DOI: https://doi.org/10.4038/sjhs.v3i1.49

Received: 07 October 2022 *Accepted: 20 January 2023*

Identifying Ordinal Nature Inherited Proteins Associated with a Certain Disease

Samarawickrama, O.[1*], Jayatillake, R.[2*], & Amaratunga, D.[3]

ovinrams@gmail.com, rasika@stat.cmb.ac.lk, damaratung@yahoo.com
[1,2]*Department of Statistics, Faculty of Science, University of Colombo, Sri Lanka.*
[3]*Princeton Data Analytics LLC, USA.*

Abstract

Proteomic studies are studies of protein expression levels. They are growing swiftly with the steady improvement in technology and knowledge of cell biology. Since differentially expressed proteins have an influence on overall cell functionality, this improves discrimination between healthy and diseased states. Identifying prime proteins offers prospective insights for developing optimized and targeted treatments. This research involves analyzing data from an early-stage study of which the main purpose was to identify differentially expressed proteins. There are three progressively serious disease states (healthy to mild to severe) in this study. The analysis can be categorized into 2 stages as univariate and multi-protein analysis. The approach of the univariate analysis was to implement continuation ratio modeling considering one protein at a time to pick those that exhibit potential ordinality. Penalized continuation ratio modeling using lasso regularization incorporated with bootstrapping proteins was performed as the next stage to identify protein combinations that perform well together. Combining results of the univariate and multi-protein analyses identified 20 proteins that join forces to discriminate disease severity with an ordinal setting and 21 proteins that are effective each on its own.

Keywords: Bootstrapping, Lasso regularization, Ordinal nature, Proteomic studies, Trend tests.

Introduction

Being an essential substrate of living matter, proteins have a distinctive significance in biology. Wherever growth and reproduction take place in living organisms, proteins play a vital role in building cellular structures, in mediating metabolic pathways, in protection mechanisms, in transporting materials and so forth, simply being the major building block and the major functional bio-polymer to maintain life on Earth. Differential expression of proteins leads to over-functioning, under-functioning, normal functioning, or unrelated/ silencing effects on the overall functionality of the cells; allowing to discriminate between the healthy and diseased conditions leading to a huge diversity among phenotypes of organisms (Alberts, 2002). As such, high throughput protein assays are useful for

parallel screening of multiple proteins and are now being used in many research areas for expression profiling (Amaratunga & Cabrera, 2012; Kingsmore, 2006; Lubomirski et al., 2007; Stoevesandt et al., 2009).

In the development of drugs for diseases, understanding the expression patterns and the differential expression of proteins is of vital importance (Yang et al., 2018). Hence, the study discussed in this paper will help to systematically profile the protein expression related to disease progression which would be important in the medical field to align the obtained data more authentically with the disease condition. This approach would help identify potential protein biomarkers that could be used to categorize diseased individuals or subtype of disease, the stage and progression of the disease (early, mid or late), hence improving diagnosis for patient management and treatment, and also to identify potential drug targets which could be used for drug development (Rusling et al., 2010).

The data we analyse here are from a clinical study in which the patients were categorized into 3 groups based on their disease status: healthy, mild and severe; and the expression levels of 144 proteins were measured. To study ordinal categorical data, several methods have been used in the existing literature. Parametric tests such as Helmert / Reverse Helmert test, proportional odds model, continuation ratio model and non-parametric tests such as Jonckheere Terpstra trend test have been used to explore the ordinal nature. To spot extreme trend patterns across the ordinal categorical levels, Helmert or reverse Helmert contrasts can be applied (Sundström, 2010). Using proportional odds model to discriminate the behaviour across the ordinal levels has its advantage due to its highly interpretable nature (Liu, 2010). But one major limitation is its assumption where the odds ratios of a protein has to be constant across all the target group cut-offs (Brant, 1990). Jonckheere Terpstra test which contrasts across population medians has been sensitive even for modest changes across the ordinal levels (Ali et al., 2015). For high-throughput data, the general methodology has been to segment into one or more dichotomous response analyses. However, this method does not make use of the entire data, thus leading to reduced power (Ananth & Kleinbaum, 1997). Archer and Williams (2012)the common approach to analyzing ordinal response data has been to break the problem into one or more dichotomous response analyses. This dichotomous response approach does not make use of all available data and therefore leads to loss of power and increases the number of type I errors. Herein we describe an innovative frequentist approach that combines two statistical techniques, L 1 penalization and continuation ratio models, for modeling an ordinal response using gene expression microarray data. We conducted a simulation study to assess the performance of two computational approaches and two model selection criteria for fitting frequentist L 1 penalized continuation ratio models. Moreover, we empirically compared the approaches using three application datasets, each of which seeks to classify an ordinal class using microarray gene expression data as the predictor variables. We conclude that the L 1 penalized constrained continuation ratio model is a useful approach for modeling an ordinal response for datasets where the number of covariates (p describe fitting a L1

