¬Application Note

Phoenix Enhancer: an online service/tool for proteomics data mining using clustered spectra

Mingze Bai[1,2,*], Chunyuan Qin[2], Kunxian Shu[2], Johannes Griss[3], Yasset Perez-Riverol[3], Weimin Zhu[1] and Henning Hermjakob[1, 3]

[1] Beijing Proteome Research Center, National Center for Protein Science, Beijing, China.

[2] Chongqing Key Laboratory on Big Data for Bio Intelligence, Chongqing University of Posts and telecommunications, Chongqing.

[3] European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK.

*To whom correspondence should be addressed.

Abstract

Motivation: Spectrum clustering has been proved to enhance proteomics data analysis: some originally unidentified spectra can be potentially identified and individual peptides can also be evaluated to find potentially mis-identifications by using clusters of identified spectra. The Phoenix Enhancer spectrum service/tool provides an infrastructure to perform data analysis on tandem mass spectra and the corresponding peptides against previously identified public data. Based on previously released PRIDE Cluster data and a newly developed pipeline, four functionalities are provided: i) evaluate the original peptide identifications in an individual dataset, to find low confident peptide spectrum matches (PSMs) which could correspond to mis-identifications; ii) provide confidence scores for all originally identified PSMs, to help users to evaluate their quality (complementary to getting a global false discovery rate); iii) identified potentially new PSMs to originally unidentified spectra; and iv) provide a collection of browsing and visualization tools to analyze and export the results. The code is open-source and easy to re-deploy on local computers using Docker containers.

Availability: The service of Phoenix Enhancer is available at http://enhancer.ncpsb.org.
Contact: baimingze@gmail.com, hhe@ebi.ac.uk
Supplementary information: Supplementary data are available at online.

1 Introduction

Mass spectrometry (MS) has become the mainstream technology in proteomics research activities, consequently, the proteomics data has growing swiftly with this speedy developing of MS technology. With more and more researchers sharing their data on public data repositories like PRIDE Archive (Perez-Riverol, et al., 2019) and iProx (Ma, et al., 2019), the proteomics data became a “big data” field in the aspect of volume. By 1 September 2019, PRIDE Archive hosts up to 13389 datasets, representing nearly 94757 assays for acquiring proteomics data.

The number of unidentified spectra in public datasets (“dark matter”) is on average 75% of spectra measured in each MS experiment (Griss, et al., 2016; Perez-Riverol, et al., 2018). The main reason behind the low number of identified spectra, is that during the peptide identification step (Vaudel, et al., 2014) many derived peptides are either not present in the sequence database (e.g. sequence variants, or incomplete genome sequences) or they contain unexpected PTMs. In 2016, by clustering all PRIDE Archive data, we were able to identify 20% of previously unidentified spectra (Griss, et al., 2016).

In this manuscript, we presented the Phoenix Enhancer service/tool; a platform that enable researchers to perform meta-comparison of their identified and unidentified spectra against previously published datasets in PRIDE Cluster. Four functionalities are provided: i) evaluate the original peptide identifications in an individual dataset, to find low confident peptide spectrum matches (PSMs) which could correspond to mis-identifications; ii) provide confidence scores for originally identified PSMs, to help users to evaluate their quality (complementary to getting a global false discovery rate); iii) identified potentially new PSMs to originally unidentified spectra; and iv) provide a collection of browsing and visualization tools to analyze and export the results. All the source code are freely available in GitHub (https://github.com/phoenix-cluster/) and can be deploy in the Cloud and HPC architectures.

Design and Implementation

Phoenix Enhancer is designed for enhancing the proteomics identification by evaluating the quality of original PSMs or find new potential PSMs by searching the query spectra against the clustering results (Figure 1). It provides three core components, i.e. the front-end Web, a Restful API and data analysis pipeline. Users can upload the query files which include MS/MS spectra and/or sequences, and set the searching parameters (**Supplementary information, Note 1)**. Then, the pipeline which searches the query spectra against the spectral cluster archives and then score the previous PSMs and/or the new recommend PSMs, as well as writing the scored PSMs into the MySQL database. The Restful-API or Phoenix Enhancer web can be used to retrieve or browse the final results. The analysis pipeline informed three major reports: i) potential incorrect identifications with a confidence score and a new recommending sequence if possible; ii) new identified PSMs for previous unidentified spectra; iii) high confident previous identifications which also got a high-confidence score.

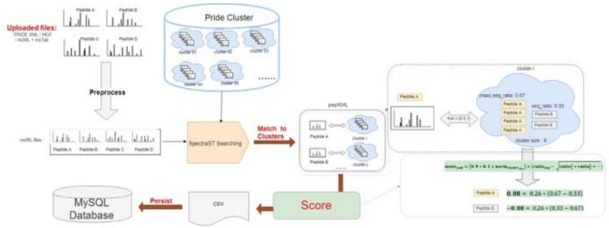

Figure 1: The Phoenix Enhancer Pipeline is performing the actual analyzing of the incoming spectra data by matching them to spectral cluster archive and scoring the matches.

Data analysis pipeline

When a file containing the MS/MS spectra with/without identifications is uploaded, the Phoenix Enhancer pipeline converts the files to mzML; the MS/MS spectra is search against the PRIDE cluster archives using the SpectraST (Lam, et al., 2007) in "archive searching model" (**Supplementary information, Note 2**). After picking the matches whose "fval" scores are higher or equal than SpectraST's default threshold (0.5), Phoenix Enhancer then calculates the confidence scores for the 3 types of matched spectra/PSMs: previous PSMs get confidence scores for their previous assigned peptides; spectra in low confident PSMs or unidentified group will get a new recommend peptides with confidence scores too, the new coming sequences are the dominate sequences in matched clusters.

