Journal of Proteomics 245 (2021) 104282

Contents lists available at ScienceDirect

Journal of Proteomics

journal homepage: www.elsevier.com/locate/jprot

Leveraging the partition selection bias to achieve a high-quality clustering of mass spectra

André R.F. Silva [a,*], Diogo B. Lima [b], Louise U. Kurt [a], Mathieu Dupré [c], Julia Chamot-Rooke [c], Marlon D.M. Santos [a], Carolina Alves Nicolau [e,f], Richard Hemmi Valente [e], Valmir C. Barbosa [d,*,1], Paulo C. Carvalho [a,*,1]

[a] Laboratory of Structural and Computational Proteomics, Carlos Chagas Institute, Fiocruz Paraná, Brazil
[b] Department of Chemical Biology, Leibniz – Forschungsinstitut für Molekulare Pharmakologie (FMP), Berlin, Germany
[c] Mass Spectrometry for Biology Unit, CNRS USR 2000, Institut Pasteur, Paris, France
[d] Systems Engineering and Computer Science Program, Federal University of Rio de Janeiro, Rio de Janeiro, Brazil
[e] Laboratory of Toxinology, Oswaldo Cruz Institute, Fiocruz, Rio de Janeiro, Rio de Janeiro, Brazil
[f] Centre de Recherche en Cancérologie et Immunologie Nantes-Angers (CRCINA), Team SOAP, INSERM U1232, Nantes, France

ARTICLE INFO

Keywords:
Clustering
Tandem mass spectra
Partition assessment tool

ABSTRACT

In proteomics, the identification of peptides from mass spectral data can be mathematically described as the partitioning of mass spectra into clusters (i.e., groups of spectra derived from the same peptide). The way partitions are validated is just as important, having evolved side by side with the clustering algorithms themselves and given rise to many partition assessment measures. An assessment measure is said to have a selection bias if, and only if, the probability that a randomly chosen partition scoring a high value depends on the number of clusters in the partition. In the context of clustering mass spectra, this might mislead the validation process to favor clustering algorithms that generate too many (or few) spectral clusters, regardless of the underlying peptide sequence. A selection bias toward the number of peptides is desirable for proteomics as it estimates the number of peptides in a complex protein mixture. Here, we introduce an assessment measure that is purposely biased toward the number of peptide ion species. We also introduce a partition assessment framework for proteomics, called the Partition Assessment Tool, and demonstrate its importance by evaluating the performance of eight clustering algorithms on seven proteomics datasets while discussing the trade-offs involved.

Significance: Clustering algorithms are widely adopted in proteomics for undertaking several tasks such as speeding up search engines, generating consensus mass spectra, and to aid in the classification of proteomic profiles. Choosing which algorithm is most fit for the task at hand is not simple as each algorithm has advantages and disadvantages; furthermore, specifying clustering parameters is also a necessary and fundamental step. For example, deciding on whether to generate "pure clusters" or fewer clusters but accepting noise. With this as motivation, we verify the performance of several widely adopted algorithms on proteomic datasets and introduce a theoretical framework for drawing conclusions on which approach is suitable for the task at hand.

1. Introduction

Data clustering techniques are fundamental, widely adopted components of machine learning systems. They are used in such disparate fields as genomics [1], computer vision [2], social networks [3], hydrology [4], geochemistry [5], marketing [6], and psychology [7], to name a few. In proteomics, the clustering of tandem mass spectra has many applications, such as generating consensus spectra, speeding up database searches [8,9], quality control of wrongly annotated spectra [10], discovering novel molecules, and classification of proteomic profiles [11]. However, a considerable fraction of the mass spectra remains unidentified [12]. The lack of annotated spectra, especially in massive proteomics applications, hampers further insights into the data. It is challenging to reinterrogating large-scale repositories, such as PRIDE,

* Corresponding authors.
E-mail addresses: andrerfsilva@gmail.com (A.R.F. Silva), valmir@cos.ufrj.br (V.C. Barbosa), paulo@pcarvalho.com (P.C. Carvalho).
[1] Equal contributions.

https://doi.org/10.1016/j.jprot.2021.104282
Received 12 April 2021; Received in revised form 25 May 2021; Accepted 30 May 2021
Available online 2 June 2021

without resorting to spectral clustering algorithms as the same peptide are present multiple times.

The art of clustering requires essential and often difficult choices to be made, each involving the critical consideration of several possibilities as the approach to be used is set up. First, a feature selection function must be chosen so that the elements to be clustered can be mapped as multidimensional vectors (whose components are the features in question) in some appropriate space. Then a similarity function becomes necessary to compare vectors, such as the normalized dot product that is typically used in proteomics [9]. Next, the choice of a clustering algorithm, along with its parameters, must be undertaken. Last, and no less critically important, comes the choice of validation criteria to assess the resulting data partition.

During validation, a reference partition is often used, to which all candidate partitions under consideration are compared. By definition, all clusters in a reference partition are correct, which makes it a model for the assessment of any partitions output by the clustering algorithms in use. In proteomics, the reference partition can be created by assigning each spectrum to the cluster matching its peptide identification. Partition validation criteria relying on a reference partition are called external criteria. They can be useful when some domain-related expectation exists regarding the assignment of certain elements of the dataset to certain clusters. Partition assessment measures that do not use a reference partition are called internal criteria.

The need for correction for chance in validation criteria comes from the observation that even random algorithms can artificially score high values in assessment measures. Such was the case of the Rand Index [13] (RI), studied by Fowlkes and Mallows [14] and the subject of a critical comment by Wallace [15], which eventually led to Hubert and Arabie's [16] Adjusted Rand Index (ARI), a correction for chance inspired by Cohen's kappa [17,18]. In general, the κ-correction of an assessment measure f is given by

$$\kappa_0(f) = \frac{f - \mathrm{E}_0[f]}{max[f] - \mathrm{E}_0[f]},$$

where max[f] is the maximum possible value of f and $\mathrm{E}_0[f]$ is the expected value of f assuming that the possible partitions of the data in hand follow a certain distribution H_0. (Notably, ARI = κ_0(RI)). Thus, if the partition output by some random clustering algorithm follows distribution H_0, then the expected value of $\kappa_0(f)$ is zero for that algorithm.