penalized continuation ratio model to predict an ordinal class using gene expression data. Penalized models have demonstrated to be effective when applied to high throughput genomic datasets (Zhu & Hastie, 2004). However, there is no clear-cut guaranteed method for analysing ordinal categorical data.

This study explores the behavioural pattern of protein expression levels across the three progressive disease states. Proteins of most interest are those that inherit an ordinal setting. Thus, the primary objective of this research is to study the differential expression of proteins taking into account the ordinal nature of disease severity. seven million students in America receiving special education services between the year 2019-2020 (Lewis et al., 2021).

Materials and Methods

In this study, the expression levels of 144 proteins were measured for 45 subjects using a multiplex procedure in which all the proteins were studied simultaneously. The levels of the target group were ordered as healthy, mild and severe. Out of the 45 subjects, 15 are healthy, 15 are mildly diseased and the remaining 15 are severely diseased.

The analysis of this data can be segregated into two sections, namely univariate and multi-protein analysis. Since proteins of most interest are those that follow an ordinal pattern, methods that take this into account were utilized. The approach of the univariate analysis was to fit a forward continuation ratio (CR) model which compares a particular level to higher levels, to each protein individually (Liu, 2010). The CR model, which is also known as the stage approach model, focuses on transitions of successive levels and it assumes that the lower levels have been reached first. The odds of being in a particular level in comparison to being above that level is estimated using maximum likelihood (Liu, 2010). Considering the expression levels of one protein at a time as the explanatory variable, the model implementation was iterated across all the 144 proteins. For the categorical response variable that has i levels, there will be (i-1) models. i^{th} model is as follows:

