Comparison of Proteomic Approaches in Autoinflammatory Disease Classification

Orestis D. Papagiannopoulos
Unit of Medical Technology and Intelligent Information Systems, Dept. of Materials Science and Engineering, University of Ioannina, Ioannina GR45110, Greece, orepap@uoi.gr

Costas Papaloukas
Dept. of Biological Applications and Technology University of Ioannina & Dept. of Biomedical Research, Foundation for Research and Technology-Hellas, Institute of Molecular Biology and Biotechnology (FORTH-IMBB), Ioannina GR45110, Greece, papalouk@uoi.gr

Vasileios C. Pezoulas
Unit of Medical Technology and Intelligent Information Systems, Dept.of Materials Science and Engineering, University of Ioannina, Ioannina GR45110, Greece, v.pezoulas@uoi.gr

Harmen J.G. van de Werken
Dept. of Immunology, Erasmus MC Cancer Institute, University Medical Center, Dr. Molewaterplein 40, 3015 GD, Rotterdam, the Netherlands, h.vandewerken@erasmusmc.nl

Christophe Poulet
Laboratory of Rheumatology, GIGA-Research CHULiège, ULiège, 4000, Liège, Belgium, christophe.poulet @chuliege.be

Yvonne M. Mueller
Dept. of Immunology, Erasmus MC Cancer Institute University Medical Center, Dr. Molewaterplein 40, 3015 GD, Rotterdam, the Netherlands, y.muller@erasmusmc.nl

Peter D. Katsikis
Dept. of Immunology, Erasmus MC Cancer Institute, University Medical Center, Dr. Molewaterplein 40, 3015 GD, Rotterdam, the Netherlands, p.katsikis@erasmusmc.nl

Dominique de Seny
Laboratory of Rheumatology, GIGA-Research CHULiège, ULiège, 4000, Liège, Belgium, ddeseny@chuliege.be

Dimitrios I. Fotiadis
Unit of Medical Technology and Intelligent Information Systems, Dept. of Materials Science and Engineering & Dept. of Biomedical Research, Foundation for Research and Technology-Hellas, Institute of Molecular Biology and Biotechnology (FORTH-IMBB), Ioannina GR45110, Greece, fotiadis@uoi.gr

2022 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI) | 978-1-6654-8791-7/22/$31.00 ©2022 IEEE | DOI: 10.1109/BHI56158.2022.9926877

***Abstract*— A cross-analysis study was conducted to compare proteomic platforms in classifying patients with Systemic Autoinflammatory diseases, using proteins extracted from different profiling experiments. The datasets used were obtained from SomaScan assays and Mass Spectrometry (MS). A separate analysis was performed to each dataset based on the false discovery rate (FDR) in order to extract statistically important proteins. Conventional machine learning algorithms were subsequently employed to evaluate the denoted proteins as candidate biomarkers and compare the predictive capabilities of the two proteomic platforms. Using the SomaScan assay, we managed to achieve higher classification metrics compared to the MS dataset. An improvement was also attained on the classification results when the features used were extracted from the MS data and applied on the SomaScan dataset, compared to the opposite combination. Finally, the proteins derived from the FDR analysis in both datasets proved to be highly correlated regarding their importance score.**

Keywords—Proteomics, SomaScan, Mass Spectrometry, Systemic Autoinflammatory Diseases

I. Introduction

Systemic autoinflammatory diseases (SAIDs) is a group of no-age specific conditions that encompasses several rare disorders and are characterized by extensive inflammation [1]. Physical manifestations of SAIDs mainly include fever, rash, joint pain or swelling and in most cases, have a strong genetic mutation background underlying the dysregulation of the innate immune system [2-3]. Inability for a distinct diagnosis to be met, amounts for at least 40 - 60% of patients with phenotypes typical for SAIDs [4]. Mainly, the diagnosis process consists of clinical evaluation along with the exclusion of other possible disorders and as such, delays and inadequate treatment decisions are prevalent to patients with SAID-related conditions. Contrary to autoimmune pathology whose autoantibodies are diagnostic tools, even today, there is no SAID-defining biomarkers.

