Text mining and computational chemistry reveal trends in applications of laser desorption/ionization techniques to small molecules

Bergman, N.P.; Bergquist, J.; Hedeland, M.; Palmblad, M.

Taylor & Francis
Taylor & Francis Group

Expert Review of Proteomics

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/ieru20

Tackling reproducibility: lessons for the proteomics community

Julian Uszkoreit, Magnus Palmblad & Veit Schwämmle

To cite this article: Julian Uszkoreit, Magnus Palmblad & Veit Schwämmle (2024) Tackling reproducibility: lessons for the proteomics community, Expert Review of Proteomics, 21:1-3, 9-11, DOI: 10.1080/14789450.2024.2320166

To link to this article: https://doi.org/10.1080/14789450.2024.2320166

EXPERT REVIEW OF PROTEOMICS
2024, VOL. 21, NOS. 1–3, 9–11
https://doi.org/10.1080/14789450.2024.2320166

Taylor & Francis
Taylor & Francis Group

EDITORIAL

Tackling reproducibility: lessons for the proteomics community

Julian Uszkoreit [a], Magnus Palmblad [b] and Veit Schwämmle [c]

[a]Medical Bioinformatics, Ruhr University Bochum, Bochum, Germany; [b]Center for Proteomics and Metabolomics, Leiden University Medical Center, Leiden, Netherlands; [c]Department of Biochemistry and Molecular Biology, University of Southern Denmark, Odense, Denmark

ARTICLE HISTORY Received 23 November 2023; Accepted 3 February 2024

KEYWORDS proteomics; reproducibility; quality control; standardization; workflows

1. Introduction

The much reported 'reproducibility crisis' [1] refers to the increasing perception across scientific domains, that many published results are not or only partly reproducible. In proteomics, particularly affinity methods suffer from poor reliability and unknown specificity of antibodies [2]. Interpretation of common readouts such as Western blots is partially subjective. Mass spectrometry (MS) methods struggle with high variability in sample preparation, instrumentation and data analysis workflows, resulting in many false positives and poor reproducibility unless proper statistical controls are applied. Across branches of proteomics, there is growing emphasis on developing standardized protocols to reduce variability, including antibody validation/mass spectrometry system suitability testing, quality control and sharing data and procedures. A 2020 review of studies conducted by ABRF (Association of Biomolecular Resource Facilities) Research Groups [3] specifically highlighted the importance of community-based standards and sustainable frameworks for sharing best practices in methodology, standard operating procedures, and data management across all technologies and domains, including proteomics. In this editorial, we highlight the efforts made by the proteomics community and suggest future activities to tackle reproducibility problems and improve the quality of the scientific results generated. While we touch on some of the sample processing steps prior to data acquisition, the main focus is on the computational analysis.

2. Heterogeneity

Proteomics is faced with considerable sample and data variety, both having a large impact on the outcome and biological interpretation of a proteomics experiment. While there are many sources challenging reproducibility, we highlight some of the most prominent cases below.

Sample heterogeneity stems from methods that are only partially reproducible. For bottom-up proteomics, proteins are digested into peptides. While specific standards are available for the commonly used tryptic digestion, it is still a variable biochemical process leaving missed cleavage sites at varying degrees. This minor problem is overshadowed, however, by the irreproducibility of data dependent analyses: while the MS1 spectra are relatively stable for repeated analyses or measurements, depending on the chromatography setup, the derived MS2 spectra are not. For the most abundant peptides, MS2 spectra can be obtained reproducibly, but less abundant peptides suffer from stochastic changes in their identification between different runs. Data-independent methods such as DIA or PRM improve inter-run stability to some extent, but require careful calibration and adjustment of the collision energy to allow comparison of measurements, even on the same instrument [4].

Data heterogeneity ranges over most steps of the data analysis. Raw data files are different for each instrument vendor and sometimes across machine types. Analysis algorithms are based on different hypotheses and assumptions, use different parameters and parameter settings. This leads to differently formatted and interpreted intermediate and end results. Examples are incomparable PSM tables created by different data analysis pipelines and published data sets in nearly arbitrary table formats. Although this variability is not uncommon in any scientific field, the need to convert data at several stages is an unnecessary additional task, usually resulting in loss of information and burdening the development and establishment of data analysis pipeline, the reproducibility and large-scale re-processing of public data.

3. Standardization

Proper standardization is a crucial step to enable reproducibility as it simplifies the comparison of different results. Many efforts have been taken by the proteomics community to create data and reporting standards for different data types. The efforts of the Proteomics Standards Initiative (PSI) of HUPO have had an impact far beyond human studies [5] and are applicable to many different stages of a project. Starting with the minimal information about a proteomics experiment (MIAPE), community standard formats and guidelines exist covering gel and MS-based proteomics, as well as peptide and protein identification and quantification. All formats are linked by controlled vocabularies to make them interoperable. A more widespread adoption of these formats will facilitate the assembly, modification and extension of data analysis pipelines.