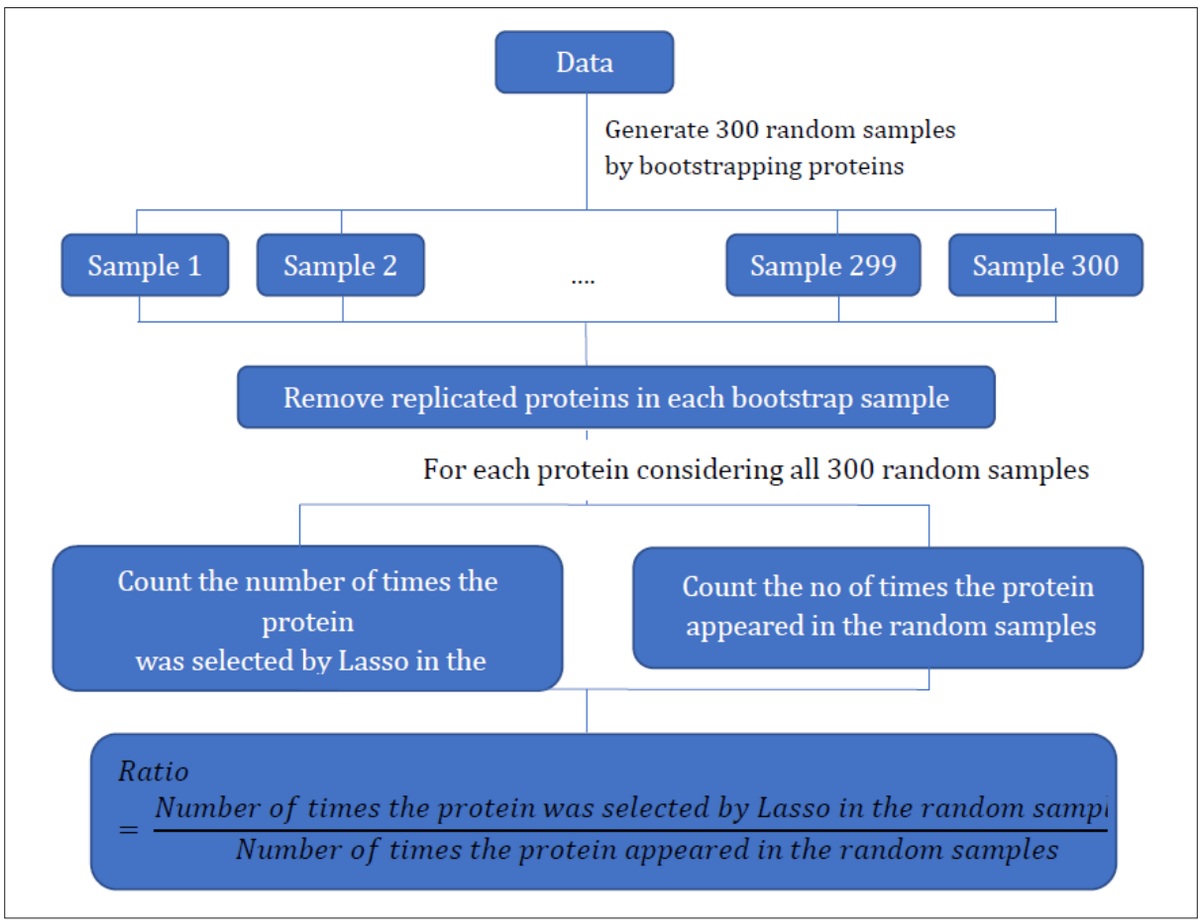
$$\log_e\left(\frac{p_i}{\sum_{1>i}^{I} p_I}\right) = \alpha_i + \beta_i X$$

where, $_i P_i$ = Probability (response i for an observation with explanatory variable X)., $\alpha_i$is the intercept and β_i is the logit coefficient. Here, each regression is fitted using regular binary logistic regression and the sum of G^2 statistics can be used to assess the protein's importance. However, the issue of multiple comparisons may occur since this hypothesis test is implemented for all the proteins. Thus, False Discovery Rate (FDR) adjusted p values were examined (Jafari & Ansari-Pour, 2019).

In the multi-protein context, penalized continuation ratio modelling using lasso regularization (Tibshirani, 1996) incorporated with bootstrapping proteins was implemented.

Figure 1.

Flow chart representation of the multiprotein implementation.



The flow chart in Figure 1 is an abstraction of the multi-protein implementation. Lasso regularization strives to achieve a balance between minimizing the value of the logistic regression loss function and the size of the coefficients. The reason for using lasso regularization was because of its implicit feature selection capability. Thus, redundant variables that do not add any information can be removed. In bootstrapping, random sampling with replacement is used to generate novel samples of the same sample size (Efron, 1992; Hesterberg, 2011). The approach of this study was to bootstrap proteins. If a protein appeared more than once in a sample, the replicates were disregarded. The reason for bootstrapping proteins was to identify different combinations of proteins which seem to do quite well together (Amaratunga et al., 2012). Considering 300 bootstrap samples, lasso penalized continuation ratio model was implemented for each bootstrap sample. To analyse the results of this approach, a ratio was derived: the number of times the protein was selected by lasso divided by the number of times that particular protein appeared in the bootstrap sample. The higher the ratio, the more important the protein is likely to be.

Results and Discussion

Univariate analysis

From the application of continuation ratio model in univariate context, 59 proteins were declared significant based on the p value. However, as discussed in the methodology,

the problem of multiple comparisons was addressed. False Discovery Rate (FDR) adjusted p values were used rather than the original p values to declare the significance. Thus, when adjusted p values were contrasted with 5% significance level, the number of significant proteins reduced from 59 to 41. A comparison of p values and adjusted p values for the protein expression levels is shown in Table 1.