This correction for chance is usually calculated assuming H_0 to be the so-called Permutation Model, which assigns nonzero uniform probability to all partitions having as many clusters as the reference partition and with the same sizes. However, the question of its plausibility was raised by Wallace [15] and discussed in depth by Gates and Ahn [19], who studied two additional possibilities: Fixed Number of Clusters and All Partitions (see Table 1).

However, correction for chance alone may not be enough. In fact, it can be shown that ARI has a selection bias. An assessment measure is said to have a selection bias if, and only if, the probability that a randomly chosen partition scoring a high value depends on the number of clusters in the partition. Controlling the selection bias is therefore fundamental in some assessment measures. In this regard, Romano et al. [20] showed statistical evidence that standardizing scores under the H_{perm} distribution may reduce the selection bias of assessment measures. For assessment measure f, standardization is achieved via

Table 1
Possibilities for H_0 (Gates and Ahn [19]).

H_0	Name	Definition
H_{perm}	Permutation Model	Uniform on all partitions having a given number of clusters and given cluster sizes.
H_{num}	Fixed Number of Clusters	Uniform on all partitions having a given number of clusters.
H_{all}	All Partitions	Uniform on all possible partitions.

$$\mathrm{S}_0(f) = \frac{f - \mathrm{E}_0[f]}{\sqrt{\mathrm{Var}_0[f]}},$$

where $\mathrm{Var}_0[f]$ is the variance of f under H_0.

On the other hand, it has been argued that some selection bias toward the number of clusters in the reference partition may be desirable in some applications [21]. This applies, for example, to the discovery of novel molecules. In this case, an ideal algorithm would output a partition with as many spectral clusters as there are peptides in the data, each cluster comprising spectra from the same peptide only.

In this work, we build on Rieder et al.'s [10] validation framework for the clustering of tandem mass spectra and discuss the problem of selection bias in assessment measures. We introduce an assessment measure that is purposely biased toward the number of peptide ion species and correspondingly a cluster assessment framework for proteomics, the Partition Assessment Tool. We demonstrate the framework's usefulness by evaluating the performance of eight widely-adopted clustering algorithms on seven proteomics datasets and discussing the trade-offs involved. Notably, in principle the validation criteria we present can be applied well beyond the clustering of mass spectra.

Before we proceed, we remark that our focus in this work is part of the broader area of quality control (QC) for spectral clustering, which has been reviewed in-depth by Perez-Riverol and collaborators [22] and gone through important landmarks that are worth noting. For example, after PRIDE Cluster had its efficiency verified against all identified spectra in PRIDE (some 21 million at the time), in 2016 its authors went on to cluster 256 million spectra and, thanks to robust QC, obtained clusters of unidentified peptides with unexpected PTMs. Spectral libraries also serve as the foundation for successful Data Independent Analysis (DIA), with more advanced approaches even generating a spectral library together with the DIA data [23]. QC of spectral clustering algorithms and search engines has also been achieved through synthetic peptides [24]; for example, proteometools.org has the ambitious goal of providing spectral libraries of tryptic synthetic peptides covering all canonical human proteins in UniProtKB/Swiss-Prot [25]. While on the one hand datasets are critical for the objective evaluation of proteomic algorithms [26], on the other hand advances in theoretical measures for assessing clustering effectiveness are also necessary. In the present work, our contribution to QC comes in the form of both a well-founded validation framework and its use in verifying a broad spectrum of algorithms on several datasets.

2. Materials and methods

2.1. Selection probability and selection bias

Given a reference partition V and an i-cluster partition U_0^i sampled according to distribution H_0, the selection probability of i relative to assessment measure f applied to the pair (U_0^i, V) is the probability that $f(U_0^i, V) \geq f(U_0^j, V)$ for any other partition size j in a given interval. A selection bias exists for f if, and only if, the selection probability is not the same for all i in the interval of possible partition sizes.

In this work, we used $H_0 = H_{\mathrm{num}}$ exclusively. The selection probability of assessment measures was estimated with Partition Assessment Tool's Selection Bias Simulator, created by us for this work. Simulations were run assuming a set of 100 elements and one fixed, random reference partition with 35 clusters. In each simulation a random candidate partition was generated with i clusters for i ranging within one of two possible intervals, 2 through 99 or 15 through 55 (this one centered at 35). The candidate partitions' scores on the assessment measure in use were then computed relative to the reference partition, and the highest-scoring partition (i.e., its number i of clusters) was selected. After 5,000 of these simulations, the selection probability was estimated for every partition size in the range in use. We proceeded in this way to look for assessment measures that are biased toward the number of peptide ion

species of the reference partition.

2.2. Primary and secondary assessment measures

Partition assessment measures can be divided into two classes, depending on the stage of clustering-algorithm evaluation they are used in. Primary assessment measures are used to select the highest-scoring partitions out of those produced by a group of algorithms. Secondary assessment measures are used to evaluate different aspects of the partitions selected. In this work, we propose Gaussian Biased True Similar Pairs as the primary assessment measure (see definitions below). For secondary assessment measures, we propose Jaccard Index, Variation in Partition Size, and Purity as external criteria, and Mean Pairwise Similarity and Dunn Index as internal criteria. We also show that the widely adopted ARI to be inappropriate as either a primary or a secondary assessment measure.

2.3. Partition assessment measures

2.3.1. External assessment measures

Let $X = \{x_1, x_2, \ldots, x_N\}$ be a set of N spectra, V (with C clusters) the reference partition of X obtained by the identification of peptide spectrum matchings, and U (with R clusters) the candidate partition to be evaluated with respect to V. A pair of spectra $\{x_i, x_j\}$ can be classified as one of four types regarding partitions U and V:

- *True similar pair*: x_i and x_j belong to the same cluster in U and to the same cluster in V;
- *False similar pair*: x_i and x_j belong to the same cluster in U but not to the same cluster in V;
- *False dissimilar pair*: x_i and x_j belong to the same cluster in V but not to the same cluster in U;
- *True dissimilar pair*: x_i and x_j belong to the same cluster neither in U nor in V.

We henceforth let a be the number of true similar pairs, b the number of false similar pairs, c the number of false dissimilar pairs, and d the number of true dissimilar pairs. A pair-counting index is simply a function of a, b, c, and d.

