PRiSM: A prototype for exhaustive, restriction-free database searching for mass spectrometry-based identification

Peer-reviewed author version

VAN HOUTVEN, Joris; Boonen, Kurt; Geert Baggerman, |; Askenazi, Manor; Laukens, Kris; HOOYBERGHS, Jef & VALKENBORG, Dirk (2020) PRiSM: A prototype for exhaustive, restriction-free database searching for mass spectrometry-based identification. In: Rapid communications in mass spectrometry,.

DOI: 10.1002/rcm.8962
Handle: http://hdl.handle.net/1942/32829

Supporting Information: PRiSM: exhaustive, restriction-free database searching for mass spectrometry-based identification.

Joris Van Houtven,[†,‡,¶] Kurt Boonen,[†,¶] Geert Baggerman,[†,¶] Manor Askenazi,[§] Kris Laukens,[‖,⊥] Jef Hooyberghs,[†,#] and Dirk Valkenborg[*,‡,¶,†]

[†]*VITO NV, Applied Bio & molecular Systems, Boeretang 200, Mol, BE 2400*

[‡]*Universiteit Hasselt, Data Science Institute (DSI), Interuniversity Institute for Biostatistics and Statistical Bioinformatics (I-BioStat), Agoralaan, Diepenbeek, BE 3590*

[¶]*Universiteit Antwerpen, Centre for Proteomics, Groenenborgerlaan 171, Antwerpen, BE 2020*

[§]*Biomedical Hosting LLC*

[‖]*Universiteit Antwerpen, Biomedical Informatics Network Antwerp (Biomina), Middelheimlaan 1, Antwerpen, BE 2020*

[⊥]*Universiteit Antwerpen, ADReM Data Lab, Department of Computer Sciences, Middelheimlaan 1, Antwerpen, BE 2020*

[#]*Universiteit Hasselt, Data Science Institute (DSI), Theoretical Physics, Agoralaan, Diepenbeek, BE 3590*

E-mail: dirk.valkenborg@uhasselt.be

The supporting information has the following contents

The `chargeSep` filter

From the main text: *”The `chargeSep` filter tries to infer the peaks’ charge by recognizing the charge state of its isotopic envelopen. If successful, the mono-isotopic peaks are only taken into consideration when comparing with the mass pattern representation of that particular charge state, and the others are removed altogether.”*

The recognition of an isotopic envelope of charge Z is done by running the very same CPC algorithm to compare the spectrum with mock spectrum $[0, 1/Z]$. Whenever two observed spectrum peaks match this pattern, they are considered possible isotopes from an envelope with charge Z. The processing of the scores of all the alignments is done in such a way that any peak is only used in a comparison of charge Z, if it appears as the mono-isotopic peak in an isotopic envelope (i.e., the first of its kind in a sequence of isotopes) *or* if it never appears in any isotopic envelope of any charge (because then its charge cannot be determined).

Filters and blind spots

Two of the filters we use (`multiDetections`, `chargeSep`) are *not* entirely harmless. In fact, they may both reduce the sensitivity through the presence of a noise peak (which, in case of a mixed spectrum, may even correspond to a signal peak from a co-eluting peptide). The multiDetections filter may accidently use such noise peaks to unjustly shift nearby signal peaks outside of their match window. The `chargeSep` filter may use such noise peaks to wrongly classifying a nearby mono-isotopic fragment as non-mono-isotopic, thus removing it from the match window.

So strictly speaking, these filters are actually heuristics – which we intend to avoid – that negatively affect sensitivity locally (by *unjustly* reducing a small amount of CPC alignment scores), but their net effect on global sensitivity (by *justly* reducing a large amount of CPC scores) and runtime is so overwhelmingly positive that we decided to keep them in a first approximation. Moreover, both of these errors can only cause a decrease in the amount of matching fragments, so there is no risk of introducing false positives, though of course there is a risk of introducing blind spots. This risk should be evaluated in a future version of PRiSM where the burden of statistical power can be shifted away from these filters towards improvements in other components.