Table 1.
p value and FDR adjusted p values for the 144 proteins.

Protein	Estimate	p value	FDR adjusted p-value	Significance
P1	-0.0068	0.9826	0.9894	Not Significant
P2	-3.2028	0.0009	0.0057	Significant
P3	-3.3566	0.0037	0.0163	Significant
P4	2.9982	0.0001	0.011	Significant
- **-**	**-** **-**	**-** **-**	**-** **-**	**-** **-**
P141	2.6003	0.0000	0.0004	Significant
P142	0.0980	0.9445	0.9696	Not Significant
P143	2.4304	0.0393	0.0992	Not Significant
P144	1.4673	0.0320	0.0870	Not Significant

The expression levels of these 41 proteins discriminate across the three disease states. However, there could be correlations among the proteins. Thus, a multi-protein analysis was performed to determine prime protein combinations.

Multi-protein analysis

With Lasso regularization incorporated with bootstrapping proteins, the corresponding ratio was obtained for each protein. Shown in Table 2 are the proteins whose expression levels were picked at least 10% of the time. There were 20 such proteins.

Table 2.
Results of lasso regularization with bootstrapping.

Protein	Freq.	Significant	Ratio	Protein	Freq.	Significant	Ratio
P92	190	190	1.00	P140	197	68	0.35
P22	189	189	1.00	P115	187	56	0.30
P139	186	186	1.00	P7	192	55	0.29
P141	179	179	1.00	P85	199	49	0.25
P10	186	183	0.98	P128	193	48	0.25
P134	171	162	0.95	P123	181	42	0.23
P20	184	156	0.85	P101	203	29	0.14
P69	173	101	0.58	P62	199	29	0.15
P38	183	84	0.46	P100	184	24	0.13
P132	185	84	0.45	P76	177	21	0.12

Columns "Frequency" and "Significant" of table 2 represent the number of times a protein appeared in the bootstrap sample and the number of times it was picked by lasso respectively. These 20 proteins are prime since it indicates that these are doing quite well together. Considering the top four dominant proteins, each time the corresponding protein appeared in the bootstrap sample, its expression levels have become striking all the time. However, when exploring low ratio proteins where its expression levels have not been picked each time it appeared in the bootstrap sample, it is apparent that the contribution from the presence of more dominant proteins became prominent. To further elaborate, if two proteins are important but they are correlated, and if both proteins are present in the bootstrap sample; lasso regularization would pick only one because the second protein adds to the trend only, but it is not vital on its own. Thus, the multi-protein analysis revealed protein combinations that join forces.

Combining univariate and multi-protein results

Results from univariate and multi-protein approaches were linked. Considering the 41 proteins whose expression levels were declared significant in the univariate context and the 20 proteins that were picked in the multi-protein context, the following were considered:

1. Intersection of the two sets of proteins
2. Union of the two sets of proteins

All 20 proteins that were picked by the multi-protein analysis were also significant in the univariate scenario; thus, the intersection consists of just these 20 proteins. These 20 prime proteins join forces to discriminate disease severity with an ordinal setting both by themselves and in combination with other proteins. A visual representation of the ordinal disease separation was also obtained. Principal component analysis (PCA) was applied to these 20 proteins and a summary of its results are as follows.

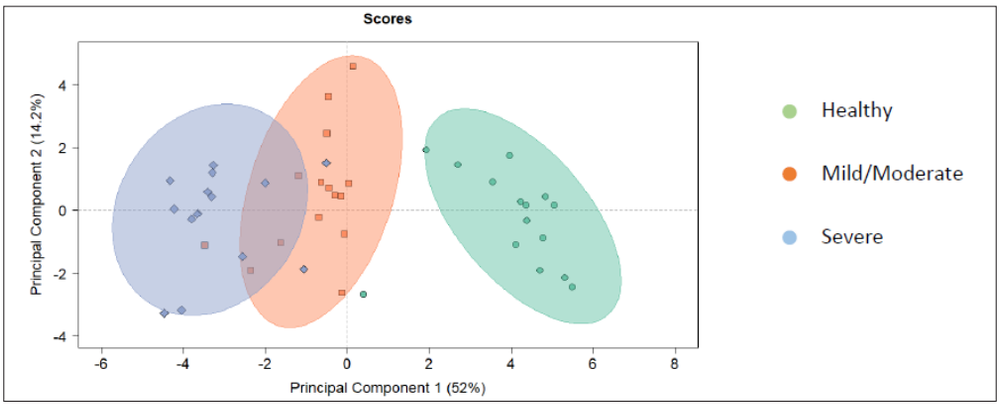
Table 3.
PCA results of first six PCs for 20 proteins.

