, Goodman A, Smyth P 2001:321–328.
30. An Z, Harris G, Zink D, Giacobbe R, Lu P, Sangari R, Bills G, Svetnik V, Gunter B, Liaw A, Masurekar P, Liesch J, Gould S, Strohl W: **Expression of Cosmid-Size DNA of Slow-Growing Fungi in *Aspergillus Nidulans* for Secondary Metabolite Screening.** In *Handbook of Industrial Mycology*. Edited by An Z, New York: Marcel Dekker 2003:167–187.
31. Li J, Zhang Z, Rosenzweig J, Yang Y, Chan D: **Proteomics and bioinformatics approaches for identification of serum biomarkers to detect breast cancer.** *Clinical Chemistry* 2002, **48**(8):1296–1304.
32. Jeffries N: **Performance of a genetic algorithm for mass spectrometry proteomics.** *BMC Bioinformatics* 2004, **5**:180, [<http://www.biomedcentral.com/1471-2105/5/180>].
33. Adam B, Qu Y, Davis J, Ward M, Clements M, Cazares L, Semmes O, Schellhammer P, Yasui Y, Feng Z, Wright G: **Serum protein fingerprinting coupled with a pattern-matching algorithm distinguishes prostate cancer from benign prostate hyperplasia and healthy men.** *Cancer Research* 2002, **62**:3609–3614.
34. Petricoin E, Ardekani A, Hitt B, Levine P, Fusaro V, Steinberg S, Mills G, Simone C, Fishman D, Kohn E, Liotta L: **Use of proteomic patterns in serum to identify ovarian cancer.** *Lancet* 2002, **359**:572–577.
35. Wan KX, Vidavsky I, Gross ML: **Comparing similar spectra: from similarity index to spectral contrast angle.** *J Am Soc Mass Spectrom* 2002, **13**:85–88.
36. Herwig R, Poustka AJ, Müller C, Lehrach H, O'Brien J: **Large-scale Clustering of cDNA Fingerprinting Data.** *Genome Res.* 1999, **9**:1093–1105.
37. Alfassi ZB: **On the normalization of a mass spectrum for comparison of two spectra.** *J Am Soc Mass Spectrom* 2004, **15**(3):385–387.
38. Rasmussen GT, Isenhour TL: **The Evaluation of Mass Spectral Search Algorithms.** *Journal of Chemical Information and Computer Sciences* 1979, **19**(3):179 – 186.
39. Perkins DN, Pappin DJ, Creasy DM, Cottrell JS: **Probability-based protein identification by searching sequence databases using mass spectrometry data.** *Electrophoresis* 1999, **20**(18):3551–3567.
40. Tabb DL, Eng JK, Yates JR: **Protein Identification by SEQUEST** 2001.
41. Ross S: *A First Course In Probability*. Prentice Hall 2003.
42. Schmidt F, Schmid M, Jungblut PR, Mattow J, Facius A, Pleissner KP: **Iterative data analysis is the key for exhaustive analysis of peptide mass fingerprints from proteins separated by two-dimensional electrophoresis.** *J Am Soc Mass Spectrom* 2003, **14**(9):943–56.
43. Zhou XH, McClish DK, Obuchowski NA: *Statistical Methods in Diagnostic Medicine*. Wiley 2002.

44. Bland JM, Altman DG: **Measuring agreement in method comparison studies.** *Stat Methods Med Res.* 1999, **8**(2):135–60.
45. Fox J: *Applied Regression Analysis, Linear Models, and Related Methods.* Sage Publications 1997.
46. Zhang N, Aebersold R, Schwikowski B: **ProbID: A probabilistic algorithm to identify peptides through sequence database searching using tandem mass spectral data.** *Proteomics* 2002, **2**:1406–1412.
47. Hageman JA, Wehrens R, Gelder RD, Buydens LMC: **Powder Pattern Indexing Using the Weighted Crosscorrelation and Genetic Algorithms.** *Journal of Computational Chemistry* 2003, **24**(9):1043–1051.