The Gaussian Biased True Similar Pairs, denoted by GBTSP, is the pair-counting index defined as

$$\mathrm{GBTSP} = e^{-\frac{1}{2}\left(\frac{R-C}{\sigma}\right)^2} \mathrm{S}_{\mathrm{num}}(a),$$

where $\mathrm{S}_{\mathrm{num}}(a)$ is the standardized value of a under the H_{num} distribution. In the formula for $\mathrm{S}_{\mathrm{num}}(a)$, obtained from the formula for $\mathrm{S}_0(f)$ given above with a substituting for f and H_{num} for H_0, the expected value $\mathrm{E}_{\mathrm{num}}[a]$ and the variance $\mathrm{Var}_{\mathrm{num}}[a]$ are as given in Supporting Section S1. GBTSP, therefore, is to be regarded as a version of $\mathrm{S}_{\mathrm{num}}(a)$ that is modulated by a Gaussian centered at $Z = C$ with width parameter σ, where Z is an integer spanning all possible partition sizes. The parameter σ is simply the RMSE relative to C of the partition sizes Z that lie within a 10% difference of C,

$$\sigma = \sqrt{\sum_{Z=0.9C}^{1.1C} \frac{(Z-C)^2}{0.2C}}.$$

As we discuss below, $\mathrm{S}_{\mathrm{num}}(a)$ has a reduced selection bias when compared to other assessment measures (therefore a selection probability that varies relatively little with partition size), whereas the Gaussian amplifies this bias ever more strongly as the partition size approaches the number of clusters in the reference partition.

Another pair-counting index that is useful is the Jaccard Index (JI), given by

$$\mathrm{JI} = \frac{a}{a+b+c}.$$

That is, JI is the proportion of true similar pairs (counted by a) in relation to the larger set of similar pairs, be these true similar (a), false similar (b), or false dissimilar (c), between candidate partition U and the reference partition V.

Variation in Partition Size (VPS) is a straightforward assessment measure of how much the difference in partition size from V to U (i.e., $R - C$) represents in relation to C, the number of clusters of the reference partition V. That is,

$$\mathrm{VPS} = \frac{R-C}{C}.$$

Algorithms that are good estimators of C will score a VPS close to zero.

Purity is defined as the proportion of spectra identified as the most frequent peptide in the cluster to which it was assigned by the clustering algorithm being considered. In relation to other assessment measures in Rieder et al. [10] we have the following. Assuming that the most frequent annotation in each cluster is correct, the Proportion of Incorrectly Clustered Spectra is in fact the complement of Purity to 1 and could also be called Impurity. The Proportion of Spectra Remaining is the number of clusters R of U divided by the number of spectra, N. The Retainment of Identified Peptides is the number of most frequent peptides in proportion to the total number of identified peptides. And the Proportion of Clustered Spectra is the proportion of spectra that do not belong to a singleton cluster.

2.3.2. Internal assessment measures

To assess partitions of the full datasets (see below) with unidentified spectra, we employed two internal assessment measures on partition U: the Mean Pairwise Similarity (MPS), defined as the mean cosine similarity of all pairs of spectra in the same cluster, and the Dunn index [27,28] (DI), defined as

$$\mathrm{DI} = \frac{\min\limits_{1 \le i < j \le R} \delta(u_i, u_j)}{\max\limits_{1 \le k \le R} \Delta_k},$$

where u_i and u_j are clusters in U, $\delta(u_i, u_j)$ is the dissimilarity (measured as the sine) between the representative spectra of clusters u_i and u_j, and Δ_k is the maximum dissimilarity between two spectra in cluster u_k. A good partition maximizes both MPS and DI.

2.3.3. RMSE

In the following analysis, a clustering algorithm is run multiple times on a dataset, each time with different parameters – such as similarity threshold, precursor tolerance, etc. We define JI_+ as the JI score of the top-scoring partition in terms of GBTSP, the primary assessment measure. We define VPS_+, Purity_+, MPS_+, and DI_+ in a similar way.

In general, we want to evaluate the overall performance of K algorithms on L datasets. Let f_+^{ij} be one of JI_+, VPS_+, Purity_+, MPS_+, or DI_+ for the ith algorithm on the jth dataset, $1 \le i \le K$ and $1 \le j \le L$. If f is an external assessment measure, let $\widehat{f}$ be its best possible value (that is, $\widehat{f}$ is obtained by letting $U = V$). If f is an internal assessment measure, let $\widehat{f}$ be the best score obtained by all K algorithms on all L datasets. Thus, in regard to f, the performance of the ith algorithm on all L datasets can be summarized as

$$\mathrm{RMSE}_i(f) = \sqrt{\frac{1}{L}\sum_{j=1}^{L}\left(f_+^{ij} - \widehat{f}\right)^2}.$$

Likewise, computing

$$\mathrm{RMSE}_j(f) = \sqrt{\frac{1}{K}\sum_{i=1}^{K}\left(f_+^{ij} - \widehat{f}\right)^2}$$

can help evaluate how challenging the *j*th dataset is to the *K* algorithms.

2.4. Proteomics datasets

We demonstrated the importance of our approach by reanalyzing datasets from Rieder et al. [10], deposited in the PRIDE repository with identifier PXD004824. Here, our goal was to produce a set of reference spectra with highly confident identifications so we could properly assess the clustering quality of the algorithms. As such, we extracted only spectra that were confidently identified for downstream analysis. The datasets contain spectra from roundworm (*Caenorhabditis elegans*), fruit fly (*Drosophila melanogaster*), human (*Homo sapiens*, HeLa cell line), mouse (*Mus musculus*, C2C12 cell line), and yeast (*Saccharomyces cerevisiae*). We also evaluated our strategy on a snake (*Bothrops jararaca*) venom subproteome (> 10 kDa) dataset (PRIDE PXD022124) generated by our group (motivated by the fact that this type of proteomic data is particularly challenging to be analyzed by the standard PSM procedure, as typically there are many splice variants and PTMs), and on an *Escherichia coli* K12 dataset (PRIDE PXD015367).

2.5. Generation of reference datasets with PatternLab for proteomics

A reference dataset for evaluating the clustering algorithms can be generated by using a PSM tool and filtering the identification results with high confidence. To achieve proteomic identifications, we followed the steps provided in the bioinformatic protocol [29] of PatternLab for proteomics 4 (PL4); this software is freely available at patternlabforproteomics.org.