SomaScan assays are highly sensitive and reproducible tools for clinical diagnosis across a vast range of diseases, capable of measuring up to 7000 protein analytes in only a very small amount of biological matrix [5]. Slow Off-rate Modified Aptamers (SOMAmers) are chemically constructed reagents, from the transformation of each individual protein concentration of native in-matrix proteins to corresponding reagent concentrations. This process grants to SOMAmers the properties of a protein affinity-binding reagent along with the properties of a unique recognizable nucleotide sequence. Being single-stranded DNA-based protein affinity reagent, also renders them highly suitable for aptamer discovery technology [6]. SomaScan assays have enabled significant advances in the identification of plasma proteome signatures. In [7] the proteomic signature of surgery is characterized in association with postoperative surgery outcomes, while in [8] the plasma brain natriuretic peptide measurements were correlated with the Alzheimer's disease. Biomarker identification and evaluation through SomaScan assays have also yielded promising results in the last couple of years in kidney disease [9-10], as well as, in systemic juvenile idiopathic arthritis [11].

Mass spectrometry-based proteomics is a complex high-throughput technology used widely to analyze biological samples. It measures the mass-to-charge ratio (*m/z*) of molecules, quantifying known compounds, and determining their chemical properties and structure. In the field of SAIDs, proteomic mass spectrometry (MS) techniques have enabled the protein analysis of serum exosomes in the post intravenous immunoglobin therapy period of Kawasaki patients [12] and the identification of novel candidate biomarkers [13]. Regarding Behçet's disease, MS methods have proven crucial for the profiling of the peripheral blood mononuclear cell proteome [14], as well as for the analysis of metabolomic alteration associated with the disease [15].

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 779295 (ImmunAID – Immunome project consortium for AutoInflammatory Disorders).

In this study, two recently acquired proteomic datasets from the European Union's Horizon 2020 ImmunAID project were exploited, corresponding to the two platforms described above. We implemented a novel cross-analysis able to identify non-previously examined candidate protein biomarkers between SAIDs and control patients and further classify them. Therefore, we could assess the efficacy of the two platforms in the classification of SAID patients, along with the importance of the extracted features in a cross-platform schema.

II. Materials And Methods

A. SOMAcan Assay

To remove systematic biases in the raw SomaScan assay data, normalization and calibration procedures were applied. Hybridization control normalization was performed to correct for systematic effects on the data introduced during the hybridization readout, followed by median signal normalization across calibrators to adjust for systematic variability within a single plate. Plate scaling is accomplished based on unique-to-SOMAmers scale factors, due to the idiosyncratic nature of SOMAmers binding reagents. The final assay is provided in the form of a tab-delimited ASCII file containing the measurements in Relevant Fluorescent Units for a series of analytes across a set of samples.

B. Mass Spectrometry

The Liquid Chromatography procedure (LC-MS/MS) was performed for the acquisition of the mass spectrometry results and the Andromeda search engine [16] was utilized for the MS spectra analysis. Extracted proteins from plasma samples were injected on a 2D-nanoAquity UPLC (Waters, Corp., Milford, MA, USA) coupled online with an ESI-Q-Orbitrap (Q Exactive, Thermo Fisher Scientific, Waltham, MA, USA) in positive ion mode. The final dataset is provided as a comma separated values file containing the raw measurements in parts per-notation units for a series of analytes across the samples. The missing values were imputed with zeroes for the analysis.

C. Datasets

For the purpose of this study, the datasets were preprocessed to include only the common samples and proteins. The identification of the proteins was based on the UniProt IDs, as provided from both approaches, taking also into account the corresponding protein chains. The final datasets have both dimensions $24x460$, with 24 common sample labels (subjects) and 460 common proteins. Overall, they include four Adult-Onset Still's Disease samples, three Schnitzler disease, two Systemic-Onset Juvenile Idiopathic Arthritis, two Chronic Osteitis, two Cryopyrin-Associated, two Takayasu, one Behçet's, one Tumor Necrosis Factor Receptor-Associated Periodic Syndrome, one Kawasaki, one Recurrent Pericarditis, one inflammation disease of unknown origin and four negative control (healthy) samples.

