

RESEARCH ARTICLE

Open Access



# DISMS2: A flexible algorithm for direct proteome-wide distance calculation of LC-MS/MS runs

Vera Rieder<sup>1</sup>, Bernhard Blank-Landeshammer<sup>2</sup>, Marleen Stuhr<sup>3</sup>, Tilman Schell<sup>4</sup>, Karsten Biß<sup>2</sup>, Laxmikanth Kollipara<sup>2</sup>, Achim Meyer<sup>3</sup>, Markus Pfenninger<sup>4,5</sup>, Hildegard Westphal<sup>3</sup>, Albert Sickmann<sup>2,6,7</sup> and Jörg Rahnenführer<sup>1\*</sup> 

## Abstract

**Background:** The classification of samples on a molecular level has manifold applications, from patient classification regarding cancer treatment to phylogenetics for identifying evolutionary relationships between species. Modern methods employ the alignment of DNA or amino acid sequences, mostly not genome-wide but only on selected parts of the genome. Recently proteomics-based approaches have become popular. An established method for the identification of peptides and proteins is liquid chromatography-tandem mass spectrometry (LC-MS/MS). First, protein sequences from MS/MS spectra are identified by means of database searches, given samples with known genome-wide sequence information, then sequence based methods are applied. Alternatively, de novo peptide sequencing algorithms annotate MS/MS spectra and deduce peptide/protein information without a database. A newer approach independent of additional information is to directly compare unidentified tandem mass spectra. The challenge then is to compute the distance between pairwise MS/MS runs consisting of thousands of spectra.

**Methods:** We present DISMS2, a new algorithm to calculate proteome-wide distances directly from MS/MS data, extending the algorithm compareMS2, an approach that also uses a spectral comparison pipeline.

**Results:** Our new more flexible algorithm, DISMS2, allows for the choice of the spectrum distance measure and includes different spectra preprocessing and filtering steps that can be tailored to specific situations by parameter optimization.

**Conclusions:** DISMS2 performs well for samples from species with and without database annotation and thus has clear advantages over methods that are purely based on database search.

**Keywords:** Proteomics, LC-MS/MS, Mass spectrometry, Comparison of MS/MS spectra, Peptide identification, Distance of LC-MS/MS runs

## Background

In recent years proteomics has become of great interest in biochemistry. New methods and technologies are constantly being developed [1, 2]. Qualitative and quantitative proteomics strategies are particularly useful to analyze samples measured under different conditions or samples from different phenotypes. Another application was presented by Palmblad and Deelder [3] who reconstructed

the unique correct phylogenetic tree for the great apes and other primates based solely on proteome-wide measurements. Furthermore, Yilmaz et al. [4] have recently proposed a pipeline for differential proteomics in unsequenced species.

Most of the workflows in proteomics nowadays are based on mass spectrometry [5–7], replacing 2D gel electrophoresis. A great variety of instruments is being used and refined. Dealing with the high complexity of protein or peptide samples, liquid chromatography as separation technique is often combined with mass spectrometry. Tandem mass spectrometry, MS/MS, involves at least two

\*Correspondence: rahnenfuehrer@statistik.tu-dortmund.de  
<sup>1</sup>Department of Statistics, TU Dortmund University, Dortmund, Germany  
Full list of author information is available at the end of the article

stages of mass analysis and an intermediate fragmentation step. MS/MS spectra comprise of detected intensities of occurring masses corresponding to peptide fragments.

The identification of peptides and proteins using MS and MS/MS data is typically performed by database-dependent search algorithms, e.g., with Mascot [8]. Peptide sequences are verified by large and continuously updated databases that are derived from genome sequencing projects. These databases are usually well curated and often manually annotated. However, novel peptides nonexistent in databases cannot be identified with this approach. De novo peptide sequencing algorithms [6, 9, 10] are independent of database searches, but robust error estimation strategies are still lacking.

By omitting the peptide identification step mass spectra from different samples can be directly compared. Palmblad and Deelder [3] introduced a basic algorithm differentiating between blood samples. Two components are required for comparing samples on a proteome-wide scale. First, appropriate distance measures for mass spectra are needed [6, 11, 12]. Second, the information of thousands of spectra has to be aggregated to compute the global distance between pairwise LC-MS/MS runs.

Our main goal is to provide a general method for the comparison of different samples using data of LC-MS/MS runs. As a proof of concept, our new algorithm DISMS2 is applied to real data of LC-MS/MS runs. This includes both known species with established protein databases as well as two organisms with no prior comprehensive genomics and proteomics studies, namely *Radix* and *Amphistegina*, for which only direct spectra comparisons are feasible.

The resulting distances of DISMS2 are compared to a standard database search with Mascot evaluating the competitiveness of our flexible approach.

## Methods

We introduce the new flexible DISMS2 algorithm that calculates explicit distances between pairs of LC-MS/MS runs. First, a collection of 27 MS/MS runs from different species is presented. Second, a conventional Mascot database search as competitive method is explained. Third, an overview of appropriate distance measures between single spectra is given. Then the algorithm DISMS2 is explained in detail, and finally an approach for parameter optimization for DISMS2 is presented.

### Samples and LC-MS/MS analysis

Proteolytic (tryptic) digests of five sequenced organisms, i.e. (i) human (*Homo sapiens*, H, HeLa cell line), (ii) mouse (*Mus musculus*, M, C2C12 cell line), (iii) yeast (*Saccharomyces cerevisiae*, Y), (iv) roundworm (*Caenorhabditis elegans*, C), and (v) fruit fly (*Drosophila melanogaster*, D) and of four organisms without sequenced genome, i.e. (vi)

fresh water snail *Radix* species: molecular operational taxonomic unit (MOTU) 2 (R2), 4 (R4) and (vii) foraminifera species *Amphistegina lessonii* (Al), *Amphistegina gibbosa* (Ag) were analyzed using an Ultimate 3000 nano RSLC system coupled to a Q Exactive HF mass spectrometer (both Thermo Scientific). Each sample was measured in triplicate (1  $\mu$ g each) resulting in a dataset of 27 MS/MS runs. The samples were analyzed in randomized order to minimize systematic errors. Detailed information about sample preparation and LC-MS/MS analyses is provided in a document about Material and Methods [see Additional file 1].

### Database search

MS data interpretation was conducted using Proteome Discoverer 1.4 (Thermo Scientific) and Mascot 2.4 (Matrix Science). Database searches of the five model organisms (i.e. human, mouse, yeast, roundworm and fruit fly) were performed in a target/decoy mode against their respective protein sequence (FASTA) databases [see Additional file 1]. Trypsin was selected as enzyme, and two missed cleavage sites were allowed. Carbamidomethylation of cysteine was set as fixed and oxidation of methionine was set as dynamic modifications. MS and MS/MS tolerances were set to 10 ppm and 0.02 Da respectively, and only peptide-to-spectrum matches (PSMs) with search engine rank 1 and a false discovery rate (FDR) < 1% (Percolator setting) were considered.

### Distance measures

For any MS/MS run  $i$  containing  $n_i$  MS2 spectra define spectrum  $S_{k_i}$  with rank  $k$  in run  $i$  as a set of two vectors  $\mathbf{x}_{k_i}$  and  $\mathbf{I}_{k_i}$  with length  $p_{k_i}$ :

$$S_{k_i} = \{\mathbf{x}_{k_i}, \mathbf{I}_{k_i}\} = \left\{ \left( x_{k_i,1}, \dots, x_{k_i,p_{k_i}} \right)', \left( I_{k_i,1}, \dots, I_{k_i,p_{k_i}} \right)' \right\}$$

The  $m/z$  (mass-to-charge) ratios  $\mathbf{x}_{k_i}$  are sorted in ascending order, and corresponding peak intensities are labeled with  $\mathbf{I}_{k_i}$ .

