Published in final edited form as:

Nat Methods. 2013 February ; 10(2): 95–96. doi:10.1038/nmeth.2343.

PRIDE Cluster: building the consensus of proteomics data

Johannes Griss*, Joseph M. Foster, Henning Hermjakob, and Juan Antonio Vizcaíno
EMBL-European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK.

Correspondence

To the editor: The amount of mass spectrometry (MS) proteomics data in public repositories is growing rapidly¹ but its (re-)use to increase the reliability of newly performed experiments is still limited. Two of the major obstacles are the high heterogeneity of the data present in repositories, and the inflation of false positive identifications when combining datasets. Here we present ‘PRIDE Cluster’: a novel method to identify reliable identifications in heterogeneous MS proteomics experiments. It is used to highlight reliable peptide identifications in the PRIDE database² (<http://www.ebi.ac.uk/pride>) and generate constantly updated, reliable spectral libraries based on these identifications.

The current state of the art to estimate the false discovery rate of proteomics experiments is the target-decoy strategy³. While this approach can estimate the overall error rate of an experiment, it cannot estimate the reliability of individual peptide identifications when combining independent experiments. Search engines’ identification scores are based on different statistical models and are therefore not suited to assess identification reliabilities beyond a single experiment⁴.

‘PRIDE Cluster’ uses spectral clustering to identify reliable identifications in highly heterogeneous data, taking advantage of the wealth of data present in PRIDE. It uses a modified version of the MS-Cluster algorithm⁵, which we refined to increase the clustering quality (Supplementary Note 1). In our opinion, this was necessary to make the approach applicable to highly heterogeneous data, where the original algorithm may lead to inaccurate results (Supplementary Fig. 1). The new algorithm is freely available as a Java Application Programming Interface at <http://pride-spectra-clustering.googlecode.com> (Supplementary Note 2).

We tested our clustering algorithm using three large, highly heterogeneous datasets which we searched against a target-decoy database (Supplementary Protocol and Supplementary Note 3). The proportion of clusters that contained spectra identified as multiple different peptides proved to be too dataset-dependent for a reliable assessment of clustering quality⁵ (Supplementary Fig. 2). Instead, we assessed it by looking at the precursor ion m/z range of spectra that were clustered together and found that the algorithm was robust for every test dataset (Supplementary Note 4). We observed that larger clusters contained more spectra identified as the same peptide (Supplementary Fig. 3) and that classical search engines identified their consensus spectra more reliably (Supplementary Protocol and Supplementary

*Corresponding author. jgriss@ebi.ac.uk European Bioinformatics Institute. Wellcome Trust Genome Campus CB10 1SD, Cambridge, UK Tel: +44 (0) 1223 492686.

Author contributions J. Griss designed and implemented the algorithm, ran the experiments, performed the analysis, and developed the ‘PRIDE Cluster’ application. J. M. Foster contributed to the development of the algorithm and the data analysis. J. Griss and J. A. Vizcaíno wrote the manuscript. H. Hermjakob and J. A. Vizcaíno supervised the project. All authors discussed, commented and contributed to the final version of the manuscript.

Figs. 4 and 5). This suggested that the larger a (clustered) repository like PRIDE becomes, the more distinct reliable and unreliable clusters get.

The relative proportion of spectra within a cluster identified as the same peptide, called 'ratio' (Supplementary Fig. 6), was the most predictive attribute to distinguish between target and decoy peptide identifications (Fig. 1a and Supplementary Note 5). Furthermore, the number of target and decoy identifications with low 'ratios' was basically identical while nearly only target identifications had high 'ratios' (Supplementary Fig. 7). This is in line with the target-decoy strategy which states that incorrect identifications are evenly distributed among the target and the decoy database³. It therefore seemed conclusive that identifications with low 'ratios' represented the random incorrect identifications of low quality spectra.