D. False Discovery Rate Analysis

On each dataset, a False Discovery Rate (FDR) analysis was applied using the SelectFdr class [17] of the scikit learn Python package, which implements the Benjamini-Hochberg procedure [18]. The statistical importance of the dataset features was computed based on the Analysis of Variance (ANOVA) value, by examining the null hypothesis of no significant difference between the feature vectors (proteins) and the binary target vector (SAIDs or control). For the SomaScan dataset, an alpha value threshold 0.001 was selected, resulting in ten statistically important proteins regarding the differentiation of SAID samples from controls. In order to get a similar number of features, as that from the SomaScan dataset, an alpha value threshold 0.01 was selected for the MS dataset, resulting in seven statistically important proteins. The ANOVA values of the features extracted from the SomaScan dataset ranged from 15 - 27, compared to the 8 - 20 range for the features extracted from the MS dataset.

E. Classification

Machine learning techniques were implemented to explore whether the extracted proteins can be utilized as SAIDs biomarkers. The statistically important proteins of each FDR analysis were used as the features for a binary classification between SAIDs and healthy controls on their corresponding dataset. Moreover, we compared the data quality of the two platforms in a classification accuracy context. This was managed by conducting the same classification task, using on each dataset, the extracted proteins of the other dataset as input features. Lastly, by using both sets of extracted features on both datasets for the purposes of classification, we were able to evaluate the feature importance score for each dataset. Thus, two vectors of feature importance for each set of extracted proteins were obtained. For those two sets of vectors, the Pearson's correlation index [19] was calculated. This is to assess the linear correlation between the vectors, meaning how similar the importance of the features is between both datasets, for each set of extracted features.

For the tasks of classification and the feature importance score estimation, the Random Forest Classifier (RF) [17] was used. The Gradient Boosting Classifier (GB) and the Logistic Regression (LR) estimators [17] were also employed for classification scrutiny. RF and GB are robust supervised ensemble machine learning algorithms, which utilize a multitude of decision trees for training, improving their capabilities by ensemble learning and through error reduction boosting. LR is a simpler and more efficient estimator that excels in binary outcome scenarios and determines whether a new sample that fits best into a category is also important [20]. All estimators were employed with their default parameter settings, as we opted for classification comparison and not performance optimization.

To assess the significance of the extracted proteins, the accuracy, sensitivity, and specificity metrics were used, based on the leave one out cross validation schema. To address the high sample imbalance, for each dataset, the classification process was repeated five times, each with a different set of 4 SAID samples against the 4 controls. Hence, the classification metrics are given as the means along with the standard errors. Due to the stochastic nature of the RF classifier, the process of calculating the feature importance scores and the Pearson's index was repeated fifty times for all samples, resulting in the mean score for those values. Fig. 1 summarizes the workflow of the study.

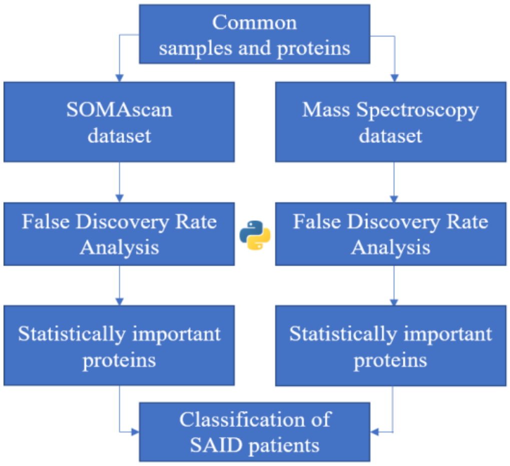
Common samples and proteins
SOMAscan dataset
Mass Spectroscopy dataset
False Discovery Rate Analysis
False Discovery Rate Analysis
Statistically important proteins
Statistically important proteins
Classification of SAID patients

Figure 1. Workflow of the study.