According to the resolution of the experiment the range of  $m/z$  values can be subdivided into small intervals so that every peak can be assigned to exactly one interval. Then an alternative definition of spectrum  $S_{k_i}$  is a vector  $\tilde{\mathbf{I}}_{k_i} = (\tilde{I}_{k_i,1}, \dots, \tilde{I}_{k_i,\tilde{p}})'$  with  $\tilde{p}$  entries, where the entry at a specific position is the peak intensity, if one peak was assigned, and otherwise 0.

The most commonly used distance measure for the pairwise comparison of mass spectra is the cosine distance  $d_{\cos}$  [11]. For a pair of vectors of intensities  $\tilde{\mathbf{I}}_{k_i}$  and  $\tilde{\mathbf{I}}_{l_j}$ , the cosine similarity of the spectra  $k_i$  and  $l_j$  is the ratio of the dot product and the product of the Euclidean norms of the intensity vectors, according to the alternative definition. The cosine distance  $d_{\cos}$  is then calculated by subtracting the cosine similarity from 1:

$$d_{\cos}(S_{k_i}, S_{l_j}) = 1 - \frac{\langle \tilde{\mathbf{I}}_{k_i}, \tilde{\mathbf{I}}_{l_j} \rangle}{|\tilde{\mathbf{I}}_{k_i}| |\tilde{\mathbf{I}}_{l_j}|}$$

$$= 1 - \frac{\sum_{q=1}^{\tilde{p}} \tilde{I}_{k_i,q} \cdot \tilde{I}_{l_j,q}}{\sqrt{\sum_{q=1}^{\tilde{p}} \tilde{I}_{k_i,q}^2} \cdot \sqrt{\sum_{q=1}^{\tilde{p}} \tilde{I}_{l_j,q}^2}}$$

Depending on preprocessing of the spectra, e.g. only considering the top `topn` (`topn` ∈ N) peaks of each spectrum, a cosine distance that neglects intensities, is appropriate. Novak and Hoksza [12] have introduced the angle distance, a distance corresponding to cosine distance, with the original spectrum definition given by:

$$d_{\text{angle}}(S_{k_i}, S_{l_j}, \epsilon)$$

$$= \arccos \left( \frac{\sum_{q=1}^{p_{k_i}} \max_{q^*=1, \dots, p_{l_j}} 1_{\{|x_{k_i,q} - x_{l_j,q^*}| \leq \epsilon\}}}{\sqrt{p_{k_i} \cdot p_{l_j}}} \right)$$

Several other distance measures have been discussed, such as Pearson's or Spearman's correlation [11]. Novak and Hoksza [12] have introduced the Parametrized Hausdorff distance  $d_{\text{PH}}(S_{k_i}, S_{l_j}, \delta, k) = \max(h(S_{k_i}, S_{l_j}, \delta, k), h(S_{l_j}, S_{k_i}, \delta, k))$  with

$$h(S_{k_i}, S_{l_j}, \delta, k)$$

$$= \frac{1}{p_{k_i}} \sum_{q=1}^{p_{k_i}} \left( \min_{q^*=1, \dots, p_{l_j}} |x_{k_i,q} - x_{l_j,q^*}| 1_{\{|x_{k_i,q} - x_{l_j,q^*}| > \delta\}} \right)^{1/k}.$$

Given an error tolerance  $\delta$ ,  $h$  averages the  $k$ -th root of the minimal absolute distance greater than  $\delta$  of the position of all peaks of spectrum  $S_{k_i}$  compared to all peaks of spectrum  $S_{l_j}$ .

### DISMS2

The pseudo code of our new algorithm DISMS2 is shown in Algorithm 1. Calculating the pairwise **DI**Stances of  $N$  MS/MS (**MS2**) runs is a four-step procedure, consisting of spectra filtering, checking constraints for matching, matching of MS2 spectra, and calculation of the distance matrix with pairwise distances of MS/MS runs. The algorithm has been implemented in the statistical programming language R [13].

The first step is preprocessing and filtering of MS/MS spectra. Additionally it can be specified if only peaks of MS2 spectra with top `topn` highest intensities are included in the analysis. Then all spectra are binned with a flexible binsize `bin`. Binning with a fixed binsize `bin = 0.2` has been applied in compareMS<sup>2</sup> [3]. All intensities with  $m/z$  ratio  $x$  in a small interval  $[n^* \cdot \text{bin}, (n^* + 1) \cdot \text{bin})$  ( $n^* \in \mathbb{N}_0$ ) are replaced by one

---

### Algorithm 1: DISMS2: DIStance of MS/MS (MS2) runs

---

**Input:**

- $N$  MS/MS runs.
- Parameters `topn`, `bin`, `ret`, `prec`, `dist` and `cdis`.

**Output:** Distance matrix containing the pairwise distances of  $N$  MS/MS runs.

First carry out step 1 for each run  $i$  ( $i = 1, \dots, N$ ) separately. Steps 2 and 3 are carried out for each pair  $(i, j)$  of MS/MS runs separately and repeated with exchanged  $i$  and  $j$ . Finally pool all results in step 4.

- 1: If only the top `topn` peaks of spectra shall be considered, filter all MS2 spectra in run  $i$ . Bin all MS/MS spectra in run  $i$  with bin size `bin`.
  - 2: For each MS/MS spectrum in run  $i$  check if the following constraints are fulfilled for each MS/MS spectrum in run  $j$ . If no spectrum fulfills all constraints, count as match with distance greater than `cdis`. Let  $k$  be the rank of an MS2 spectrum in run  $i$  (temporal order) and  $l$  be the rank of an MS/MS spectrum in run  $j$ . Constraints:
    - (a)  $k - \text{ret} \leq l \leq k + \text{ret}$ .
    - (b) Same precursor charge of spectra  $k$  and  $l$ .
    - (c) Similar precursor mass:
 
$$m_{k_i} \cdot (1 - 10^{-6} \cdot \text{prec}) \leq m_{l_j} \leq m_{k_i} \cdot (1 + 10^{-6} \cdot \text{prec})$$
  - 3: Match MS2 spectrum with rank  $k$  in run  $i$  with MS/MS spectrum in run  $j$  with smallest distance `dist` of all MS/MS spectra fulfilling constraints (a)–(c) in step 2. As a directed distance measure  $d^*(i, j)$  between MS/MS runs  $i$  and  $j$  calculate the frequency of spectra in run  $i$  with no match (all distances greater than `cdis`) in run  $j$ .
  - 4: Create a distance matrix  $d$  whose entry at position  $(i, j)$  is the distance between MS/MS runs  $i$  and  $j$ :  $d(i, j) = (d^*(i, j) + d^*(j, i))/2$ ,  $d^*(i, j) = \# \{\text{spectra in } i \text{ with no match in } j\} / \# \{\text{spectra in } i\}$ .
- 

representative, the maximum intensity at the central  $m/z$  ratio value  $(n^* + 0.5) \cdot \text{bin}$ .

The concept of the procedure is to match all MS/MS spectra in run  $i$  with the most similar spectra in run  $j$  and vice versa. Due to a long computing time and for reasons of content the number of possible matching candidates is reduced by checking three constraints in step 2.

Constraints (a)–(c) are checked in the following order. First, only spectra with a similar retention time are considered. The usage of a HPLC before MS analysis justifies the constraint of a similar retention time since peptides with the same properties elute from the column at the same time. While retention time alignment is necessary for an optimal analysis, the following method is fast and adequate for filtering. To save computing time, instead of determining a time window the tolerance `ret` limits the number of MS/MS spectra in run  $j$  before and after any MS/MS spectrum in run  $i$  by the rank of the scan number. By filtering all candidate spectra with rank  $l$  in an interval  $[k - \text{ret}, k + \text{ret}]$  for matching with spectrum  $k$  in run  $i$ , at most  $2 \cdot \text{ret} + 1$  MS2 spectra remain. The tolerance `ret` can be increased to ensure that no best matching spectra are missed.