CONTACT Julian Uszkoreit julian.uszkoreit@rub.de Medical Faculty, Medical Bioinformatics, Ruhr University Bochum, Universitaetsstr. 105, Bochum D-44789, Germany

With the ProteomeXchange consortium [6], the community has a standardized way to publish and store scientific data in repositories. The consortium partner repositories around the world enable uploading, sharing and reanalyzing data and thus make it feasible for every scientist to store results in a FAIR (findable, accessible, interoperable and reusable [7]) way.

4. Quality control

In all analytical sciences, rigorous quality control (QC) of the measured data should be the default. Unfortunately, this is often neglected [8] and thus published data may be either irreproducible or lead to wrong conclusions. Several checkpoints for QC have been established, but a gold standard for assessing data quality is so far lacking.

The need for QC starts already during the sample preparation. If a quantitative analysis should be performed, a known amount (or concentration) of peptides respectively proteins in the sample must be injected into the chromatography system or mass spectrometer (in the case of direct infusion). To achieve this, an exact measurement of the amino acid concentration is required. Here, the most accurate method is amino acid analysis. However, this analysis is both expensive and sample consuming, and hence not feasible to perform on every sample. Less expensive albeit less accurate methods could be applied instead [9].

After sample preparation, an instrument operator should monitor the chromatogram and acquired fragmentation spectra in real time. However, there appears to be no common method for such instrument control, resulting in different thresholds for assessing quality during operation.

Several QC characteristics calculated by computational post-processing tools can detect possible problems in any step of the analysis [10]. For this purpose, the HUPO-PSI is creating a community standard called mzQC (https://hupo-psi.github.io/mzQC/). Also, during the computational analysis of a sample, the resulting quality should be assessed and at least the tools or algorithms used, databases and their versions should be recorded to ensure reproducibility.

5. Benchmarking

For the analysis of proteomics data, many software tools and algorithms have been developed over the last years. While this highlights the importance and the progression of the field, each approach gives slightly different results and the original assumptions for the respective applicability, on which any calculation or interpretation is performed, should always be considered, e.g. most quantification and normalization algorithms assume that only a few proteins are regulated. Here, benchmarking and comparisons are very useful approaches to highlight the applicability of specific algorithms and pipelines.

To nurture the lack of available benchmarking platforms, we implemented WOMBAT-P for the comparison of complete proteomics pipelines as part of a study within the ELIXIR infrastructure consortium [11]. This platform allows benchmarking of full workflows starting from the raw file processing to statistical analysis of the quantified results. Here, the goal was not to find the best performing single tool for a specific subtask of the full analysis (in the past often the one returning the highest numbers of proteins), but the unbiased comparison of workflows as tool combinations. This comparison showed that workflow and tool performance depends on multiple factors, and mainly depends on the data set the analysis was applied on, the used parameter settings and the used benchmarking metric to assess workflow performance.

A similar study indicated [12] that for the still existing problem of protein inference the single best tool cannot be determined, and even on the PSM identification level the tools vary to sometimes concerning degrees [13].

6. Expert opinion

An essential aspect of improving reproducibility in the field is the rigorous annotation of samples, data acquisition methods and analyses. Adopting the recent SDRF standard for proteomics [14] improves metadata accuracy and completeness, and allows for additional annotation of already deposited data. Encouraging or enforcing these practices by publishers and data repositories would significantly improve the rigor and reproducibility of MS-based proteomics research. Recent community efforts now allow the deposition of complete computational workflows, including all parameterization and results, into Research Object Crates (RO-Crates) [15]. If a respective analysis is performed using a workflow environment and in an optimal case using software containers, the RO-crates can be deposited together with the analytical raw data into a repository and published, allowing currently the best approach in maximum reproducibility. Proteomics was at the forefront of reproducibility with the introduction of central repositories regarding FAIR data principles ten years ago. Increasing the re-usability of data would help to keep this status.

One important aspect to increase reproducibility is thorough benchmarking of different tools or workflows on the same datasets. With platforms like OpenEBench (https://openebench.bsc.es/) and WOMBAT-P, we hope to see an increase in objective benchmarking in the future.

Reusing established workflows on data generated by new technologies is greatly facilitated by vendors allowing data to be exported in community standard formats. The latter require periodic updates to keep up with technological developments. Although these standard formats are still based on XML and not optimal for storage and processing, they can be read and written by most common programming languages, which has allowed their wide adoption in community software.

While some above-mentioned areas still need considerable improvement and solutions proposed by experts deserve a wider acceptance by the community, many important steps were taken in the last years to improve the reproducibility in proteomics, and as such make it a pivotal component in highly rigorous and transparent research.

Funding

This paper was not funded.

Declarations of interest

The authors have no relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript. This includes employment, consultancies, honoraria, stock ownership or options, expert testimony, grants or patents received or pending, or royalties.

Reviewer disclosures

Peer reviewers on this manuscript have no relevant financial or other relationships to disclose.

Author contributions

All three authors contributed to the manuscript, reviewed it and share any responsibility.