III. Results And Discussion

A. Classification

Tables I and II show the metrics of the classification, using the extracted set of proteins of each FDR analysis, as features in their corresponding dataset. Tables III and IV show the same metrics, but in the case where the extracted set of proteins of each FDR analysis were used as input features in the other dataset. Fig. 2 depicts the feature importance score in both SomaScan and MS datasets of the proteins extracted from the FDR analysis of the first dataset. While Fig. 3 shows the same feature importance scores but for the proteins extracted from the FDR analysis of the MS dataset. These figures include also the Pearson's correlation index of the feature importance scores between both datasets, for the two sets of extracted proteins.

B. Discussion

Our results show the improved classification metrics using the SomaScan assay in comparison to the MS one, when selecting features based on FDR analysis on both datasets. More specifically, the derived scores were equal or higher than 90% for all estimators on the SomaScan data (Table I), in comparison to the 55% - 80% range on the MS dataset (Table II). It should be noted that the above results correspond to analyses performed on the same proteins for both datasets, as the scope of this work is to compare the data quality and accuracy of the two platforms. This is further stressed by the cross-classification results.

TABLE I. SOMA dataset classification metrics (mean and standard error) using the proteins derived from the FDR analysis of the SOMA dataset.

Classifier	Accuracy	Sensitivity	Specificity
GB	0.925 (0.045)	0.950 (0.045)	0.900 (0.089)
RF	0.975 (0.022)	0.950 (0.045)	1.000 (0.000)
LR	0.950 (0.027)	0.900 (0.055)	1.000 (0.000)

TABLE II. MS dataset classification metrics (mean and standard error) using the proteins derived from the FDR analysis of the MS dataset.

Classifier	Accuracy	Sensitivity	Specificity
GB	0.675 (0.076)	0.800 (0.084)	0.550 (0.084)
RF	0.725 (0.055)	0.800 (0.084)	0.650 (0.055)
LR	0.775 (0.042)	0.800 (0.084)	0.750 (0.000)

TABLE III. SOMA dataset classification metrics (mean and standard error) using the proteins derived from the FDR analysis of the MS dataset.

Classifier	Accuracy	Sensitivity	Specificity
GB	0.850 (0.042)	0.800 (0.045)	0.900 (0.055)
RF	0.850 (0.045)	0.800 (0.045)	0.900 (0.055)
LR	0.900 (0.022)	0.850 (0.055)	0.950 (0.045)

TABLE IV. MS dataset classification metrics (mean and standard error) using the proteins derived from the FDR analysis of the SOMA dataset.

Classifier	Accuracy	Sensitivity	Specificity
GB	0.600 (0.089)	0.550 (0.164)	0.650 (0.055)
RF	0.750 (0.035)	0.750 (0.071)	0.750 (0.000)
LR	0.725 (0.074)	0.700 (0.084)	0.750 (0.071)

Using the proteins extracted by the MS FDR analysis as classification features in the SOMA dataset, the accuracy and specificity (Table III) are notably higher than those of the same-dataset classification for the MS (Table II). To discard the possibility that the SomaScan dataset inherently allows for high classification metrics, a number of tests using random proteins as input features were conducted, none of which produced such high results. Furthermore, the feature importance scores across the datasets for both sets of extracted proteins follow the same high correlation trend, but with the features extracted for the SomaScan FDR analysis having a slightly higher Pearson index. This implies that both platforms produce measurements that are highly comparable and reliable.

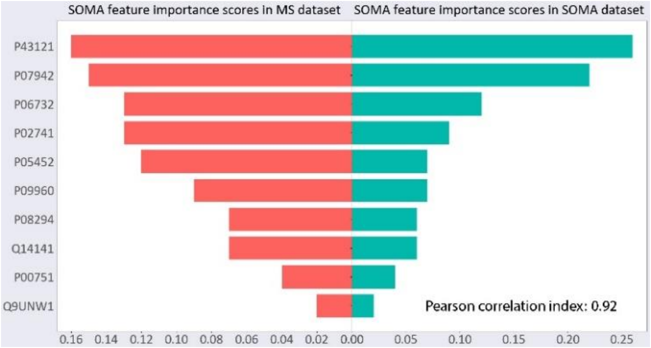

Figure 2. Feature importance score in both datasets of the proteins derived from the FDR analysis on the SomaScan dataset.