Although mass to charge ratios are displayed to allow for comparisons of spectra with different charge states, spectra might look different due to different properties. For this reason constraint (b) guarantees that only matches of spectra with the same precursor charge state are considered. Again the number of possible candidates is reduced.

Apparently peptides with the same amino acid sequences and the same post-translational modifications have same masses. This leads to constraint (c) allowing only the comparison of spectra with similar precursor masses. For any spectrum  $k$  in run  $i$  with precursor mass  $m_{k_i}$  only spectra  $l$  of run  $j$  inside a small interval  $[m_{k_i} \cdot (1 - 10^{-6} \text{prec}), m_{k_i} \cdot (1 + 10^{-6} \text{prec})]$  are considered. This ensures a maximum precursor mass accuracy of `prec` ppm (parts per million):  $|(m_{l_j} - m_{k_i})/m_{k_i}| \cdot 10^6 \leq \text{prec}$ . If no spectrum of run  $j$  fulfills all constraints for spectrum  $k$  in run  $i$ , no match is available. Then spectrum  $k$  is assigned a distance greater than a threshold `cdis`.

Finally in step 4 sequentially for each spectrum  $k$  in run  $i$  distances `dist` to all remaining candidate spectra in run  $j$  are calculated. As mentioned above several appropriate distance measures can be chosen, for example the most commonly used cosine distance  $d_{\text{cos}}$ , the angle distance  $d_{\text{angle}}$ , or the parametrized Hausdorff distance  $d_{\text{PH}}$ . Depending on the chosen distance measure `dist` a binary cut-off threshold `cdis` for a hit, i.e. same or different spectrum, has to be fixed. For the cosine distance  $d_{\text{cos}}$  Palmblad and Deelder [3] have shown that `cdis` = 0.2 is a reasonable choice. A distance  $d^*(i, j)$  between runs  $i$  and  $j$  is the frequency of spectra in run  $i$  with no match (all distances greater than `cdis`) in run  $j$ . Match means that the spectrum in run  $j$  with smallest distance `dist` is considered and this minimal distance is smaller than `cdis`. The distance  $d^*$  is not symmetric because the process is directed. Thus the procedure is repeated with exchanged runs  $i$  and  $j$  and finally the mean  $d(i, j) = (d^*(i, j) + d^*(j, i))/2$  of the directed distances is

calculated.  $N \cdot (N - 1)$  directed distances  $d^*$  have to be calculated to fill the distance matrix  $d$ .

### Parameter optimization for DISMS2

In our new algorithm DISMS2 several parameters have to be set and therefore an appropriate parameter optimization is needed. Given the data as mentioned above, a distance matrix between runs with size  $27 \times 27$  is computed using DISMS2. Distances within groups of technical replicates of organisms should be smaller than distances between different organisms. However, due to the random selection of precursor ions for MS/MS analysis, data dependent acquisition (DDA) is biased to the most abundant peptides present in a complex sample. Further, in DDA mode the intra-sample variation of peptide identifications between technical replicates is high ( $\geq 50\%$ ) [14]. For this reason, an ANOVA like approach for distance matrices, `adonis` (R package `vegan` [15]), is used for evaluation. Anderson [16] has introduced this non-parametric multivariate analysis of variance which is applicable for a distance matrix explaining different sources of variation. Variations of distances are divided into two parts, one representing variation of technical replicates within species and the other variation between species. A permutation test (by default with 10 000 permutations) with pseudo-F ratio between the mean sum of squares of distances between and within species is used. As goodness-of-fit measure the partial R-squared between groups of technical replicates, i.e. the ratio of sum of squares of distances between species and the sum of squares of all distances, is used. Values close to 1 are desired.

### Results and discussion

First, the application of the new algorithm DISMS2 on real data is presented in detail, including data preprocessing and choice of parameters. Second, DISMS2 is compared to a common Mascot database search (on spectra and peptide level). Finally, distances between species are visualized by dendrograms using average linkage hierarchical clustering.

#### Application of DISMS2 on real data with parameter optimization

We implemented DISMS2 in R [13] and applied it on 27 MS/MS runs to compare samples of human, mouse, yeast, roundworm, fruit fly, two different *Radix* species and two different foraminifera species. The ProteoWizard tool `MSConvertGUI` [17] was used to convert Thermo RAW files into `mzXML` files. The open data format `mzXML` can be read with the R package `readMzXMLData` [18].

To find appropriate parameter settings in DISMS2 we used a full factorial design. Due to time and memory costs the number of parameter combinations was limited. The values of factor levels in design 1 were set

according to preliminary investigations and expert advice (see Table 1). Especially the accepted precursor mass shift  $prec$  is set constant (10 ppm). Since the angle distance is not bounded in the interval  $[0, 1]$ , in design 2 more factors were added with higher values for the cutoff  $cdis$ . In total 81 factor combinations were compared by means of the partial R squared as goodness of fit measure, based on *adonis*, an ANOVA like approach for distance matrices.

The results of the parameter optimization are summarized in Table 2. In particular distance measures are of great interest, so bold printed lines correspond to the optimized parameters for the different distance measures.

We requested 3GB RAM on one core of an eight-core Intel Xeon E5-2630 (2.4 GHz, 128 GB RAM, Debian Linux 8.3.0 operating system). The median runtime of DISMS2 for 27 MS/MS runs for one of in total 81 factor combinations (settings) is 13.41 hrs (range 4.49 - 20.01 hrs). For the best ranked setting (see Table 2) 15.73 hrs are needed. The use of constraints (a) – (c) in step 2 of Algorithm 1, especially an appropriate choice of retention time tolerance  $ret$ , drastically reduces the runtime of DISMS2. Increasing  $ret$  for the best ranked setting from 3000 to 40000 leads to a runtime of more than 26 hrs, an increase by a factor of 1.65. Removing all constraints would result in a dramatic runtime increase in the range of 2 years. The use of constraints leads to a small number of candidates for matches. In case of H1 and H2, for example, on average only 2.5 out of more than 35000 candidates remain.

To quantify the influence of the parameters a regression tree was constructed (Fig. 1, [19]). The choice of the distance measure has the highest influence. The cosine distance outperforms  $d_{PH}$  and  $d_{angle}$ . Still, the parametrized Hausdorff distance is competitive when a higher cutoff is chosen.

The impact of the choice of a distance measure for spectra is shown in Fig. 2 illustrating two similar and two diverse spectra, respectively. The value of the angle distance in case of two similar spectra is almost as large as the value of the cosine distance in case of two distinct spectra.