Actual CPC implementation

The actual CPC implementation is different from the one lined out in the main text, but equivalent: first, a matrix of pairwise distances between the fragment peaks (primary index) and the mass pattern (secondary index) is constructed. Immediately after, the

$\Delta m/z$-values are sorted into an array, but the intermediate matrix structure allows to trace back the identity of the peaks that generated a specific mass shift. Then, a dedicated in-house 1D-clustering algorithm is used on the array to identify clusters with a maximum width equal to twice the fragment tolerance. Different clusters may overlap but can never be a strict subcluster of another cluster. Each cluster with S elements then corresponds to a distinct alignment where S peaks are matching within a margin equal to the fragment tolerance. The final $\Delta m/z$ value associated with a cluster is chosen as the middle between the largest and smallest value in that cluster, which ensures that all values in the cluster deviate no further from this final value than the fragment mass tolerance.

Motifs and the negative binomial fit

To the best of our knowledge, the empirical null distributions are quite well characterized by negative binomial distributions, except for spectra containing motifs or polymers or AA repeats in general. Polymers create a mixture of fragment ladders, each with a different fixed mass difference, which together may be interpreted as repeats of some amino acid which by chance happens to match the mass difference between two fragments of any two ladders. AA repeats on the other hand are genuinely present in the spectrum, but both behave just like motifs. The issue with spectra containing motifs is that they will - by definition - match with a disproportionate amount of (protein) sequences in the database. Hence, their empirical distribution of PIDAS will be disproportionately heavy-tailed and thus deviate from the corresponding negative binomial distribution *NB* which will underestimate the amount of high-scoring alignments, as shown in Figure S1.

However, this need not deter us from using PRiSM since this is a known issue which state-of-the-art search engines also suffer from and moreover we developed an automated method to detect such cases and discard them from the results. To prevent this phenomenon from causing too many false hits, we developed a method to flag such spectra by measuring how much *NB* underestimates the empirical distribution.

First, we measure the *SPFVU* (sign-preserved fraction of variance unexplained) and *UFVU* (underestimating fraction of variance unexplained) of an empirical null distributions d as

$$\begin{aligned} \textit{SPFVU}_d &= \sum_P \frac{\epsilon_{d,P}^2}{\lambda_d} \text{sign}(\epsilon_i) && (1) \\ \textit{UFVU}_d &= \sum_P \frac{\epsilon_{d,P}^2}{\lambda_d} \mathcal{H}(\epsilon_{d,P}) && (2) \end{aligned}$$

where the ϵ_P are the differences between the observed and theoretical number of occurrences of PIDAS values P on the log scale, λ_d is the variance of the empirical occurrences(P) (also log scale), and $\mathcal{H}$ is the Heaviside step function. Note that only those PIDAS values P where occurrences$(P) > 0 \forall i < P$ may contribute to *SPFVU*, *UFVU* and λ_d in order to avoid capturing the contributions of outliers which do not originate from the true null distribution. Note also that the theoretical number of occurrences on log scale is set to 0 whenever the actual value on the regular scale is smaller than 0.5, in order not to disproportionately penalize such cases where the theoretical number of occurrences on log scale would run off to $-\infty$.

Secondly, we determine a threshold *UFVU** by looking at the distribution of all *UFVU* values that do not correspond to cases of an strictly over-estimating *NB*, meaning we look only at cases where *SPFVU* > 0 OR *UFVU* $>$ median$($*UFVU*$)$. The second member of the OR-clause is necessary because we want to take cases into account where there is both under- and overestimation by the *NB* fit which causes *SPFVU* < 0 while the *UFVU* is still of considerable magnitude. The resulting distribution of *UFVU* values is used to determine outliers by means of boxplot whiskers (excluding 0.35% on either side in case of normally distributed data). However, since the distribution is very right-skewed – most values lie close to zero, indicating the *NB* fits well – we do not use regular boxplot whiskers but rather the ones based on the medcouple statistic *MC* for skewed distributions[S1]. Thus, we obtain

$$\mathit{UFVU}^{*} = 1.5 \cdot IQR \cdot e^{3\mathit{MC}} \tag{3}$$

where IQR is the inter-quartile range.

