Journal of Statistics on Health Decision

P2

A non-parametric analysis for the identification of differentially expressed proteins in proteomic data

Pedro O. Corda[1], Tiago Costa[2], Fábio Trindade[3], Vera Afreixo[2]

[1] Institute for Research in Biomedicine (iBiMED), Medical Sciences Department, University of Aveiro, Aveiro, Portugal.

[2] Department of Mathematics & Center for Research and Development in Mathematics and Applications (CIDMA), University of Aveiro, Portugal

[3] UnIC – Cardiovascular Research and Development Centre, Department of Surgery and Physiology, Faculty of Medicine, University of Porto, Porto, Portugal

Introduction

Nowadays, proteomic techniques allow obtaining a large set of raw data that must be processed in order to obtain results with biological significance and clinical applicability. This technique has been particularly relevant in clinical research for the identification of biomolecules that can function as biomarkers (diagnosis, prognosis and risk prediction) of certain pathologies (1). Thus, over the years a set of statistical techniques has been applied to the proteomics analysis in order to identify differentially expressed proteins (DEPs) that can contribute to the understanding of biological processes and pathological mechanisms. Statistically speaking, one of the first problems that proteomics studies face is the small number of samples available relatively to the number of measured variables, largely reducing analysis' power (2). Despite the use of tools that allow the calculation of sample size (e.g. power calculation), the limitations derived from costs, ethical issues or even the reduced epidemiological penetrance of certain diseases, do not allow the inclusion of the necessary number which will question the validity of statistical inference tests (2). A set of methods has been described to effectively determine proteins differentially expressed from these data, based largely on parametric models. The use of parametric models implies the assumption of a normal distribution of data, which can be extremely inappropriate in situations where the sample number is small (2,3). Contrary to what is seen in genomics where non-parametric models are applied, non-parametric tests are rarely used in proteomic analysis. Comparatively to parametric tests, non-parametric tests are distribution-free and, therefore, are more robust than tests based on a distribution. Recently, a non-parametric model was proposed for proteomic data analysis revealing effectiveness and sensitivity in the identification of DEPs (4). The aim of this work was to apply a non-parametric analysis to proteomic data to understand if this approach can improve the stringency in the identification of DEPs.

Methods

To achieve our goal, we used the dataset available from Zhang et al. 2018 (5) where a parametric approach was used to identify DEPs. Briefly, the authors identified proteomic profile changes in frontal cortex tissue samples between Alzheimer's disease (AD) patients (n=8) and healthy controls (n=8) through LC-MS/MS (label-free quantitative). In this dataset, only proteins without missing data across all samples were included (n=1968), being represented by the respective relative abundance (after data scaling). To identify DEPs using a non-parametric approach, we used the Mann-Whitney U test ($P<0.05$). Up- and down-regulated proteins were selected based on AD/control ratio (up-regulated >1.3 and down-regulated <0.77) and a false discovery rate (FDR) <0.11, as set by Zhang and colleagues. The q values were calculated using the "q value" package in R to correct for multiple comparisons and estimate false discovery rates. To perform multiple correction analysis, we used the Benjamini-Hochberg correction both for parametric and non-parametric approaches. To understand if the number of protein identifications was significantly different between parametric and non-parametric approaches the Z test was performed. All statistical analysis was performed in R (version 3.6.1.) though the RStudio, using the "stats" package. To compare the proteins deemed differentially expressed by the non-parametric analysis with that of the parametric approach, we performed a Venn diagram analysis using JVenn tool (6).

Keywords:
proteomics, statistical analysis, non-parametric analysis

Corresponding author:
Pedro O. Corda
pedrocorda@ua.pt

Supplementary material:
https://proa.ua.pt/index.php/jshd/article/view/18252/15276

First published: 23 OCT 2020

Results

Zhang et al. observed, through the Student's t-test, 529 DEPs in AD patients. From those, 262 proteins were up-regulated and 225 were down-regulated, according to the criteria (±1.3 fold-change of over the control and FDR <0.11). Through the Mann-Whitney U test, we found 515 DEPs in AD patients, of which 193 were up-regulated and 156 were down-regulated, under the same criteria. Table 1 summarises the number of proteins found by each statistical approach.

Table 1 – Number of proteins identified by each statistical approach. The proportion of the proteins in relation to the entire proteome is given in brackets. * p < 0.05 (Z test)

Approach	Differential Expressed Proteins (p<0.05)	Up-Regulated Proteins	Down-regulated Proteins
Parametric	529 [26.9%]	262 [13.3%]*	225 [11.4%]*
Non-parametric	515 [26.2%]	193 [9.8%]*	156 [7.9%]*

To inquire the level of redundancy of the two statistical approaches, we performed a Venn diagram analysis (figure 1). Most of the DEPs identified by the parametric approach were also disclosed by the non-parametric approach. Interestingly, the non-parametric approach retrieved some proteins that were not considered in the study conducted by Zhang et al.