**Table 2** Results of parameter optimization for DISMS2 with partial R-squared (*adonis*)

Rank	$topn$	$bin$	$ret$	$dist$	$cdis$	partial R-squared
1	$\infty$	<b>0.20</b>	<b>3000</b>	$d_{cos}$	<b>0.3</b>	<b>0.923</b>
2	50	0.20	3000	$d_{cos}$	0.3	0.923
3	20	0.20	3000	$d_{cos}$	0.3	0.923
4	$\infty$	0.20	3000	$d_{cos}$	0.1	0.892
5	50	0.20	3000	$d_{cos}$	0.1	0.892
6	20	0.01	3000	$d_{cos}$	0.3	0.890
7	20	0.20	3000	$d_{cos}$	0.1	0.890
8	50	0.01	3000	$d_{cos}$	0.3	0.890
9	$\infty$	0.01	3000	$d_{cos}$	0.3	0.890
<b>10</b>	<b>20</b>	<b>0.01</b>	<b>3000</b>	$d_{PH}$	<b>0.3</b>	<b>0.879</b>
11	$\infty$	0.20	1000	$d_{cos}$	0.3	0.878
⋮	⋮	⋮	⋮	⋮	⋮	⋮
<b>23</b>	<b>20</b>	<b>0.20</b>	<b>3000</b>	$d_{angle}$	<b>0.6</b>	<b>0.808</b>
⋮	⋮	⋮	⋮	⋮	⋮	⋮
81	$\infty$	0.20	3000	$d_{angle}$	0.3	0.308

Accepted precursor mass shift constant ( $prec = 10$  ppm)

Bold printed lines correspond to the optimized parameters for the different distance measures

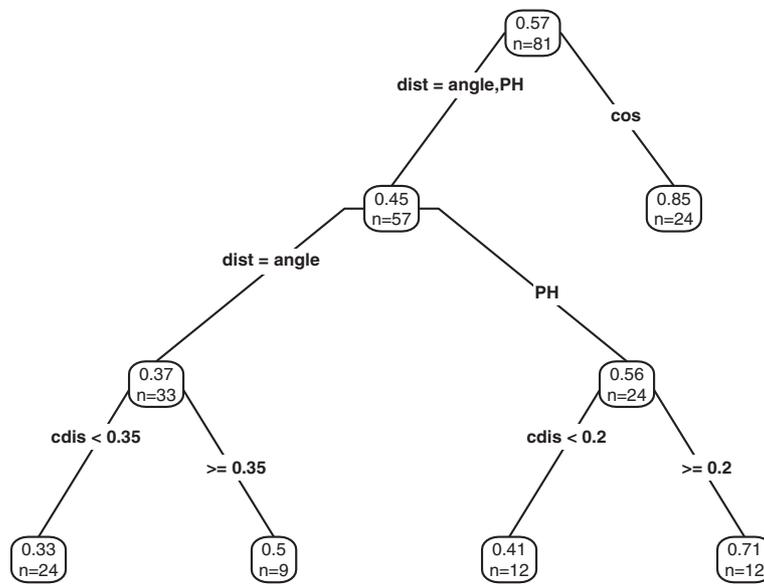
### Comparison of DISMS2 with state of the art database search

To evaluate the quality of DISMS2 we compared it to database search algorithms. DISMS2 requires a filter check, whereas database search requires an annotation check. We analyzed several methods combining algorithmic steps in different ways, see Table 3 for a list of all compared algorithms.

For all algorithms we calculated the average relative number of no hits of a directed search. The general principle is that a list of spectra associated with an MS/MS run is compared one by one with a second list of candidate spectra associated with a second MS/MS run. The algorithms differ in the definition of a hit, the spectrum universe, possible annotation and filter checks, and potential removal of duplicates.

**Table 1** Design of factor levels of parameters in evaluation of DISMS2. All 81 different combinations of all parameter values, 72 combinations for design 1, and 9 combinations for design 2

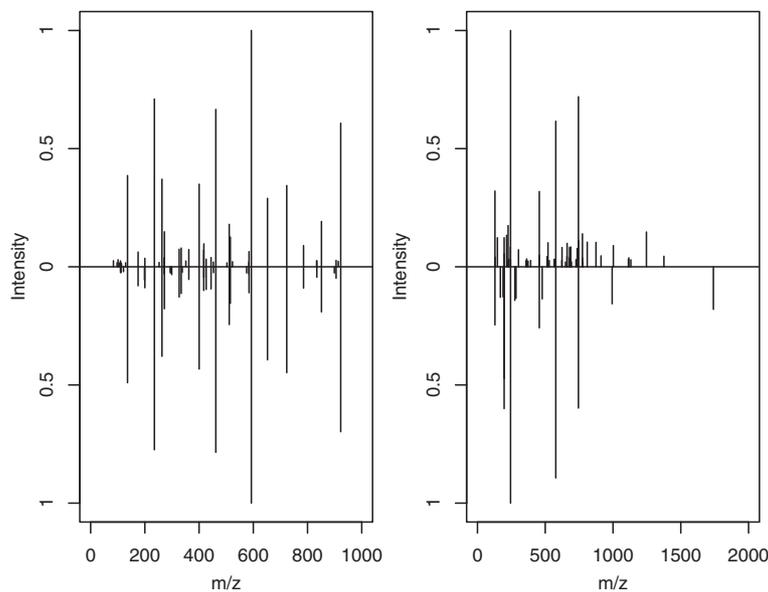
Parameter	Description	Design 1	Design 2
$topn$	$topn$ highest peaks, $\infty$ means no selection	20, 50, $\infty$	20, 50, $\infty$
$bin$	binning with binsize $bin$	0.01, 0.2	0.2
$ret$	size of retention time window	1000, 3000	3000
$prec$	accepted precursor mass shift (ppm)	10	10
$dist$	distances between spectra	$d_{angle}(\epsilon = 0.05)$ , $d_{cos}$ , $d_{PH}(\delta = 0.05; k = 50)$	$d_{angle}(\epsilon = 0.05)$
$cdis$	cutoff for distance of spectra	0.1, 0.3	0.4, 0.5, 0.6



**Fig. 1** Regression tree fitted to explain the partial R-squared (adonis) based on combinations of parameter settings. Different Parameter settings of `topn`, `bin`, `ret`, `dist`, and `cdis` were used. Each node displays the average partial R-squared in the node (top) and the number of observations that fall in the node (down). Classification was performed using the statistical programming language R, R package `rpart` [19]

Search method *database* means that a hit is counted if in the list of candidate spectra the same peptide is annotated. For the search method *distance* a hit means that the distance between the spectrum and a candidate spectrum is smaller than `cdis = 0.3`.

The spectrum universe is the union of all MS/MS spectra in respective runs. If *all* spectra are included in the analysis complete runs are used. A *reduced* spectrum universe stands for a selection of MS/MS spectra with peptide annotations.



**Fig. 2** Comparison of two similar (left) and two diverse (right) spectra. Preprocessing (`topn=20`, `bin=0.01`) was used and parameters for distance measures are as in Table 1. Distances for left example:  $d_{cos} = 0.010$ ,  $d_{angle} = 0.451$ ,  $d_{PH} = 0.105$ , distances for right example:  $d_{cos} = 0.396$ ,  $d_{angle} = 1.204$ ,  $d_{PH} = 0.746$

**Table 3** Overview over all algorithms compared

Name	Search method	Spectrum universe	Annotation check	Filter check	Duplicates
DB.ra	Database	Reduced	Yes	No	Kept
DB.ra.nodup	Database	Reduced	Yes	No	Removed
DISMS2.f	Distance	All	No	Yes	Kept
DB.a	Database	All	Yes	No	Kept
DISMS2.af	Distance	All	Yes	Yes	Kept
DB.af	Database	All	Yes	Yes	Kept

Annotation check means that for matching candidates both spectra have to be annotated by a Mascot peptide-to-spectrum-match. Possibly matching candidates are downgraded as no hit. For about 30 to 60% of all spectra (on average for roundworm 29%, human 34%, mouse 40%, fruit fly 45% and yeast 57%) peptide hits are missing, i.e. the database search with Mascot resulted in no hit, meeting a 1% FDR criterion.

Filtering means that for matching candidates the constraints for retention time, precursor mass and charge state have to be fulfilled (see step 3 in DISMS2).

Usually duplicated peptide annotations are *kept* meaning that all spectra are considered in the spectrum universe. *Removing* duplicates (nodup) of peptide annotations means that a hitlist of all peptides annotated at least once is used.