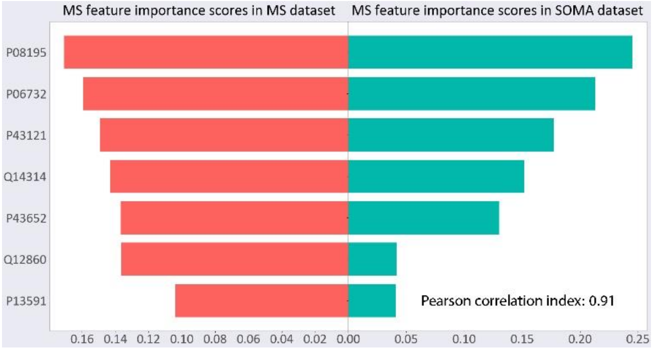

Figure 3. Feature importance score in both datasets of the proteins derived from the FDR analysis on the MS dataset.

Two proteins with UniProt IDs P06732 and P43121 were found in both sets of the FDR extracted proteins. They refer to the proteins Creatine kinase M-type and Melanoma-associated antigen MUC18 (Table V). The scores yielded from the Analysis of Variance were 27 and 20, respectively, in the SomaScan dataset, and 17 and 11 in the MS. According to the National Center for Biotechnology Information (NCBI) [21], Creatine kinase M-type is a cytoplasmic enzyme, reversibly catalyzing the transfer of phosphate between ATP and various phosphogens. It is an important serum marker for myocardial infarction [21] and elevated levels have been associated with neuromyopathy [22] and myopathy in Behçet's disease [23]. Likewise, Melanoma-associated antigen MUC18 is a biomarker of uveal melanoma and chronic obstructive pulmonary disease [21] and plays a role in lymphocyte endothelium interaction [24].

TABLE V. The statistically most important proteins of the two FDR extracted sets of proteins examined in the present study, alongside their description and known biological associations [21].

Protein	Description/Association
CK-MM	• Member of the ATP: guanido phosphotransferase protein family. • Biomarker for myocardial infarction.
MUC18	• Located in the external side of the plasma membrane, acting upstream of or within angiogenesis. • Biomarker of uveal melanoma and chronic obstructive pulmonary disease.
LAMA2	• Extracellular protein and a major component of the basement membrane. • Organizes cells into tissues during embryonic development. • Biomarker of congenital merosin-deficient muscular dystrophy.
SLC3A2	• Cell surface transmembrane protein, member of the solute carrier family. • Participates in the intracellular calcium levels regulation and L-type amino acids transportation.

IV. Conclusions

We presented a novel cross-analysis study in which we utilized proteomic data from two different high-throughput platforms, SomaScan and Mass Spectrometry, to extract statistically important proteins for the classification of SAID patients. We compared the two approaches in terms of accuracy, specificity and sensitivity, using protein measurements of the same sample labels and proteins. The SomaScan assay provided higher classification results than MS, using both same and cross-dataset features. We believe this is due to the more sensitive and less noisy measurements compared to MS. Both methods proved to be highly correlated regarding the significance of the derived features. Furthermore, we managed to identify informative proteins, based on the results of both the FDR analysis and the feature importance scores of the machine learning estimator, which could potentially be characterized as candidate SAID biomarkers. Additional data for SAIDs are currently being collected under the framework of the ImmunAID project, and consequently an extensive validation of this study's approach in terms of both performance optimization and candidate biomarkers identification will be carried out in our immediate future work, to provide further clinical insight to the insufficiently examined SAIDs.

V. References

[1] P. Efthimiou, P. Paik, L. Bielory, "Diagnosis and management of adult onset Still's disease", Annals of the rheumatic diseases, 65.5, pp. 564-572, 2006.