Any null distribution d with *UFVU*$_d >$ *UFVU** is flagged and subsequently discarded, though the corresponding spectrum need not be unidentifiable if another one of its distributions corresponding to another ion type was not flagged. If the proportion of spectra representing motifs or polymers becomes large (more than, say, 1%) they run a risk of not being discarded by this procedure, resulting in an increase in false discoveries. At present, therefore, this is an additional hidden requirement for reliably using PRiSM . In the future, we will further develop PRiSM to prevent this possibility as well as rigorously remedy flagged cases and try to identify them correctly.

PIDAS and filters curb random matches

This section is quite complex and intricately takes multiple subtle observations to paint a detailed, consistent picture of the scoring of a CID tandem-MS experiment using PIDAS, as opposed to using regular peak counting.

Figure S3 displays the highest PIDAS score observed in each combination, aggregated across all proteins *and* spectra from the Closed data set, but split according to charge (Figure S3a) or orientation (Figure S3b) of the pseudo-ion series. It shows that mass pattern representations with lower charge states are more prone to generating ’best alignments’ with high scores, and also the same can be said for the C-terminal ones as opposed to their N-terminal counterparts. Note that the latter happens despite the fact that the means of the boxplots in Figure S3b are approximately equal.

These phenomena seem to replicate two ‘commonly known facts’ in mass spectrometry. Namely, they affirm that C-terminal ions in CID experiments – mostly y-ions – are more abundant than their N-terminal counterparts (mostly b-ions), and that ions with lower charges are more abundant than highly charged ions. These two affirmations follow from the fact that the extent of the (upper) tail of each distribution of maxPIDAS values acts as a proxy for the abundance of non-trivial scores: the more observations in your distribution, the further its tail extends. The fact that the means of the boxplots in Figure S3b are equal also affirms the idea that the driving process behind such score distributions is a random process – which does not distinguish between N-terminal and C-terminal ions.

Interestingly, however, Figure S2 (made without using filters) suggests that pseudo-ion series with a higher charge state increase the average number of matching peaks in an alignment, but decrease the number of matching peaks in the very best alignments. A plausible explanation is that the density of peaks in a pseudo-ion series – and therefore also the chance of random matching peaks – is directly proportional to its charge state. The protein mass patterns are much 'longer' than our observed spectra, so when its charge state is increased a bigger portion gets 'squeezed' into the m/z range of an observed spectrum, thus increasing the mean alignment score. The highest observed score on the other hand is usually not determined by the random matching procedure, but rather by whether the AA of the spectrum actually originates from the protein it is being compared to. The higest scores tend to decrease with charge, consistent with the fact that in most spectra low fragment charge states are more abundant than high charge states.

As such, the introduction of filters (only partly confounded with the use of PIDAS) significantly reduced the occurrence of random matches, as Figure S3a shows a monotonically decreasing trend instead. Moreover, Figure S3b also shows a larger discrepancy in the most extreme scores between b- and y-ions. Since in this type of experiment C-terminal ions (y-ions) are slightly more abundant (the COOH group is more likely to retain electrical charge during CID) this seems to confirm that fragment peak matching is now dominant over random peak matching, especially for the best alignments. Still, however, the mean maxPIDAS values are equal because the driving process behind most combinations is random matching, since only one or at most a handful of mass patterns in the database is expected to truly be represented by the observed spectrum.

More future work

Some (but not all) additional future improvements which are less prominent are listed below.