Confidence score can be used to assess the quality of the PSMs and to help the users finding the novel and new PSMs. The confident score is calculated based on these rules: i) the bigger the cluster size is, the higher score is getting, since it is less important than the cluster ratio, so we chose (0.9 + 0.1* normalized size of the cluster) to measure it's contribution to the final score (using a cut off at 1000 to get highest value); ii) for each sequence in a cluster, if it's ratio is higher than 0.5, then gets positive score; if lower than 0.5 gets negative score; iii) the distribution of the other sequences' ratios also matters: more other sequences, or more equally they distributed both means the major sequence has higher confidence. **Equation 1** is used to compute the score; and the square root is used to measure how much the rest of the cluster is dominated by a few sequences: this part attains highest value ratio1 for the case when the cluster only have one sequence in the others, and a lowest value when the contribution to the score is evenly distributed among all ***n*** sequences with an *equal_ratio*, and get down with higher number of sequences *n*: $\sqrt{n * equal\_ratio^2}$ ·

$$\mathrm{score}_{conf} = \left(0.9 + 0.1 * norm_{cluster_{size}}\right) * \left(ratio_{seq} - \sqrt{ratio_1^2 + ratio_2^2 + \ldots}\right)$$ **Equation 1**.

Restful API and Web

The mission of the Web and Web Service is to provide interface to do the analysis and access to the analysis results: i) upload files for analysis, set analysis parameters; ii) show the results PSMs in tables and charts (**Supplementary information, Note 3**); iii) filter the results using species; iv) compare the spectra pairs of query spectrum and matched cluster consensus spectrum (**Figure 2**); (v) checking the details of the matched cluster, includes comparing the spectra inside a cluster to its consensus spectrum; vi) download analysis result files for further analyses.

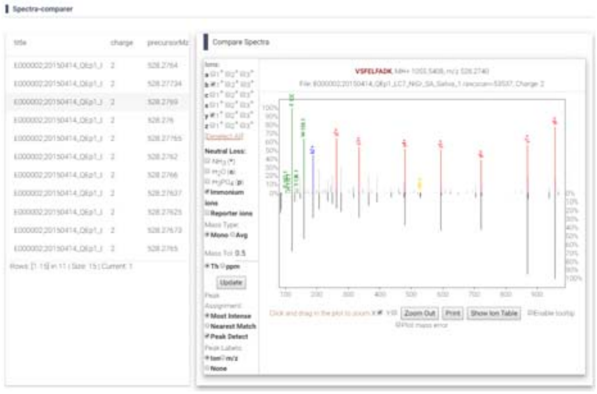

Figure 2: Spectra Comparer, for comparing the spectra pairs of query spectrum and matched cluster consensus spectrum.

The complete source code from the Pipeline, Web and the Web Service is available on GitHub (https://github.com/phoenix-cluster). We provide our four components (Web, Web service, Pipeline and MySQL server) as BioContainers images at Docker Hub (da Veiga Leprevost, et al., 2017).

Benchmark datasets

We tested the accuracy of the Phoenix Enhancer's pipeline using 71 datasets from PRIDE Archive, 39 out of 71 are already included in PRIDE Cluster (**Supplementary information, Note 4**). We found than less than 1% of the peptides is incorrectly identified (**Supplementary information, Note 4**) and the total error rate is 0.172% among the 39 datasets (already included in PRIDE Cluster). In addition, we have found that the number of identifications ranges from 0 to 167% for the 71 datasets (**Supplementary information, Note 4**). Interestedly, for datasets PXD000529, PXD000533, PXD000535; the previously submissions 290 distinct peptides with phosphorylation sites; while we can identify 1027 distinct new peptides with phosphorylation sites.

Conclusion

In summary, we believe that Phoenix Enhancer is competent to take the advantage of spectral clustering results to dig more information in proteomics data analysis, especially in validating the confidence of specific peptide biomarkers, and in finding interesting new potential biomarkers in repository datasets. The proposed Phoenix Enhancer is an easy-to-deploy and reuse in local HPC and Cloud environments. The framework will enable smaller and less computing consuming research effort to utilize the data mining potential on repository scale spectral clusters, and in doing better visualization and manual inspection into an easily retrievable manner for data analysis and validation.

Acknowledgments

This work was funded by the State Key Laboratory of Proteomics [SKLP-K201705] and National Key Research and Development Program of China (2017YFA0505002 and 2017YFC0906602].

References

da Veiga Leprevost, F., *et al.* BioContainers: an open-source and community-driven framework for software standardization. *Bioinformatics* 2017;33(16):2580-2582.

Griss, J., *et al.* Recognizing millions of consistently unidentified spectra across hundreds of shotgun proteomics datasets. *Nat Methods* 2016;13(8):651-656.

Lam, H., *et al.* Development and validation of a spectral library searching method for peptide identification from MS/MS. *Proteomics* 2007;7(5):655-667.

Ma, J., *et al.* iProX: an integrated proteome resource. *Nucleic Acids Res* 2019;47(D1):D1211-D1217.

Perez-Riverol, Y., *et al.* The PRIDE database and related tools and resources in 2019: improving support for quantification data. *Nucleic Acids Res* 2019;47(D1):D442-D450.

Perez-Riverol, Y., Vizcaino, J.A. and Griss, J. Future Prospects of Spectral Clustering Approaches in Proteomics. *Proteomics* 2018;18(14):e1700454.

Vaudel, M., *et al.* Shedding light on black boxes in protein identification. *Proteomics* 2014;14(9):1001-1005.