For most mass spectral datasets, sequences were downloaded from Swiss-Prot and then a target-decoy database was generated to include a reversed version of each sequence plus those from 123 common mass spectrometry contaminants; the exception was the snake venom dataset, which was run against the reverse-decoy database of UniProt Serpents (taxonomy id 8570–156,122 entries - May 12, 2019) plus the aforementioned contaminants. PL4 relies on the embedded Comet 2016.01 rev. 3 search engine to assign scores for comparisons between experimental mass spectra and those theoretically generated from the sequence databases in question [30]. The search parameters considered were: fully and semi-tryptic peptide candidates with masses between 550 and 5500 Da, up to two missed cleavages, 40 ppm for precursor mass, and bins of 0.02 *m/z* for MS/MS. The modifications were carbamidomethylation of cysteine and oxidation of methionine as fixed and variable, respectively.

2.5.1. Validation of PSMs

In what follows, PL4's module known as Search Engine Processor [31] was used to assess the validity of the PSMs. To achieve this, PL groups the identifications by charge state (2+ and ≥ 3+) and then by tryptic status, resulting in four distinct subgroups. For each group, the XCorr, DeltaCN, DeltaPPM, and Peak Matches values were used to generate a Bayesian discriminator. The identifications were sorted in nondecreasing order according to the discriminator score. A cutoff score was accepted with a false-discovery rate (FDR) of 2% at the peptide level based on the number of decoys [32]. This procedure was independently performed on each subgroup, resulting in an FDR independent of charge state or tryptic status. Additionally, a minimum sequence length of five amino-acid residues and a protein score greater than 3 were imposed. Finally, identifications deviating by more than 10 ppm from the theoretical mass were discarded. This last filter led FDRs, now at the protein level, to be lower than 1% for all search results [26]. For this particular study, we performed an additional stringency step by post-processing our results to eliminate any identification with XCorr bellow 3.0 and one-hit-wonders (Spectral Count = 1). These steps were followed in order to obtain a gold-standard set of identifications and create a reference partition for each proteomics dataset.

2.5.2. Reference and full datasets

The subset of each proteomic dataset containing only spectra identified according to the aforementioned procedure (and thus considered as being of high-quality PSMs) is henceforth referred to as the reference dataset. Reference datasets were used to benchmark and optimize clustering algorithms. After that, clustering algorithms were run on each dataset's full set of spectra, including unidentified and noisy spectra, henceforth referred to as the full dataset. Table 2 summarizes the number of spectra in each reference and full dataset. The precursor ion of every spectrum accounted for in Table 2 has charge 2+, 3+, or 4 + .

2.6. Clustering algorithms

We tested eight clustering algorithms: DiagnoProt's clustering algorithm, which we henceforth refer to as Online Clustering algorithm (O-Cluster) [11], complete linkage hierarchical clustering (hclust) [33], CAST [34], N-Cluster [10,35], igraph [36], DBSCAN [37] (run with *minPts* = 3), PRIDE Cluster [38], and MS-Cluster [9]. With the exception of O-Cluster, all algorithms can be run with the R scripts provided by Rieder et al. [10].

O-Cluster works as follows. A cluster's representative spectrum is defined as the spectrum with the highest Xrea [39], a score that assesses spectral quality based on cumulative intensity normalization. The algorithm starts with every spectrum constituting a singleton cluster. Then for each pair of clusters, it performs a similarity test. If the normalized dot product between the two clusters' representative spectra is greater than the similarity threshold, then the two clusters are merged. The procedure continues until no two clusters can be merged.

2.7. Processing spectra

We have very limited control of the pre-processing, quality control, and vectorization subroutines of PRIDE Cluster and MS-Cluster, so we provided the spectra as they came from the RAW files. For the other algorithms, on the other hand, it was possible to establish a common procedure, as follows. All MS2 peaks below 250 *m/z* or above 1750 *m/z* were discarded. All MS2 peaks with intensity below the *seven* most intense isotopes in each bin of size 100 *m/z* were removed. To avoid selecting MS2 peaks from the same isotopic envelope, we removed all peaks with 3.0 Da from the list of most intense peaks. Spectra were binned with Bin Offset = 0.0 and Bin Size = 0.02, as recommended by Comet [30] vectorization. Then, all binned spectra were normalized so that each of their intensity vectors had norm 1.

All algorithms were executed with a precursor tolerance of ±0.75 and a wide range of similarity thresholds, varying from 0.1 to 0.9. The similarity threshold parameter is named and used differently by each algorithm (see Table 3). Still, all algorithms have in common that they employ it as a threshold for clustering spectra based on their vectorial similarity. This similarity is the normalized dot product between the vectors representing two spectra, or equivalently the cosine of the angle between them.

Table 2
Number of spectra in reference and full datasets.

Dataset	Reference version	Full version
C. elegans	38,658	119,813
D. melanogaster	30,213	109,930
H. sapiens	52,404	112,132
M. musculus	52,349	109,887
S. cerevisiae	38,538	108,929
B. jararaca	6,498	158,860
E. coli	42,143	199,419

Table 3
Similarity threshold parameters for the clustering algorithms.

Algorithm	Similarity Threshold
O-Cluster	*similarity_threshold*
hclust	*h*
CAST	*t*
N-Cluster	*c*
igraph	*cdis*
DBSCAN	ε
PRIDE Cluster	*threshold_end*
MS-Cluster	*similarity*

2.8. Partition assessment tool

Statistical analyses were carried out with the Partition Assessment Tool and its Selection Bias Simulator, both developed for this study and available for download at patternlabforproteomics.org/partitionassessmenttool.

3. Results and discussion

3.1. Selection bias

Fig. 1 shows that ARI (with $H_0 = H_{num}$) and JI are biased toward lower numbers of clusters. Purity and Retainment of Identified Peptides are biased toward higher numbers of clusters. These assessment measures are therefore not recommended as primary scores, lest they might end up selecting partitions by number of clusters instead of quality. This notwithstanding, most of them may turn out to be useful as secondary assessment measures, provided the one used as primary has an acceptable selection bias.