- During protein assignment, score ties – which maybe frequent for homologous proteins or mixed spectra – should (and can) be investigated automatically without requiring manual curation.
- We greatly reduce the output size of each comparison by rounding the scores to get a histogram to which we fit the null distribution, but we can instead use a density estimation technique.
- Remove the additional hidden requirement that no substantial proportion (say, more than 1%) of observed spectra in a PRiSM search represents motifs. This can be done by using a data-independent goodness-of-fit statistic when fitting the null distribution.
- Benchmark PRiSM on data with very reliable identifications e.g.[S2]'s synthetic peptide data set, in order to confirm that we have no blind spots and are covering the entire search space, and in order to test the PTM detection mechanism.
- Evaluate some heuristics commonly used by state-of-the-art search engines to evaluate and report on their performance usefulness.

Possible heuristics

We expect some heuristics listed below to be very effective, i.e., sacrifice very little or no sensitivity and specificity at all in exchange for an increase in computation time. Not that they may distort the null distribution, so first additional research in that area is necessary.

- Tag-based filtering to reduce candidates proteins in the database (and even restrict matches to specific protein regions). They have a natural advantage[S3] over mass-based approaches.

- Use a small portion of spectra to iteratively calibrate an intensity filter for the observed spectra.

- Use a multi-stage scoring system, f.i. by limiting $\Delta m/z$ to regions where most matches occur and gradually relax the constraint until significant matches are found.

Figures

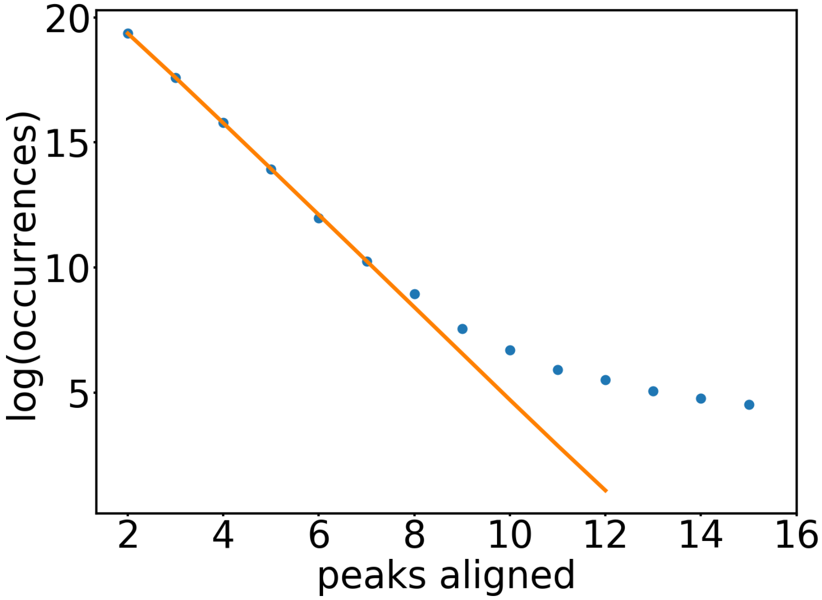
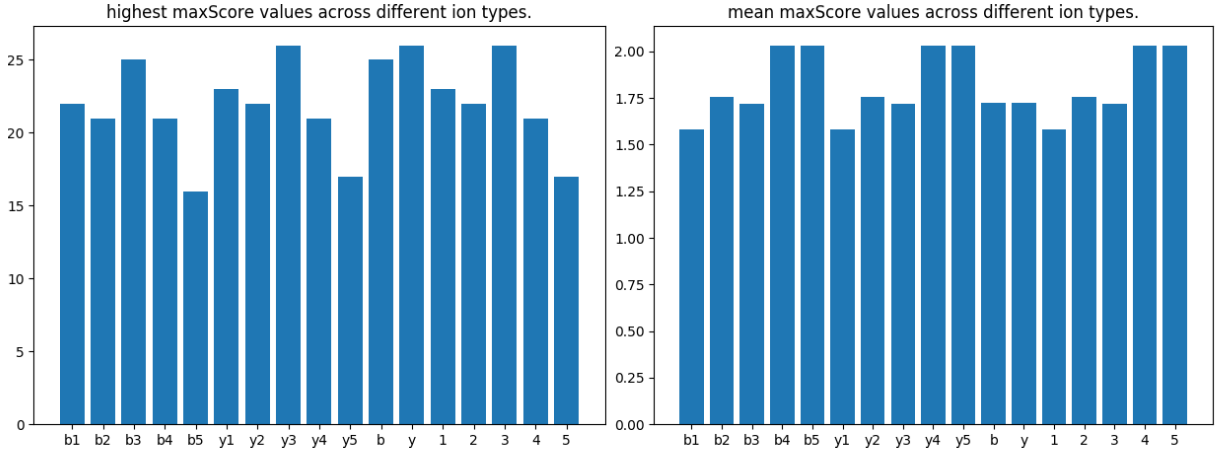