Using our clustering algorithm we built a new resource called 'PRIDE Cluster' in which we clustered all identified public spectra available in PRIDE (Supplementary Fig. 8). 'PRIDE Cluster' contained 20,666,123 identified spectra from 9,040 experiments and 40 species (June 2012, Supplementary Protocol). The clustering algorithm robustly organized the data in 3,152,393 clusters; a more than fivefold reduction compared to the original number of spectra (Fig. 1b), taking approximately 601 central processing unit days. 5.7 % of the clusters contained spectra from multiple species which were either identified as contaminants or conserved sequences (Supplementary Fig. 9).

'PRIDE Cluster' currently provides two main methods to access its data: 1) retrieve all clusters that contain a given peptide identification and 2) retrieve all clusters with a consensus spectrum similar to a query spectrum (Fig. 1c). Researchers can now easily check whether similar spectra to their experimental ones were already identified in PRIDE. As a key functionality, reliable peptide identifications based on the results from 'PRIDE Cluster' (Supplementary Note 5) are now highlighted in the classical PRIDE web interface (<http://www.ebi.ac.uk/pride>). Furthermore, 'PRIDE Cluster' provides a simple method to correct inaccurate annotations in the PRIDE database (Supplementary Note 6). The consensus spectra of all reliable clusters can be downloaded as spectral libraries (<http://www.ebi.ac.uk/pride/cluster/libraries>, Supplementary Table 1), which performed comparably to the corresponding ones from the National Institute of Standards and Technology (Supplementary Note 7).

'PRIDE Cluster' is the first step to introduce stringent quality control in a highly heterogeneous MS proteomics repository and represents a constantly updated, reliable consensus of published proteomics data.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We want to acknowledge all other members of the PRIDE team, who contributed to some aspects of this work in terms of technical implementation and discussion of the results. Additional thanks go to Matrix Science for providing a Mascot license. Special thanks go to all data submitters of the PRIDE database whose data is the foundation of the here presented work. J. Griss is supported by the Wellcome Trust [grant number WT085949MA]. J. A. Vizcaino is supported by the EU FP7 grants LipidomicNet [grant number 202272] and ProteomeXchange [grant number 260558]. J. M. Foster is supported by a Biotechnology and Biological Sciences Research Council CASE studentship, also funded by Philips.

References

1. Csordas A, et al. PRIDE: quality control in a proteomics data repository. Database (Oxford). 2012; 2012 bas004.
2. Vizcaíno JA, et al. The Proteomics Identifications database: 2010 update. Nucleic Acids Res. 2010; 38:D736–742. [PubMed: 19906717]
3. Elias JE, Gygi SP. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. Nat Methods. 2007; 4:207–214. [PubMed: 17327847]
4. Gupta N, Bandeira N, Keich U, Pevzner PA. Target-decoy approach and false discovery rate: when things may go wrong. J Am Soc Mass Spectrom. 2011; 22:1111–1120. [PubMed: 21953092]
5. Frank AM, et al. Spectral archives: extending spectral libraries to analyze both identified and unidentified spectra. Nat Methods. 2011; 8:587–591. [PubMed: 21572408]

Figure 1. A peptide identification's 'ratio' was sufficient to identify reliable identifications from the clustering results of the PRIDE repository

((a) Distribution of 'ratios' (relative number of spectra within a cluster that were identified as this sequence) between target and decoy sequences. Target sequences predominantly had 'ratios' close to one. The considerably less common (incorrect) target sequences with low 'ratios' were considered as outliers in the boxplots. Decoy identifications were clearly distinguished through lower 'ratios' (detailed distributions in Supplementary Fig. 7). (b) The precursor m/z values range of spectra within one cluster from the 'PRIDE Cluster' instance. The algorithm performed equally well as in the test datasets: 91 % of the clusters contained spectra within 1 m/z unit or less. (c) 'Result view' when searching for clusters with a similar consensus spectrum as the entered one. It presents all clusters that have similar consensus spectra as the entered one including the similarity score (normalized-dot product).