We now describe all compared algorithms listed in Table 3. The methods DISMS2.f and DB.ra are the default versions. DISMS2.f uses filtering and considers *all* spectra, whereas DB.ra uses no filtering and considers only spectra that are annotated with a peptide sequence. In DB.ra.nodup, additionally multiple spectra annotated with the same peptide are replaced by one representative (*removal* of duplicates).

Search method *distance* is associated with a filter check whereas search method *database* includes an annotation check. The differences between DISMS2.f and DB.a are only the annotation check and the filter check. All spectra are included in the spectrum universe of DB.a, in contrast to DB.ra with a reduced universe.

For a meaningful comparison we also considered two algorithms with both filtering and annotation checks. Including all MS/MS spectra in the spectrum universe, DISMS2.af and DB.af only differ by the search method.

The mean distances of runs are computed for all methods between and within species, as shown in Table 4 and in Fig. 3. In Table S1 [see Additional file 2] the total number of MS/MS spectra for different species comparisons is shown in detail. Corresponding standard errors indicating variations of technical replicates are negligibly

small, see Table S2 [see Additional file 3]. In comparisons within species standard errors are smaller than 0.008, and between species at most 0.002.

The mean distances are small within species and large between species. DB.ra and DB.ra.nodup generate smaller values, followed by DISMS2.f. DB.a is not able to keep up. Largest distances are generated for DB.af and DISMS2.af. It should be noted that the exact composition of peptides of analyzed samples is unknown as database annotation might be incorrect or incomplete. Thus the interpretation of larger or smaller values might be imprecise.

For the algorithms DB.ra and DB.ra.nodup the distances are calculated based on a reduced spectrum universe so that many spectra with low quality have been removed. Thus it is expected that many spectra without annotation implying large distances are removed, which leads to smallest values within species smaller than 20% in Table 4.

Since DB.af uses the same filtering as DISMS2.af, results are directly comparable. Annotation and filtering checks reduce the number of possible matches. As aforementioned missing annotations in list 1 have a share of 30 to 60%. Furthermore for 8 to 13% (within species) and 30 to 50% (between species) of all spectra no candidate spectra are remaining that fulfill all filtering requirements and are annotated (for further details see Table S3 [see Additional file 4]).

In case of same annotated peptides the cosine distance is typically small, see Fig. 4 (left, dark gray) with the distribution of the cosine distance of MS/MS spectra for two human samples (H1 vs. H2), with mode near 0. However, if the peptide annotation of the matching spectrum is different, the situation differs (Fig. 4, left, light gray). Most of the distances are high as one might expect. Only a few values are smaller than 0.3. In these cases very similar spectra are marked as different by the annotation approach, possibly due to missing or wrong hits. Cosine distance as a binary classifier for same or different peptides is a good choice. The ROC curve (Fig. 4, right) displays the performance, plotting the true positive rate (TPR) against the false positive rate (FPR) for different thresholds  $c_{dis}$ . The area under the curve is 0.93 indicating a good performance. For the chosen threshold  $c_{dis} = 0.3$  we obtain TPR = 92.3% and FPR = 13.3%.

The commonly used method DB.ra.nodup differs from the other methods because duplicated measured peptides are weighted different. The mean distance within species of the usual peptide list comparison DB.ra.nodup is about 18 percent, DISMS2.f based on all spectra leads to little higher values of about 30 percent.

Most of the comparisons between different species yield values above 90%, except the comparison of human and mouse samples. Again DB.ra leads to smallest values (55.1%), followed by DB.ra.nodup (62.3%) and DISMS2.f (62.4%).

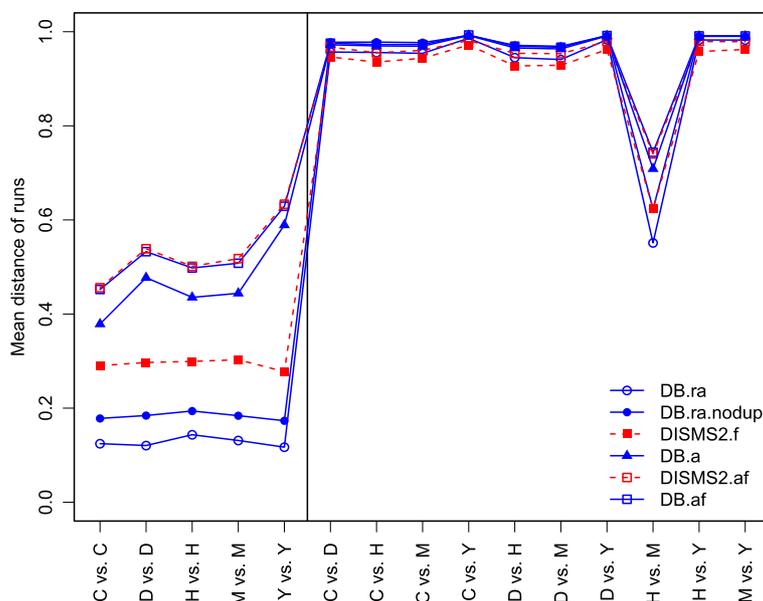
**Table 4** Mean distances of runs for different ways of proteome comparisons methods between and within species

Method	DB.ra	DB.ra.nodup	DISMS2.f	DB.a	DISMS2.af	DB.af
C vs. C	0.125	0.178	0.290	0.379	0.456	0.452
D vs. D	0.121	0.184	0.297	0.477	0.539	0.533
H vs. H	0.143	0.194	0.299	0.435	0.501	0.498
M vs. M	0.131	0.184	0.303	0.444	0.518	0.508
Y vs. Y	0.117	0.173	0.278	0.589	0.633	0.629
C vs. D	0.957	0.978	0.946	0.972	0.967	0.976
C vs. H	0.956	0.978	0.935	0.970	0.955	0.973
C vs. M	0.954	0.977	0.944	0.969	0.961	0.973
C vs. Y	0.986	0.992	0.972	0.992	0.981	0.993
D vs. H	0.945	0.971	0.927	0.966	0.954	0.969
D vs. M	0.941	0.969	0.929	0.963	0.953	0.967
D vs. Y	0.983	0.992	0.962	0.991	0.979	0.992
H vs. M	0.551	0.623	0.624	0.709	0.740	0.744
H vs. Y	0.982	0.991	0.958	0.990	0.979	0.991
M vs. Y	0.982	0.991	0.962	0.990	0.979	0.991

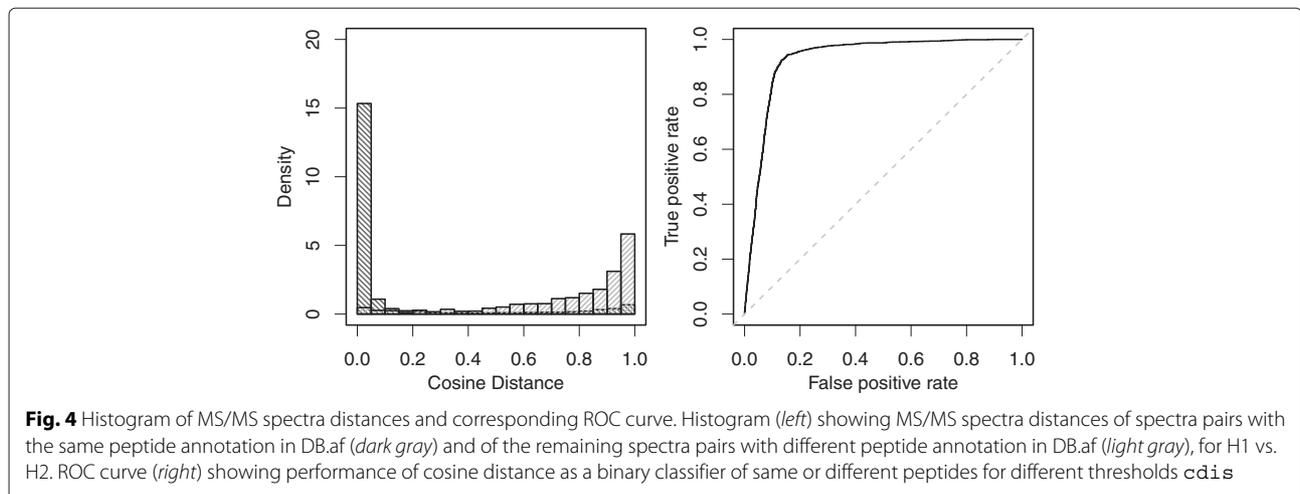
For different ways of proteome comparisons methods steps of search method, spectrum universe, annotation check, filter check and potential removal of duplicates between are combined. Mean distances of runs between (down) and within (top) species of roundworm (C), fruit fly (D), human (H), mouse (M) and yeast (Y) are shown