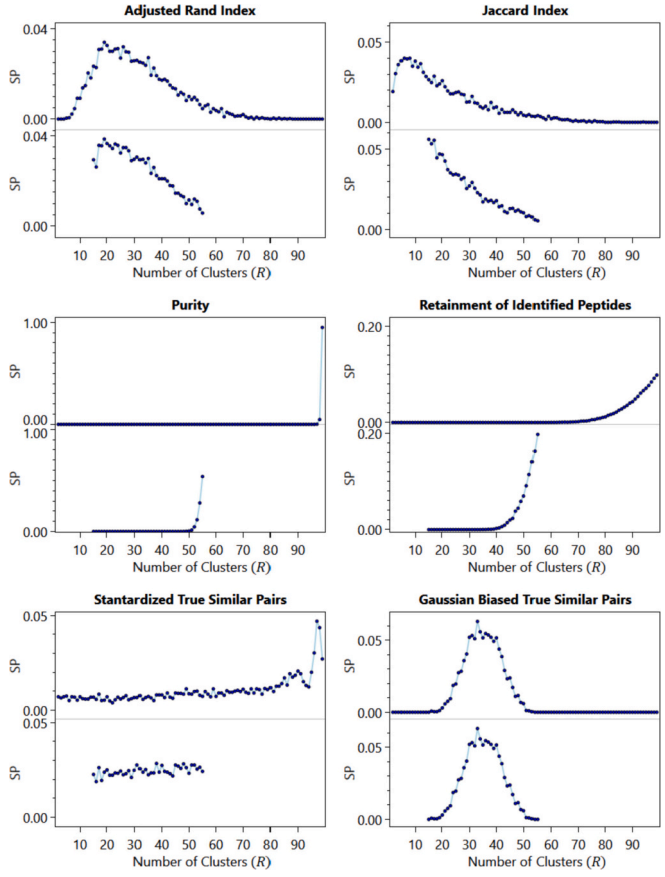

Fig. 1. Estimated selection probability (SP) as a function of the number of clusters of the candidate partition (R) in full range ($2 \leq R \leq 99$) and in practical range ($15 \leq R \leq 55$), plotted in upper and lower subplots, respectively. Adjusted Rand Index (ARI) and Jaccard Index (JI) are biased toward partitions having a lower number of clusters. Purity and Retainment of Identified Peptides are biased toward partitions having a higher number of clusters. Standardized True Similar Pairs ($S_{num}(a)$) is nearly unbiased in a practical range. Gaussian Biased True Similar Pairs (GBTSP) is biased toward the number of clusters of the reference partition.

The exception here is ARI, which remains not recommended even as a secondary assessment measure. This follows from a further flaw that is inherent to it, namely, that it assigns the same weight to both true similar and true dissimilar pairs. Wallace [15] already suspected that true dissimilar pairs are not a sure indicator of partition agreement, and in fact it can be shown that, for a sufficiently large number of clusters, $E_{num}[d] \approx \max[d]$ and $E_{num}[d] \approx \left(\frac{N}{2}\right)$ (see Supporting Section S1). Therefore, for any practical proteomics clustering application, true dissimilar pairs are useless and point to JI (which does not depend on d) as a better pair-counting secondary assessment measure.

Fig. 1 also shows that $S_{num}(a)$ has a reduced selection bias in a practical range for the number R of clusters ($15 \leq R \leq 55$, close therefore to the number of clusters of the reference partition, $C = 35$), and that GBTSP is biased toward C in the full range of all possible partition sizes ($2 \leq R \leq 99$). This makes GBTSP a good primary assessment measure.

3.2. Reference datasets

All clustering algorithms were run on each reference dataset. After undergoing screening by GBTSP, the primary assessment measure, their performances with the resulting optimized similarity thresholds (given in Supporting Table S1) were compared against those with a default similarity threshold (0.5, chosen to reflect the maximum possible uncertainty should a primary assessment measure be unavailable).

While Fig. 2(A) and Table 4 both indicate that no algorithm got ideally close to the reference partition on all assessment measures, and also that none was outstanding regarding JI, they also highlight some very positive outcomes. GBTSP optimization led to a better JI for O-Cluster, hclust, CAST, and N-Cluster, but to a worse JI for the other algorithms (worst of all for DBSCAN). All algorithms performed very well regarding VPS, which is the natural effect of the strong selection bias imposed by GBTSP. PRIDE Cluster was the best estimator of the real number of clusters. Regarding Purity, worse values were obtained in comparison with the default runs, but overall Purity values are still relatively high across all algorithms. As for the internal assessment measures, hclust obtanined the best MPS and O-Cluster the best DI. This is essentially as expected, since hclust is a natural optimizer of MPS and O-Cluster is a natural optimizer of DI. However, for most optimized runs MPS and DI are worse than for the default runs.

Fig. 2(B) and Table 5 show that the dataset RMSEs of JI, MPS, and DI turned out to be correlated (in the sense of Pearson correlation) with C/N, the ratio of the number of clusters in the reference partition to the size of the dataset. This suggests that each dataset presents a challenge to the clustering algorithms (regarding these assessment measures) that depends on C/N. Intuitively, if for a given dataset the number of clusters (C) is larger in comparison to the number of elements to be clustered (N) than it is for another dataset, then the former dataset is more challenging than the latter in the sense that classification erros become more likely.

Detailed assessment measures for each algorithm and each reference dataset, as well as the corresponding RMSEs, are available in Supporting Tables S2 and S3.

3.3. Full datasets

We also tested all algorithms on the full datasets, using the parameters provided by GBTSP optimization for the corresponding reference datasets. Performance was now evaluated only via JI, MPS, and DI, since the other secondary assessment measures are undefined for the full