48. Santner T, Williams B, Notz W: *The Design and Analysis of Computer Experiments.* Springer Series in Statistics, Springer Verlag New York 2003.
49. Klose J, Kobalz U: **Two-dimensional electrophoresis of proteins: an updated protocol and implications for a functional analysis of the genome.** *Electrophoresis* 1995, **16**(6):1034–59.
50. **R for Proteomics**[<http://r4proteomics.sourceforge.net>].
51. Gay S, Binz PA, Hochstrasser DF, Appel RD: **Modeling peptide mass fingerprinting data using the atomic composition of peptides.** *Electrophoresis* 1999, **20**(18):3527–3534.
52. Wagner Y, Sickmann A, Meyer HE, Daum G: **Multidimensional nano-HPLC for analysis of protein complexes.** *J Am Soc Mass Spectrom* 2003, **14**(9):1003–11.
53. Issel-Tarver L, Christie KR, Dolinski K, Andrada R, Balakrishnan R, Ball CA, Binkley G, Dong S, Dwight SS, Fisk DG, Harris M, Schroeder M, Sethuraman A, Tse K, Weng S, Botstein D, Cherry JM: **Saccharomyces Genome Database.** *Methods Enzymol* 2002, **350**:329–46.
54. Yates JR, Eng JK, McCormack AL, Schieltz D: **Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database.** *Anal Chem* 1995, **67**(8):1426–36.
55. Kececioğlu J: **The maximum weight trace problem in multiple sequence alignment.** In *Proceedings of the 4th Symposium on Combinatorial Pattern Matching (CPM 93)*, no. 684 in Lecture Notes in Computer Science, Springer 1993:106–119.
56. Kececioğlu JD, Lenhof HP, Mehlhorn K, Mutzel P, Reinert K, Vingron M: **A Polyhedral Approach to Sequence Alignment Problems.** *Discrete Applied Mathematics* 2000, **104**:143–186.
57. Egelhofer V, Gobom J, Seitz H, Giavalisco P, Lehrach H, Nordhoff E: **Protein identification by MALDI-TOF-MS peptide mapping: A new strategy.** *Analytical Chemistry* 2002, **74**(8):1760–1771.
58. Becker RA, Chambers JM, Wilks AR: *The New S Language.* Wadsworth & Brooks/Cole 1988.
59. Härdle W, Simar L: *Applied Multivariate Statistical Analysis.* Springer, Heidelberg 2003.
60. Lance GN, Williams WT: **Mixed-Data Classificatory Programs I - Agglomerative Systems.** *Australian Computer Journal* 1967, **1**:15–20.
61. Gower JC, Legendre P: **Metric and Euclidean Properties of Dissimilarity Coefficients.** *Journal of classification.* 1986, **3**:5–48.
62. Fowlkes EB, Mallows CL: **A method for comparing two hierarchical clusterings.** *J. Am. Stat. Assoc.* 1983, **78**:553–569.
63. Cover TM, Thomas J: *Elements of Information Theory.* New York: J.Wiley and Sons 1991.
64. Herwig R (Ed): *Large-scale information theoretic clustering with application to the analysis of genetic fingerprinting data (PhD).* Berlin: Logos 2001.
65. R Development Core Team: **R: A language and environment for statistical computing.** R Foundation for Statistical Computing, Vienna, Austria 2004, [<http://www.r-project.org>]. [ISBN 3-900051-00-3].
66. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Li FLC, Maechler M, Rossini AJ, Sawitzki G, Smith C, Smyth G, Tierney L, Yang JYH, Zhang J: **Bioconductor: Open software development for computational biology and bioinformatics.** *Genome Biology* 2004, **5**:R80, [<http://genomebiology.com/2004/5/10/R80>].
67. **Bioconductor - open source software for bioinformatics**[<http://www.bioconductor.org>].
68. **Comprehensive R Archive Network**[<http://cran.r-project.org>].