For the comparison of two methods we first calculated the absolute differences of proteome distances between the two methods and then the coefficient of variation (CV) of these values, i.e. the ratio of standard deviation and mean (see Table S4 and S5 [see Additional files 5

and 6]). A value smaller than 0.5 indicates a relevant difference between the corresponding two methods. In most cases there are relevant differences. Only DISMS2.af and DB.af perform very similar, since CVs are relatively large, some of them considerably larger than 0.5.



**Fig. 3** Mean relative number of partners for different ways of proteome comparisons methods. Steps of search method, spectrum universe, annotation check, filter check and potential removal of duplicates in different ways between (right) and within (left) species of roundworm (C), fruit fly (D), human (H), mouse (M) and yeast (Y) are compared



The filtering check is needed for the search method *distance* to speed up the computing time. Exemplary for the two human samples H1 and H2 DISMS2 is computed without filtering check. The results for all methods applied to these two samples in Table 5 show that DISMS2 without filtering is competitive even to the database methods with reduced spectrum universe.

#### Visualization of distances between species

Dendrograms using average linkage hierarchical clustering were used as trees to illustrate distances between MS/MS runs of different samples and its technical replicates. In average linkage clustering the mean distance between all pairs of elements is used for fusion of clusters.

The first dendrogram visualizes the distance matrix constructed with DISMS2.f for all 27 MS/MS runs (Fig. 5). The average distance between technical replicates is about 30%. Nodes connecting mouse and human samples (62.4%) as well as two *Radix* species (67.2%) and foraminifera species (76.1%) indicate high similarity.

**Table 5** Example of distance for two human MS/MS runs

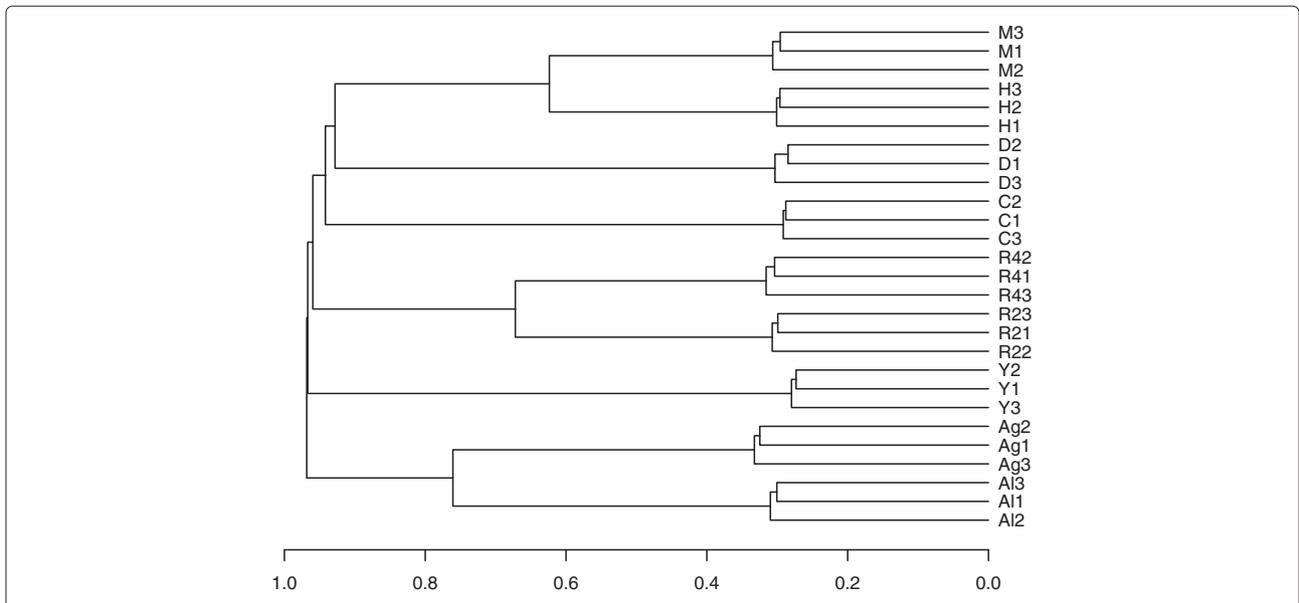
Name	$d^*(H1, H2)$	$d^*(H2, H1)$	$d(H1, H2)$
DB.ra	0.1467	0.1487	0.1477
DB.ra.nodup	0.1971	0.2010	0.1990
DISMS2.f	0.2969	0.3050	0.3009
DB.a	0.4321	0.4422	0.4371
DISMS2.af	0.4968	0.5087	0.5027
DB.af	0.4935	0.5068	0.5002
DISMS2	0.1616	0.1645	0.1631

Directed cosine distances  $d^*$  in both directions and mean of directed distances  $d$  of two human MS/MS runs H1 and H2 are compared. DISMS2 stands for DISMS2.f without filtering check

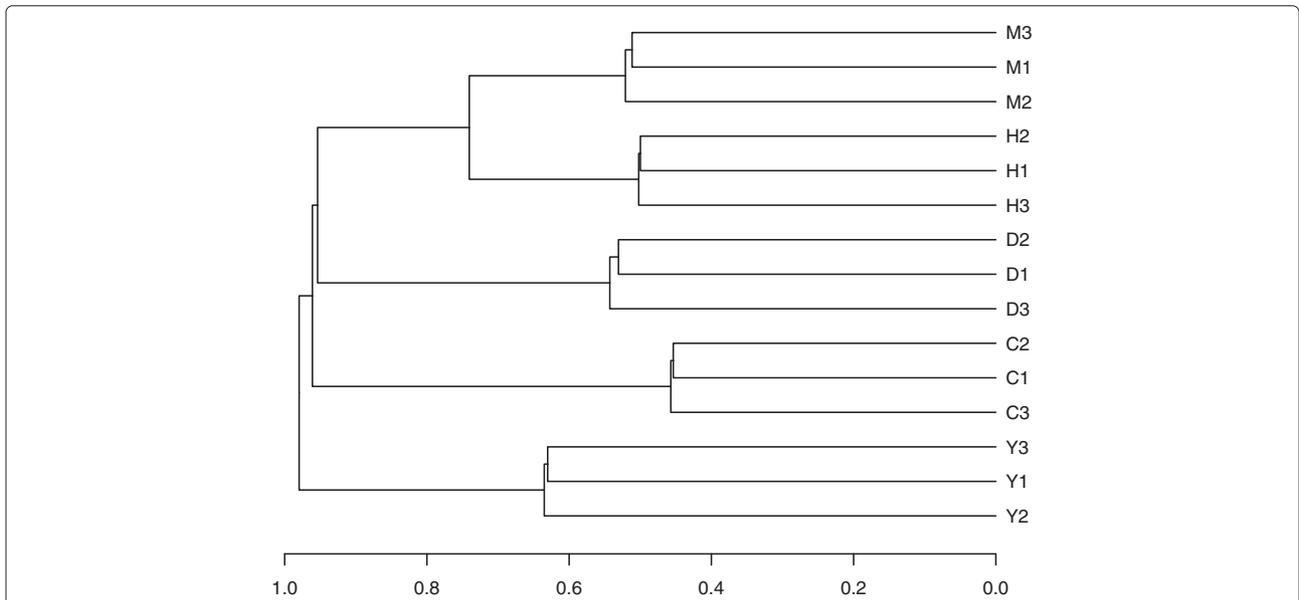
Further dendrograms (Fig. 6 and Figure S1–S4 [see Additional files 7, 8, 9 and 10]) are generated for the other methods for all 15 samples with available database annotations. The distance matrix for DB.af is almost identical to the one to DISMS2.af (Fig. 6), with notably less separation between species.