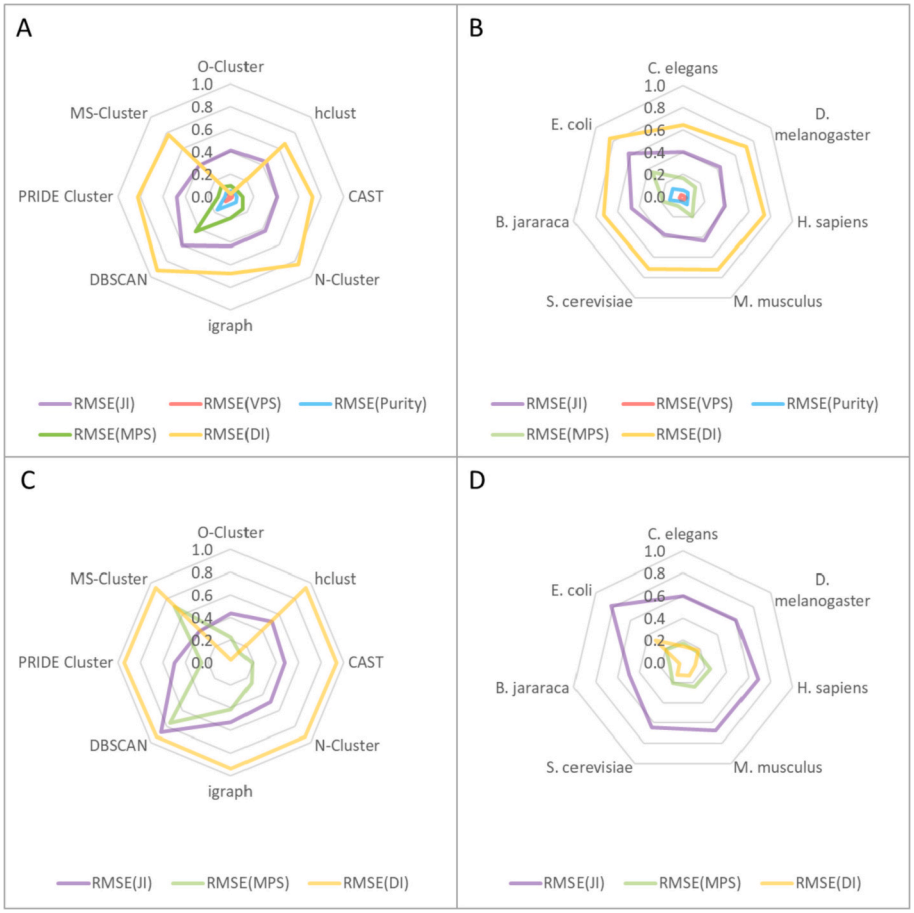

Fig. 2. RMSEs of secondary assessment measures on reference and full datasets. **A)** Algorithm RMSEs for reference datasets. **B)** Dataset RMSEs for reference datasets. **C)** Algorithm RMSEs for full datasets. **D)** Dataset RMSEs for full datasets.

Table 4
Algorithm RMSEs on referfence datasets for parameters optimized by GBTSP. Percentages in parentheses indicate growth relative to the default parameter.

Algorithm (*A*)	$RMSE_A$(JI)	$RMSE_A$(VPS)	$RMSE_A$(Purity)	$RMSE_A$(MPS)	$RMSE_A$(DI)
O-Cluster	0.412(−3.96%)	0.019(−79.3%)	0.051(+27.5%)	0.094(+5.62%)	0.024(+2300%)
hclust	0.440(−11.3%)	0.021(−89.9%)	0.052(+100%)	0.081(+179%)	0.669(+16.5%)
CAST	0.412(−5.94%)	0.016(−89.0%)	0.055(+66.7%)	0.111(+70.8%)	0.725(+12.6%)
N-Cluster	0.426(−0.93%)	0.010(−90.8%)	0.066(+65.0%)	0.153(+57.7%)	0.843(+12.1%)
igraph	0.432(+2.37%)	0.017(−81.3%)	0.066(+29.4%)	0.186(+25.7%)	0.678(+13.0%)
DBSCAN	0.603(+35.8%)	0.058(−80.7%)	0.167(+209%)	0.433(+166%)	0.918(+10.1%)
PRIDE Cluster	0.478(+0.63%)	0.008(−86.0%)	0.047(+20.5%)	0.113(−11.7%)	0.824(+10.2%)
MS-Cluster	0.389(+1.30%)	0.021(−46.2%)	0.049(+13.9%)	0.117(−5.65%)	0.781(+9.23%)

Table 5
Dataset RMSEs on reference datasets for parameters optimized by GBTSP.

Dataset (*D*)	$RMSE_D$(JI)	$RMSE_D$(VPS)	$RMSE_D$(Purity)	$RMSE_D$(MPS)	$RMSE_D$(DI)	*C/N*
C. elegans	0.401	0.012	0.059	0.165	0.645	0.286
D. melanogaster	0.425	0.010	0.057	0.138	0.718	0.265
H. sapiens	0.385	0.031	0.045	0.101	0.742	0.328
M. musculus	0.437	0.030	0.078	0.198	0.719	0.316
S. cerevisiae	0.379	0.032	0.046	0.095	0.712	0.304
B. jararaca	0.476	0.029	0.118	0.186	0.727	0.174
E. coli	0.624	0.024	0.112	0.351	0.843	0.412
Correlation with ***C/N***	0.421	0.088	−0.103	0.477	0.582	

datasets. A similar observation is in some sense applicable also to JI, since it is an external assessment measure. However, JI can consistently be computed on full datasets by considering only pairs of identified spectra, which of course are to be found also in the full datasets.

Fig. 2(C) and Table 6 show that all algorithms performed worse on the full datasets than they did on the reference datasets. While this is only to be expected, given that the full datasets contain a lot of noisy spectra, examining the table entries individually indicates substantial

Table 6
Algorithm RMSEs on full datasets for parameters optimized by GBTSP on reference datasets. Percentages in parentheses indicate growth relative to the reference datasets.

Algorithm (A)	$RMSE_A$(JI)	$RMSE_A$(MPS)	$RMSE_A$(DI)
O-Cluster	0.439(+6.55%)	0.220(+134%)	0.025(+4.17%)
hclust	0.514(+16.8%)	0.120(+48.1%)	0.934(+39.6%)
CAST	0.476(+15.5%)	0.189(+70.3%)	0.934(+28.8%)
N-Cluster	0.492(+15.5%)	0.257(+68.0%)	0.931(+10.4%)
igraph	0.529(+22.4%)	0.417(+124%)	0.934(+37.8%)
DBSCAN	0.871(+44.4%)	0.762(+76.0%)	0.931(+1.42%)
PRIDE Cluster	0.499(+4.39%)	0.253(+124%)	0.945(+14.7%)
MS-Cluster	0.395(+1.54%)	0.699(+497%)	0.938(+20.1%)

variation in how worse performance really was. The best JI was obtained by MS-Cluster, with almost no change in terms of proportion of true similar pairs. Hclust scored the best MPS and is therefore expected to have produced purer clusters than the other algorithms. O-Cluster managed to maintain an excellent ballance between inter-cluster and intra-cluster dissimilarities, as shown by its DI score. Note also that this was the only algorithm to hold a less-than-10% RMSE increase for two scores. Fig. 2(D) gives the plots for the corresponding dataset RMSEs.

Detailed assessment measures for each algorithm and each full dataset, as well as the corresponding RMSEs, are available in Supporting Table S4.