Figure S1: The fitted negative binomial (orange curve) clearly underestimates the behaviour of the empirical null distribution (blue dots), which is heavy-tailed because the associated spectrum contains a repeating sequence of 11 Alanines.



(a) Highest maxScore *decreases* with the charge of the mass pattern, regardless of terminal orientation. The highest maxScore of patterns with charge 3 is anomalously high.

(b) Mean maxScore *increases* with the charge of the mass pattern, regardless of terminal orientation. The mean maxScore of patterns with charges 4 and 5 is anomalously high.

Figure S2: Highest (left) and mean (right) maxScores highlight systematic trends when split out per mass pattern types. Note that here, maxScore refers not to the highest *PIDAS* but rather the highest *number of matching peaks* for a particular combination, and that these are results aggregated for all protein mass patterns *and* osberved spectra from a PRiSM run on the Closed data, where input filters were disabled.

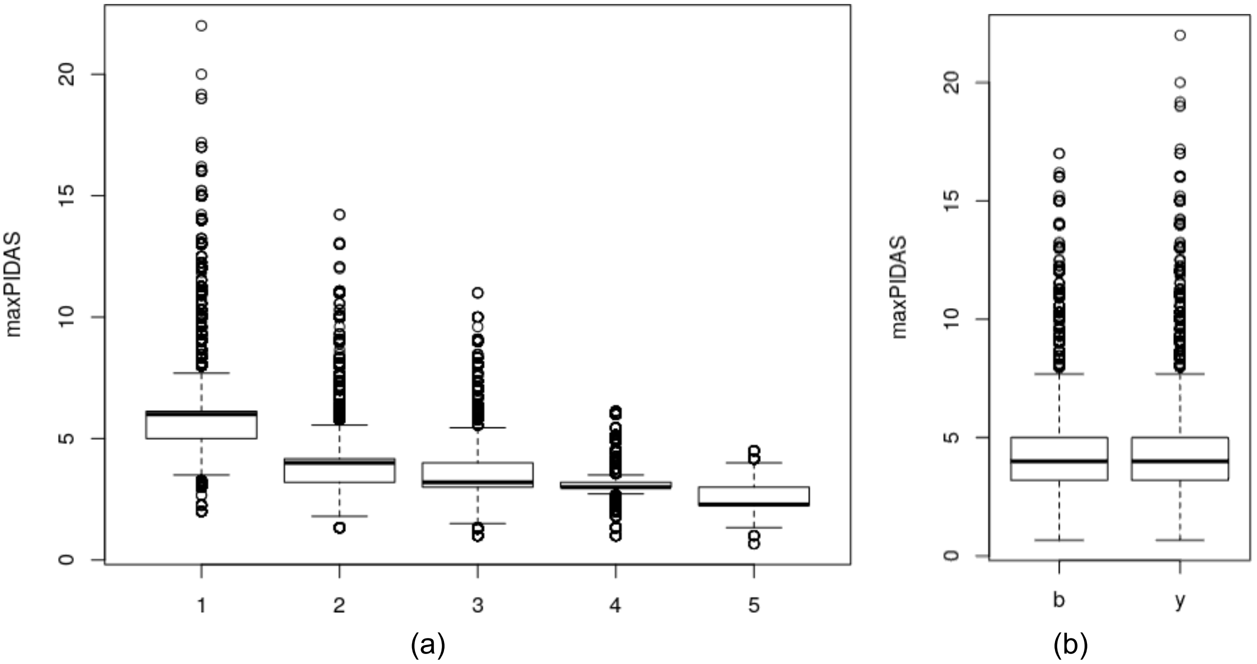
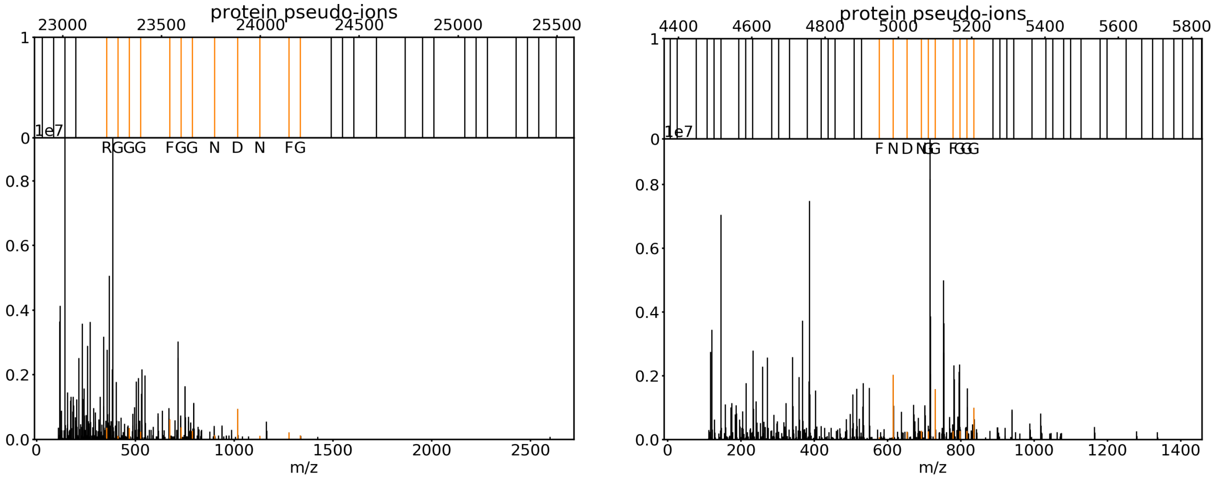