#### Conclusions

DISMS2 is a new user-friendly algorithm implemented in R for the proteome-wide distance calculation of different MS/MS runs. It performs well with data from different organisms, if parameter optimization is performed. Measuring technical replicates is the basis for selecting suitable parameters, based on an ANOVA like approach. Alternatively, prior knowledge can be used to choose adequate parameters. We carefully compared DISMS2.f (with filtering) with a state of the art method, the database search Mascot DB.a. Whereas DISMS2.f uses a filtering step that requires to set a critical distance  $cd.is$ , DB.a suffers from the large number of missing database hits. For a fair comparison, DISMS2.af and DB.af use the same spectrum universe of all spectra and perform annotation and filtering checks in the same way. A crucial difference is the different handling of duplicated spectra that are often neglected in database search methods. Thus a future goal is to perform clustering of spectra [20–23] before matching of partners, in order to further improve DISMS2 and make it directly comparable to the standard DB.ra.nodup. A huge benefit of DISMS2 is its applicability to samples from species without database annotation, as demonstrated on the fresh water snail *Radix* species (molecular operational taxonomic unit (MOTU) 2 and 4) and on the foraminifera *Amphistegina lessonii* and *gibbosa*. Further, when performing database searches in closely related species or applying de novo approaches DISMS2 can help to validate the results.



**Fig. 5** Dendrogram computed via DISMS2.f. Dendrogram for three technical replicates each of roundworm (C), fruit fly (D), human (H), mouse (M), yeast (Y), *Radix* MOTU 2 (R2), *Radix* MOTU 4 (R4), *A. gibbosa* (Ag) and *A. lessonii* (Al) using average linkage hierarchical clustering based on all pairwise distances of 27 MS/MS runs. Computed via DISMS2.f with optimized parameters:  $\text{topn} = \infty$ ,  $\text{bin} = 0.2$ ,  $\text{ret} = 3000$ ,  $\text{prec} = 10$ ,  $\text{dist} = d_{\text{cos}}$  and  $\text{cdis} = 0.3$



**Fig. 6** Dendrogram computed via DISMS2.af. Dendrogram for three technical replicates each of roundworm (C), fruit fly (D), human (H), mouse (M) and yeast (Y) using average linkage hierarchical clustering based on all pairwise distances of 15 MS/MS runs. Computed via DISMS2.af with optimized parameters:  $\text{topn} = \infty$ ,  $\text{bin} = 0.2$ ,  $\text{ret} = 3000$ ,  $\text{prec} = 10$ ,  $\text{dist} = d_{\text{cos}}$  and  $\text{cdis} = 0.3$

## Additional files

**Additional file 1:** Materials and Methods. This file provides detailed information about the sample preparation and LC-MS/MS analyses. (PDF 145 kb)

**Additional file 2: Table S1.** Total number of MS/MS spectra in Table 4. (PDF 8 kb)

**Additional file 3: Table S2.** Standard Errors belonging to mean distances in Table 4. (PDF 9 kb)

**Additional file 4: Table S3.** Additional information to Algorithm DB.af. Mean relative number of partners (same peptide), different peptides, missing annotation in list 1 and no remaining candidates after filtering in list 2 in Algorithm DB.af. (PDF 9 kb)

**Additional file 5: Table S4.** Coefficients of variation (CVs) of absolute differences of distances between two proteome comparisons (database) methods. CVs of absolute differences of distances between two methods, within species (top) and between species (bottom), for the species roundworm (C), fruit fly (D), human (H), mouse (M), and yeast (Y). Values smaller than 0.5 indicate relevant differences between the corresponding method pairs. Values greater than 0.5 are marked (\*). (PDF 8 kb)

**Additional file 6: Table S5.** Coefficients of variation (CVs) of absolute differences of distances between two proteome comparisons methods. CVs of absolute differences of distances between two methods, within species (top) and between species (bottom), for the species roundworm (C), fruit fly (D), human (H), mouse (M), and yeast (Y). Values smaller than 0.5 indicate relevant differences between the corresponding method pairs. Values greater than 0.5 are marked (\*). (PDF 9 kb)

**Additional file 7: Figure S1.** Dendrogram for three technical replicates each of roundworm (C), fruit fly (D), human (H), mouse (M) and yeast (Y) using average linkage hierarchical clustering based on all pairwise distances of 15 MS/MS runs. Computed via method DB.ra. (PDF 4 kb)

**Additional file 8: Figure S2.** Dendrogram for three technical replicates each of roundworm (C), fruit fly (D), human (H), mouse (M) and yeast (Y) using average linkage hierarchical clustering based on all pairwise distances of 15 MS/MS runs. Computed via method DB.ra.nodup. (PDF 4 kb)

**Additional file 9: Figure S3.** Dendrogram for three technical replicates each of roundworm (C), fruit fly (D), human (H), mouse (M) and yeast (Y) using average linkage hierarchical clustering based on all pairwise distances of 15 MS/MS runs. Computed via method DB.a. (PDF 4 kb)

**Additional file 10: Figure S4.** Dendrogram for three technical replicates each of roundworm (C), fruit fly (D), human (H), mouse (M) and yeast (Y) using average linkage hierarchical clustering based on all pairwise distances of 15 MS/MS runs. Computed via method DB.af. (PDF 4 kb)

## Abbreviations

Ag: *Amphistegina gibbosa*; Al: *Amphistegina lessonii*; ANOVA: Analysis of variance; C: roundworm *Caenorhabditis elegans*; CV: Coefficient of variation; D: fruit fly *Drosophila melanogaster*; DISMS2: Algorithm for calculating pairwise distances of MS/MS runs; DDA: Data dependent acquisition; DNA: Deoxyribonucleic acid; FDR: False discovery rate; FPR: False positive rate; H: human; HPLC: High-performance liquid chromatography; LC-MS/MS: Liquid chromatography-tandem mass spectrometry; M: mouse; MOTU: molecular operational taxonomic unit; MS: Mass spectrometry; MS2: Tandem mass spectrometry; MS/MS: Tandem mass spectrometry; m/z: Mass-to-charge ratio; PSM: Peptide-to-spectrum match; R2: MOTU 2; R4: MOTU 4; TPR: True positive rate; Y: yeast

## Acknowledgments

Prof. Pamela Hallock Muller (University of South Florida, St. Petersburg, USA) is warmly acknowledged for support in organizing and conducting fieldwork in Florida and Dr. Gita Roshni Narayan (Leibniz Center for Tropical Marine Ecology, Bremen) for providing the foraminifera samples from Zanzibar. The authors thank Prof. Dr. Klemens Störckuhl (AG Sinnesphysiologie,

Ruhr-Universität Bochum) for providing *Drosophila melanogaster* samples and Prof. Dr. Anna von Mikecz (Leibniz Research Institute for Environmental Medicine, Heinrich-Heine-University Duesseldorf) for providing *Caenorhabditis elegans* samples.