4. Conclusion

Clustering algorithms are not perfect, nor are clustering assessment measures. There are always trade-offs to be considered when selecting the latter and consequently when selecting the former. The usefulness of an assessment measure and of the algorithms it leads to depends on the intended application. Ultimately, it is the application that determines whether a partition is good or not. Here we introduced a new partition assessment framework, one in which the user can control the selection bias, aiming toward a partition of the data in which the number of clusters is close to the number of peptide ion species in the data. This has involved the choice of an appropriate hypothesis on the distribution of partitions (H_{num}), of a relatively bias-free assessment measure when randomness is the main force driving the generation of partitions ($S_{\text{num}}(a)$), and finally of a strongly biased primary assessment measure (GBTSP). GBTSP, in particular, was tasked with optimizing a clustering algorithm's similarity-threshold parameter, which is to be carried out on a reference version of the dataset of interest, one in which only identified spectra appear.

While depending on algorithm and dataset we were able to show good performance on the reference datasets, given the optimized parameter, the more important test was to use the same optimized versions of the algorithms on the substantially noisier, full datasets. To judge by the assessment measures we had for use in this case, most of them internal (and thus inherently self-referential), performance losses were observed for all algorithms on all datasets. Such losses were not uniform across all cases and even pointed to one of the algorithms, O-Cluster, as having incurred only modest RMSE increases in two of the three assessment measures used. Something like this occurred also for DBSCAN, PRIDE Cluster, and MS-Cluster, now relative to one of the three assessment measures. This seems to indicate that our framework has shown at least one of the ways in which leveraging the power of the selection bias can help achieve high-quality partitions. In principle, this can be expected to hold both inside and outside proteomics.

Conflict of interests

The authors declare to have no conflict of interests.

Acknowledgements

A.R.F.S., L.U.K., M.D.M.S. and V.C.B. acknowledge support from Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES). A.R.F.S. acknowledges Instituto Carlos Chagas (ICC). D.B.L., M.D. and J. C.R. acknowledge Institut Pasteur, CNRS, the Agence Nationale de la Recherche (project ANR-15-CE18-0021) and the European Joint Programme One Health EJP from the European Union's Horizon 2020 research and innovation programme (Grant Agreement 773830) for financial support. R.H.V., V.C.B. and P.C.C. acknowledge support from Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq). V.C.B. acknowledges a BBP grant from Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ). P.C. C. acknowledges Ideias Inovadoras Fiocruz and Proep ICC. The authors thank C. Malosse and M. Duchateau from the Mass Spectrometry for Biology Unit at the Institut Pasteur for their help in sample preparation and LC-MS/MS analysis.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jprot.2021.104282.

References

[1] S. Monti, P. Tamayo, J. Mesirov, T. Golub, Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data, Mach. Learn. 52 (2003) 91–118, https://doi.org/10.1023/A:1023949509487.
[2] R. Unnikrishnan, C. Pantofaru, M. Hebert, Toward objective evaluation of image segmentation algorithms, IEEE Trans. Pattern Anal. Mach. Intell. 29 (2007) 929–944, https://doi.org/10.1109/TPAMI.2007.1046.
[3] J. Xie, S. Kelley, B.K. Szymanski, Overlapping community detection in networks: the state-of-the-art and comparative study, ACM Comput. Surv. 45 (2013), https://doi.org/10.1145/2501654.2501657, 43:1–43:35.
[4] M.J. KENNARD, B.J. PUSEY, J.D. OLDEN, S.J. MACKAY, J.L. STEIN, N. MARSH, Classification of natural flow regimes in Australia to support environmental flow management, Freshw. Biol. 55 (2010) 171–193, https://doi.org/10.1111/j.1365-2427.2009.02307.x.
[5] M. Templ, P. Filzmoser, C. Reimann, Cluster analysis applied to regional geochemical data: problems and possibilities, Appl. Geochem. 23 (2008) 2198–2213, https://doi.org/10.1016/j.apgeochem.2008.03.004.
[6] K. Helsen, K. Jedidi, W.S. DeSarbo, A new approach to country segmentation utilizing multinational diffusion patterns, J. Mark. 57 (1993) 60, https://doi.org/10.2307/1252219.
[7] E.M. Pothos, N. Chater, A simplicity principle in unsupervised human categorization, Cogn. Sci. 26 (2002) 303–343, https://doi.org/10.1207/s15516709cog2603_6.
[8] M. The, L. Käll, MaRaCluster: a fragment rarity metric for clustering fragment spectra in shotgun proteomics, J. Proteome Res. 15 (2016) 713–720, https://doi.org/10.1021/acs.jproteome.5b00749.
[9] A.M. Frank, N. Bandeira, Z. Shen, S. Tanner, S.P. Briggs, R.D. Smith, P.A. Pevzner, Clustering millions of tandem mass spectra, J. Proteome Res. 7 (2008) 113–122, https://doi.org/10.1021/pr070361e.
[10] V. Rieder, K.U. Schork, L. Kerschke, B. Blank-Landeshammer, A. Sickmann, J. Rahnenführer, Comparison and evaluation of clustering algorithms for tandem mass spectra, J. Proteome Res. 16 (2017) 4035–4044, https://doi.org/10.1021/acs.jproteome.7b00427.
[11] A.R.F. Silva, D.B. Lima, A. Leyva, R. Duran, C. Batthyany, P.F. Aquino, J.C. Leal, J. E. Rodriguez, G.B. Domont, M.D.M. Santos, J. Chamot-Rooke, V.C. Barbosa, P. C. Carvalho, DiagnoProt: a tool for discovery of new molecules by mass spectrometry, Bioinformatics. 33 (2017) 1883–1885, https://doi.org/10.1093/bioinformatics/btx093.
[12] J. Griss, Y. Perez-Riverol, S. Lewis, D.L. Tabb, J.A. Dianes, N. Del-Toro, M. Rurik, M.W. Walzer, O. Kohlbacher, H. Hermjakob, R. Wang, J.A. Vizcaíno, Recognizing millions of consistently unidentified spectra across hundreds of shotgun proteomics datasets, Nat. Methods 13 (2016) 651–656, https://doi.org/10.1038/nmeth.3902.
[13] W.M. Rand, Objective criteria for the evaluation of clustering methods, J. Am. Stat. Assoc. 66 (1971) 846–850, https://doi.org/10.1080/01621459.1971.10482356.
[14] E.B. Fowlkes, C.L. Mallows, A method for comparing two hierarchical clusterings, J. Am. Stat. Assoc. 78 (1983) 553, https://doi.org/10.2307/2288117.
[15] D.L. Wallace, Comment, J. Am. Stat. Assoc. 78 (1983) 569–576, https://doi.org/10.1080/01621459.1983.10478009.
[16] L. Hubert, P. Arabie, Comparing partitions, J. Classif. 2 (1985) 193–218, https://doi.org/10.1007/BF01908075.
[17] J. Cohen, A coefficient of agreement for nominal scales, Educ. Psychol. Meas. 20 (1960) 37–46, https://doi.org/10.1177/001316446002000104.