[2] A. Betrains et al., "Systemic autoinflammatory disease in adults", Autoimmunity Reviews 20.4, 2021.

[3] G. Donato et al., "Monogenic Autoinflammatory Diseases: State of the Art and Future Perspectives", International Journal of Molecular Sciences 22.12, 2021.

[4] J. Krainer, S. Siebenhandl, A. Weinhäusel, "Systemic autoinflammatory diseases", Journal of autoimmunity 109, 102421, 2020.

[5] R. Liu et al., "Comparison of proteomic methods in evaluating biomarker-AKI associations in cardiac surgery patients", Translational Research 238, pp. 49-62, 2021.

[6] M. Dunn, R. Jimenez, J. Chaput, "Analysis of aptamer discovery and technology", Nature Reviews Chemistry 1.10, pp. 1-16, 2017.

[7] T. Fong et al., "Identification of plasma proteome signatures associated with surgery using SOMAscan", Annals of surgery 273.4, 732, 2021.

[8] E. Begic et al., "SOMAscan-based proteomic measurements of plasma brain natriuretic peptide are decreased in mild cognitive impairment and in Alzheimer's dementia patients", PloS one 14.2, 2019.

[9] O. Govaere et al., "Transcriptomic profiling across the nonalcoholic fatty liver disease spectrum reveals gene signatures for steatohepatitis and fibrosis", Science translational medicine 12.572, 2020.

[10] Y. Luo et al., "SOMAscan proteomics identifies serum biomarkers associated with liver fibrosis in patients with NASH", Hepatology Communications 5.5, pp. 760-773, 2021.

[11] G. Chen et al., "Serum proteome analysis of systemic JIA and related lung disease identifies distinct inflammatory programs and biomarkers", Arthritis & Rheumatology, 2022.

[12] L. Zhang et al., "Differential protein analysis of serum exosomes post-intravenous immunoglobulin therapy in patients with Kawasaki disease", Cardiology in the Young 27.9, pp. 1786-1796, 2017.

[13] Y. Zhou et al., "Identification of a novel anti-heat shock cognate 71 kDa protein antibody in patients with Kawasaki disease", Molecular medicine reports 21.4, pp. 1771-1778, 2020.

[14] A. Kirectepe et al., "Peripheral blood mononuclear cell proteome profile in Behçet's syndrome", Rheumatology International 40.1, pp. 65-74, 2020.

[15] W. Zheng et al., "Metabolomic alterations associated with Behçet's disease", Arthritis research & therapy 20.1, pp. 1-10 2018.

[16] J. Cox et al., "Andromeda: a peptide search engine integrated into the MaxQuant environment", Journal of proteome research 10.4, pp. 1794-1805, 2011.

[17] F. Pedregosa et al., "Scikit-learn: Machine learning in Python", the Journal of machine Learning research 12, pp. 2825-2830, 2011.

[18] Y. Benjamini, Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing", Journal of the Royal statistical society: series B (Methodological), 57.1, pp. 289-300, 1995.

[19] D. Freedman, R Pisani, R Purves, "Statistics". Pisani, R Purves, 4th edn. WW Norton & Company, New York, 2007.

[20] T. Edgar, D. Manz, "Research methods for cyber security", Syngress, 2017.

[21] G. Brown et al., "Gene: a gene-centered information resource at NCBI", Nucleic acids research 43, D36-D42, 2015.

[22] M. Altiparmak et al., "Colchicine neuromyopathy: a report of six cases", Clinical and experimental rheumatology 20.4, 2002.

[23] Y. Yamanishi et al., "A case of cyclosporin A-induced myopathy", Ryumachi.[Rheumatism] 33.1, pp. 63-67, 1993.

[24] B. Guezguez et al., "Dual role of melanoma cell adhesion molecule (MCAM)/CD146 in lymphocyte endothelium interaction: MCAM/CD146 promotes rolling via microvilli induction in lymphocyte and is an endothelial adhesion receptor", The Journal of Immunology 179.10, pp. 6673-6685, 2007.