## Funding

This work was funded by the Leibniz-Competition Fund (SAW-2014-ISAS-2-D). BBL, KB, LK and AS gratefully acknowledge the financial support by the *Ministerium für Innovation, Wissenschaft und Forschung des Landes Nordrhein-Westfalen*, the *Senatsverwaltung für Wirtschaft, Technologie und Forschung des Landes Berlin*, and the *Bundesministerium für Bildung und Forschung*.

## Availability of data and materials

R code of DISMS2 is available at <http://www.statistik.tu-dortmund.de/genetics-publications-DISMS2.html>. The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE [24] partner repository with the dataset identifier PXD004824. Data are available via ProteomeXchange with identifier PXD004824.

## Authors' contributions

VR, MP, HW, AS and JR conceived the study. VR and JR designed the statistical algorithms. VR wrote the software and carried out the statistical analysis. BBL, MS, TS, KB and LK carried out experiments. AM interpreted results and put them into context. VR, BBL and JR wrote the manuscript. All authors read and approved the final version of the manuscript.

## Competing interests

The authors declare that they have no competing interests.

## Consent for publication

Not applicable.

## Ethics approval and consent to participate

*Amphistegina gibbosa* used in this research included specimens collected under permit number FKNMS-2011-011 issued to Pamela Hallock Muller by the Florida Keys National Marine Sanctuary.

## Author details

<sup>1</sup>Department of Statistics, TU Dortmund University, Dortmund, Germany. <sup>2</sup>Leibniz-Institut für Analytische Wissenschaften - ISAS - e.V., Dortmund, Germany. <sup>3</sup>Leibniz Center for Tropical Marine Ecology (ZMT), Bremen, Germany. <sup>4</sup>Biodiversity and Climate Research Centre, Senckenberg Gesellschaft für Naturforschung, Frankfurt, Germany. <sup>5</sup>Faculty of Biological Science, Institute for Ecology, Evolution and Diversity, Department of Molecular Ecology, Goethe University, Max-von-Laue-Straße 9, 60438 Frankfurt am Main, Germany. <sup>6</sup>Department of Chemistry, College of Physical Sciences, University of Aberdeen, Aberdeen, Scotland, United Kingdom. <sup>7</sup>Medizinische Fakultät, Medizinisches Proteom-Center (MPC), Ruhr-Universität Bochum, Universitätsstraße 150, 44801 Bochum, Germany.

Received: 8 October 2016 Accepted: 31 January 2017

Published online: 03 March 2017

## References

- Fields S. Proteomics. *Proteomics in genomeland*. Science. 2001;291(5507):1221–4.
- Method of the Year 2012. *Nat Methods*. 2013;10(1):1.
- Palmblad M, Deelder AM. Molecular phylogenetics by direct comparison of tandem mass spectra. *Rapid Commun Mass Spectrom*. 2012;26(7):728–32.
- Yilmaz S, Victor B, Hulstaert N, Vandermarliere E, Barsnes H, Degroev S, et al. A Pipeline for Differential Proteomics in Unsequenced Species. *J Proteome Res*. 2016;15(6):1963–70.
- Bantscheff M, Lemeer S, Savitski MM, Kuster B. Quantitative mass spectrometry in proteomics: critical review update from 2007 to the present. *Anal Bioanal Chem*. 2012;404(4):939–965.
- Eidhammer I, Flikka K, Martens L, Mikalsen SO. *Computational Methods for Mass Spectrometry Proteomics*. Chichester: John Wiley & Sons; 2007.
- Steen H, Mann M. The ABC's (and XYZ's) of peptide sequencing. *Nat Rev Mol Cell Biol*. 2004;5(9):699–711.

8. Perkins DN, Pappin DJ, Creasy DM, Cottrell JS. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*. 1999;20(18):3551–67.
9. Allmer J. Algorithms for the de novo sequencing of peptides from tandem mass spectra. *Expert Rev Proteomics*. 2011;8(5):645–57.
10. Medzihradsky KF, Chalkley RJ. Lessons in de novo peptide sequencing by tandem mass spectrometry. *Mass Spectrom Rev*. 2015;34(1):43–63.
11. Kim S, Zhang X. Comparative Analysis of Mass Spectral Similarity Measures on Peak Alignment for Comprehensive Two-Dimensional Gas Chromatography Mass Spectrometry. *Comput Math Methods Med*. 2013;2013:509761:12. doi:10.1155/2013/509761.
12. Novak J, Hoksza D. Parametrised Hausdorff distance as a non-metric similarity model for tandem mass spectrometry. In: CEUR Proc DATESO. Stedronin-Plazy, Czech Republic: In the Proceedings of the DATESO 2010 Annual International Workshop on DAtabase, TExts, Specifications and Objects; 2010. p. 1–12.
13. R Core Team. R: A Language and Environment for Statistical Computing: Vienna, Austria; 2016. <https://www.R-project.org/>. Accessed 1 Aug 2016.
14. Canterbury JD, Merrihew GE, MacCoss MJ, Goodlett DR, Shaffer SA. Comparison of data acquisition strategies on quadrupole ion trap instrumentation for shotgun proteomics. *J Am Soc Mass Spectrom*. 2014;25(12):2048–59.
15. Oksanen J, Blanchet FG, Friendly M, Kindt R, Legendre P, McGlenn D, et al. vegan: Community Ecology Package. 2016. R package version 2.4-0. <https://CRAN.R-project.org/package=vegan>. Accessed 1 Aug 2016.
16. Anderson MJ. A new method for non-parametric multivariate analysis of variance. *Austral Ecology*. 2001;26:32–46.
17. Chambers MC, Maclean B, Burke R, Amodei D, Ruderman DL, Neumann S, et al. A cross-platform toolkit for mass spectrometry and proteomics. *Nat Biotechnol*. 2012;30(10):918–20.
18. Gibb S. readMzXmlData: Reads Mass Spectrometry Data in mzXML Format. 2015. R package version 2.8.1. <https://CRAN.R-project.org/package=readMzXmlData>. Accessed 1 Aug 2016.
19. Therneau T, Atkinson B, Ripley B. rpart: Recursive Partitioning and Regression Trees. 2015. R package version 4.1-10. <https://CRAN.R-project.org/package=rpart>. Accessed 1 Aug 2016.
20. Beer I, Barnea E, Ziv T, Admon A. Improving large-scale proteomics by clustering of mass spectrometry data. *Proteomics*. 2004;4(4):950–60.
21. Frank AM, Bandeira N, Shen Z, Tanner S, Briggs SP, Smith RD, et al. Clustering millions of tandem mass spectra. *J Proteome Res*. 2008;7(1): 113–22.
22. Salmi J, Nyman TA, Nevelainen OS, T A. Filtering strategies for improving protein identification in high-throughput MS/MS studies. *Proteomics*. 2009;9(4):848–60.
23. Tabb DL, Thompson MR, Khalsa-Moyers G, VerBerkmoes NC, McDonald WH. MS2Grouper: group assessment and synthetic replacement of duplicate proteomic tandem mass spectra. *J Am Soc Mass Spectrom*. 2005;16(8):1250–61.
24. Vizcaino JA, Csordas A, del Toro N, Dianas JA, Griss J, Lavidas I, et al. 2016 update of the PRIDE database and its related tools. *Nucleic Acids Res*. 2016;44(D1):D447–56.

Submit your next manuscript to BioMed Central  
and we will help you at every step:

- We accept pre-submission inquiries
- Our selector tool helps you to find the most relevant journal
- We provide round the clock customer support
- Convenient online submission
- Thorough peer review
- Inclusion in PubMed and all major indexing services
- Maximum visibility for your research

Submit your manuscript at  
[www.biomedcentral.com/submit](http://www.biomedcentral.com/submit)