[18] M.J. Warrens, On the equivalence of Cohen's kappa and the hubert-arabie adjusted rand index, J. Classif. 25 (2008) 177–183, https://doi.org/10.1007/s00357-008-9023-7.
[19] A.J. Gates, Y.-Y. Ahn, The impact of random models on clustering similarity, J. Mach. Learn. Res. 18 (2017) 1–28.
[20] S. Romano, N.X. Vinh, J. Bailey, K. Verspoor, Adjusting for chance clustering comparison measures, J. Mach. Learn. Res. 17 (2016) 1–32.
[21] A. Amelio, C. Pizzuti, Is normalized mutual information a fair measure for comparing community detection methods?, in: 2015 IEEEACM Int. Conf. Adv. Soc. Netw. Anal. Min. ASONAM, 2015, pp. 1584–1585.
[22] Y. Perez-Riverol, J.A. Vizcaíno, J. Griss, Future prospects of spectral clustering approaches in proteomics, PROTEOMICS. 18 (2018) 1700454, https://doi.org/10.1002/pmic.201700454.
[23] M.D.M. Santos, A.C. Camillo-Andrade, L.U. Kurt, M.A. Clasen, E. Lyra, F.C. Gozzo, M. Batista, R.H. Valente, G.V.F. Brunoro, V.C. Barbosa, J.S.G. Fischer, P. C. Carvalho, Mixed-data acquisition: next-generation quantitative proteomics data acquisition, J. Proteome 222 (2020) 103803, https://doi.org/10.1016/j.jprot.2020.103803.
[24] Y. Perez-Riverol, J.A. Vizcaíno, Synthetic human proteomes for accelerating protein research, Nat. Methods 14 (2017) 240–242, https://doi.org/10.1038/nmeth.4191.
[25] D.P. Zolg, M. Wilhelm, K. Schnatbaum, J. Zerweck, T. Knaute, B. Delanghe, D. J. Bailey, S. Gessulat, H.-C. Ehrlich, M. Weininger, P. Yu, J. Schlegl, K. Kramer, T. Schmidt, U. Kusebauch, E.W. Deutsch, R. Aebersold, R.L. Moritz, H. Wenschuh, T. Moehring, S. Aiche, A. Huhmer, U. Reimer, B. Kuster, Building ProteomeTools based on a complete synthetic human proteome, Nat. Methods 14 (2017) 259–262, https://doi.org/10.1038/nmeth.4153.
[26] J.R. Yates 3rd, S.K.R. Park, C.M. Delahunty, T. Xu, J.N. Savas, D. Cociorva, P. C. Carvalho, Toward objective evaluation of proteomic algorithms, Nat. Methods 9 (2012) 455–456, https://doi.org/10.1038/nmeth.1983.
[27] J.C. Dunn, A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters, J. Cybern. 3 (1973) 32–57, https://doi.org/10.1080/01969727308546046.
[28] J.C. Dunn, Well-separated clusters and optimal fuzzy partitions, J. Cybern. 4 (1974) 95–104, https://doi.org/10.1080/01969727408546059.
[29] P.C. Carvalho, D.B. Lima, F.V. Leprevost, M.D.M. Santos, J.S.G. Fischer, P. F. Aquino, J.J. Moresco, J.R. Yates, V.C. Barbosa, Integrated analysis of shotgun proteomic data with PatternLab for proteomics 4.0, Nat. Protoc. 11 (2015) 102–117, https://doi.org/10.1038/nprot.2015.133.
[30] J.K. Eng, M.R. Hoopmann, T.A. Jahan, J.D. Egertson, W.S. Noble, M.J. MacCoss, A deeper look into comet–implementation and features, J. Am. Soc. Mass Spectrom. 26 (2015) 1865–1874, https://doi.org/10.1007/s13361-015-1179-x.
[31] P.C. Carvalho, J.S.G. Fischer, T. Xu, D. Cociorva, T.S. Balbuena, R.H. Valente, J. Perales, J.R. Yates 3rd, V.C. Barbosa, Search engine processor: filtering and organizing peptide spectrum matches, Proteomics. 12 (2012) 944–949, https://doi.org/10.1002/pmic.201100529.
[32] R. Barboza, D. Cociorva, T. Xu, V.C. Barbosa, J. Perales, R.H. Valente, F.M. G. França, J.R. Yates 3rd, P.C. Carvalho, Can the false-discovery rate be misleading? Proteomics. 11 (2011) 4105–4108, https://doi.org/10.1002/pmic.201100297.
[33] T. Hastie, R. Tibshirani, J.H. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed, Springer, New York, NY, 2009.
[34] A. Ben-Dor, R. Shamir, Z. Yakhini, Clustering gene expression patterns, J. Comput. Biol. J. Comput. Mol. Cell Biol. 6 (1999) 281–297, https://doi.org/10.1089/106652799318274.
[35] K.U. Schork, Verbesserte Annotation von Massenspektren mit Algorithmen der Clusteranalyse, TU Dortmund University, 2016.
[36] E.D. Kolaczyk, G. Csárdi, Descriptive analysis of network graph characteristics, in: Stat. Anal. Netw. Data R, Springer New York, New York, NY, 2014, pp. 43–67, https://doi.org/10.1007/978-1-4939-0983-4_4.
[37] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, A density-based algorithm for discovering clusters in large spatial databases with noise, in: Proc. Second Int. Conf. Knowl. Discov. Data Min, AAAI Press, 1996, pp. 226–231.
[38] J. Griss, J.M. Foster, H. Hermjakob, J.A. Vizcaíno, PRIDE Cluster: building a consensus of proteomics data, Nat. Methods 10 (2013) 95–96, https://doi.org/10.1038/nmeth.2343.
[39] S. Na, E. Paek, Quality assessment of tandem mass spectra based on cumulative intensity normalization, J. Proteome Res. 5 (2006) 3241–3248, https://doi.org/10.1021/pr0603248.