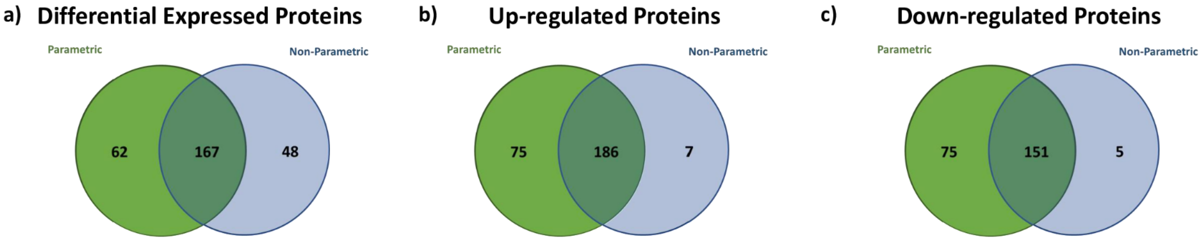

Figure 1 – Venn diagram illustrating the common proteins between parametric and non-parametric approach for (a) differentially expressed proteins: (b) up-regulated proteins and (c) down-regulated proteins.

Discussion

In proteomics studies, one of the key factors in the analysis workflow is the identification of DEPs through statistical tests (2). These proteins will subsequently be used in other analyses (experimental or bioinformatics) in order to obtain relevant information for a biological context. Most of the statistical tests used are based on parametric models, which are built on certain assumptions (e.g. normality) that are frequently violated in proteome experiments (2,3). In the present study, we reanalysed a dataset with a non-parametric test to understand whether the results obtained could be substantially different. The first difference was the smaller number of proteins identified (n=515) as differentially expressed compared to the parametric approach (n=529). Consequently, and considering the remaining selection criteria, the number of up- and down-regulated proteins was considerably smaller. In the original article, the authors used bioinformatically analysis of up- and down-regulated proteins to gain new insight into this pathology (5). Thus, by reducing these protein lists, the information that would be obtained from bioinformatics analysis could be more specific and closer to biological reality. Interestingly, we also found proteins in the non-parametric analysis that were excluded by the parametric analysis (figure 1), thus being able to lose biological information of interest. It is also important to note that in this type of analysis, the possibility of false positives can be reduced through multiple correction methods (such as Benjamini-Hochberg correction). Even in this rigorous scenario, the non-parametric test maintained its stringency in comparison to the parametric test (data not shown).

Conclusion

The non-parametric approach seems to be more stringent in the identification of differentially expressed proteins, which may bring advantages for subsequent analyses. Future studies should be conducted in order to understand the advantages of the non-parametric approach in the field of proteomics.

Acknowledgment

The authors thank FCT, Fundação para a Ciência e Tecnologia, I.P., for funding Institute for Research in Biomedicine - iBiMED -(UIDB/04501/2020 and POCI-0145-FEDER-007628), Cardiovascular R&D Center – UnIC - (UIDB/00051/2020 and UIDP/00051/2020) and Center for Research and Development in Mathematics and Applications - CIDMA - (UIDB/04106/2020 and UIDP/04106/2020). FT thanks the post-doctoral research grant from the project "New targets in diastolic heart failure: from comorbidities to personalized medicine – NETDIAMOND" financed by the European Structural and Investment Funds (ESIF), through the Programa Operacional Regional Lisboa 2020 (POCI-01-0145-FEDER-016385) and national funds by FCT (SAICTPAC/0047/2015).

References

1. Frantzi M, Bhat A, Latosinska A. Clinical proteomic biomarkers: relevant issues on study design & technical considerations in biomarker development. Clin Transl Med. 2014;3(1):7. https://doi.org/10.1186/2001-1326-3-7
2. Suppers A, van Gool AJ, Wessels HJCT. Integrated chemometrics and statistics to drive successful proteomics biomarker discovery. Proteomes. 2018;6(2). https://doi.org/10.3390/proteomes6020020
3. Lualdi M, Fasano M. Statistical analysis of proteomics data: A review on feature selection. J Proteomics. 2019;198:18–26. https://doi.org/10.1016/j.jprot.2018.12.004
4. Slama P, Hoopmann MR, Moritz RL, Geman D. Robust determination of differential abundance in shotgun proteomics using nonparametric statistics. Mol Omi. 2018;14(6):424–36. https://doi.org/10.1039/C8MO00077H
5. Zhang Q, Ma C, Gearing M, Wang PG, Chin LS, Li L. Integrated proteomics and network analysis identifies protein hubs and network alterations in Alzheimer's disease. Acta Neuropathol Commun. 2018;6(1):19. https://doi.org/10.1186/s40478-018-0524-2
6. Bardou P, Mariette J, Escudié F, Djemiel C, Klopp C. SOFTWARE Open Access jvenn: an interactive Venn diagram viewer. BMC Bioinformatics. 2014;15(293):1–7. https://doi.org/10.1186/1471-2105-15-293