ORCID

Julian Uszkoreit http://orcid.org/0000-0001-7522-4007
Magnus Palmblad http://orcid.org/0000-0002-5865-8994
Veit Schwämmle http://orcid.org/0000-0002-9708-6722

References

Papers of special note have been highlighted as either of interest (•) or of considerable interest (••) to readers.

1. Ioannidis JPA. Why most published research findings are false. PLOS Med. 2005;2(8):e124.
   - **Overview from 2005 stating that most findings in (medical) research are wrong. Some of these points have changed since then.**
2. Baker M. Reproducibility crisis: Blame it on the antibodies. Nature. 2015;521(7552):274–276.
   - **An article that significantly raised awareness about a number of issues with commercially available antibodies and poor reproducibility of experiments relying on them.**
3. Mische SM, Fisher NC, Meyn SM, et al. A review of the scientific rigor, reproducibility, and transparency studies conducted by the ABRF research groups. J Biomol Tech. 2020;31(1):11–26. doi: 10.7171/jbt.20-3101-003
   - **A review of the studies conducted by the ABRF and how good scientific practice was conducted at the cores to uphold reproducible results.**
4. Zolg DP, Wilhelm M, Yu P, et al. PROCAL: a set of 40 peptide standards for retention time indexing, column performance monitoring, and collision energy calibration. Proteomics. 2017;17(21). doi: 10.1002/pmic.201700263
5. Deutsch EW, Vizcaíno JA, Jones AR, et al. Proteomics standards initiative at twenty years: Current activities and future work. J Proteome Res. 2023;22(2):287–301. doi: 10.1021/acs.jproteome.2c00637
   - **An overview of the work conducted by the HUPO-PSI in the last 20 years. It also highlights the internally (re-)used structures like controlled vocabularies and is thus a good starting point for everybody interested to participate in the PSI efforts.**
6. Deutsch EW, Bandeira N, Perez-Riverol Y, et al. The ProteomeXchange consortium at 10 years: 2023 update. Nucleic Acids Res. 2023;51(D1):D1539–D1548. doi: 10.1093/nar/gkac1040
   - **The ProteomeXchange consortium publishes every year the current state of the repositories, regarding the number of datasets, but also new initiatives to make use of the treasure trove of data waiting to be unearthed.**
7. Wilkinson MD, Dumontier M, Aalbersberg IJJ, et al. The FAIR guiding principles for scientific data management and stewardship. Sci Data. 2016;3(1):160018. doi: 10.1038/sdata.2016.18
   - **A must read for anyone who still does not know about the FAIR (Findability, Accessibility, Interoperability, Reusability) principles and their relationships to research quality, rigor and reproducibility.**
8. Rozanova S, Uszkoreit J, Schork K, et al. Quality control—A stepchild in quantitative proteomics: a case study for the human CSF proteome. Biomolecules. 2023;13(3):491. doi: 10.3390/biom13030491
9. Kielkopf CL, Bauer W, Urbatsch IL. Methods for measuring the concentrations of proteins. Cold Spring Harb Protoc. 2020;2020(4):102277. doi: 10.1101/pdb.top102277
10. Bittremieux W, Tabb DL, Impens F, et al. Quality control in mass spectrometry-based proteomics. Mass Spectrom Rev. 2018;37(5):697–711. doi: 10.1002/mas.21544
   - **The article gives an overview of current methods for quality control in proteomics. It highlights which metrics can be used in the various steps of an analysis, ranging from sample preparation to data analysis.**
11. Bouyssié D, Altıner P, Capella-Gutierrez S, et al. WOMBAT-P: benchmarking label-free proteomics data analysis workflows. J Proteome Res. 2023;23:418–429.
   - **In this article a command line based platform for the comparison of computational workflows for the analysis of LC-MS/MS proteomics data is presented. The platform provides several workflows which allow reproducible analyses of any current DDA dataset.**
12. Audain E, Uszkoreit J, Sachsenberg T, et al. In-depth analysis of protein inference algorithms using multiple search engines and well-defined metrics. J Proteomics. 2017;150:170–182. doi: 10.1016/j.jprot.2016.08.002
13. Matthiesen R, Prieto G, Beck HC. Comparing Peptide Spectra Matches Across Search Engines. Mass Spectrometry Data Analysis in Proteomics. 2051 ed. 2020;:133–143. doi: 10.1007/978-1-4939-9744-2_5
14. Dai C, Füllgrabe A, Pfeuffer J, et al. A proteomics sample metadata representation for multiomics integration and big data analysis. Nat Commun. 2021;12(1):5854. doi: 10.1038/s41467-021-26111-3
   - **This article describes the SDRF for proteomics file format, which allows linking of the raw files deposited in a repository to metadata annotations, which allow to store all relevant data necessary to (re-)analyse a dataset, like groups and parameters, but also further biological or sample specific annotations.**
15. Soiland-Reyes S, Sefton P, Crosas M, et al. Packaging research artefacts with RO-Crate. DS. 2022;5(2):97–138. doi: 10.3233/DS-210053
   - **RO-Crates allow to store complete workflows, including all parameters, intermediate and final results into a standardized format. This approach, when combined with containerized workflows, allows easy reproduction (and inspection) of results.**