Figure S3: When using filters and PIDAS, the influence of random matching observed in Figure S2 is reduced. In (a) the mean maxPIDAS decreases monotonically with charge. In (b) the mean PIDAS is still identical for both terminal orientations, but matching an extremely high amount of y-ions is more probable than for b-ions.



(a) Suspicious alignment of spectrum 9247_43733 with mass pattern (P09651, N1).

(b) Suspicious alignment of spectrum 9260_44559 with mass pattern (P56696, C4).

Figure S4: Two significantly identified spectra from the Unidentified data appear to be false positives by visual inspection. Many matched peaks have a very low intensity, they do not span the lower end of the spectrum's m/z range, and from experience with the Closed results we know that N-terminal as well as highly charged patterns usually do not produce the best (non-random) alignments.

Additional files

1. Closed data information on spectra where PRiSM and SEQUEST disagree on the protein identification: `Closed_disagreement_information.txt`
2. Unidentified data spectral alignment plots: `Unidentified_results_annotated_spectra.pdf`

References

(S1) Hubert, M.; Vandervieren, E. An adjusted boxplot for skewed distributions. *Computational statistics & data analysis* **2008**, *52*, 5186–5201.

(S2) Marx, H.; Lemeer, S.; Schliep, J. E.; Matheron, L.; Mohammed, S.; Cox, J.; Mann, M.; Heck, A. J.; Kuster, B. A large synthetic peptide and phosphopeptide reference library for mass spectrometry–based proteomics. *Nature biotechnology* **2013**, *31*, 557.

(S3) McHugh, L.; Arthur, J. W. Computational methods for protein identification from mass spectrometry data. *PLoS computational biology* **2008**, *4*, e12.