Protein	**PC1**	**PC2**	**PC3**	**PC4**	**PC5**	**PC6**
Eigenvalue	6.98	0.82	0.57	0.41	0.26	0.21
Proportion of variance	52.0%	14.2%	8.6%	5.1%	3.3%	3.3%
Cumulative proportion of variance	52.0	66.2%	74.8%	79.9%	84.2%	87.5%

Considering eigenvalues, it was computed that a decent percentage of approximately 66% of cumulative proportion of variance could be explained by the first two principal components. A score plot was visualized considering the first two principal components as shown in Figure 2, and an ordinal separation was observed across the disease states with decent clarity. Though a minor disruption exists between the mild and severe groups, the ordinal subject separation is clearly visible.

Figure 2.

Score plot for the 45 subsets with the top selected proteins.



Furthermore, a union of proteins considering the univariate and multi-protein scenarios was explored too. The union set consists of 41 proteins. The reason for inspecting the union is because, though a protein is not working coherently with some other protein, it does not imply that the corresponding protein is not important. It can be effective on its own. Thus, this set of 41 proteins consists of the 20 proteins that work collaboratively to identify the trend pattern and the 21 proteins that is potent on its own. To visualize a score plot, PCA was applied to these 41 proteins. Its results are summarised in Table 4.

Table 4.

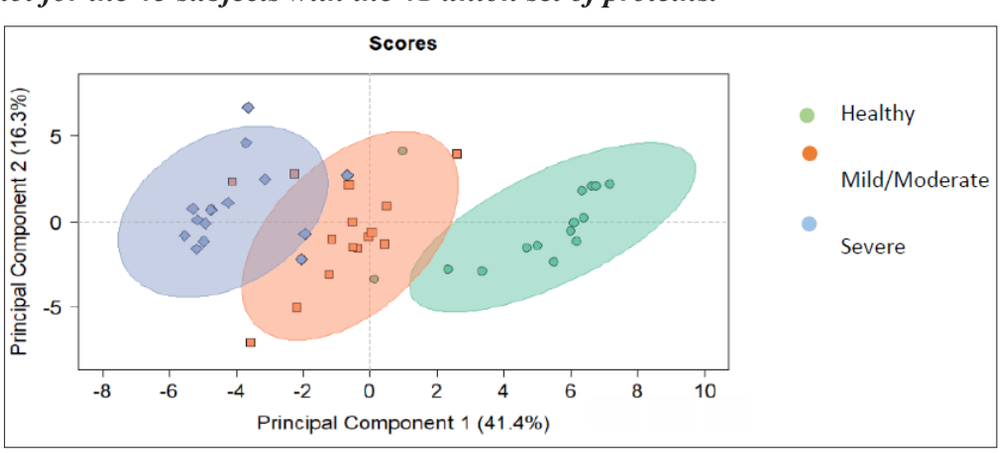
PCA results of the first 6 PCs for 41 proteins.

Protein	**PC1**	**PC2**	**PC3**	**PC4**	**PC5**	**PC6**
Eigenvalue	8.28	2.07	1.18	0.88	0.61	0.57
Proportion of variance	41.42%	16.25%	8.62%	5.22%	4.16%	3.75%
Cumulative proportion of variance	41.42%	57.67%	66.28%	71.50%	75.66%	79.41%

A decent percentage of information is explained by the first two principal components. Thus, they were used to visualize the score plot. As shown in Figure 3, the score plot visualized with these 41 proteins also confirms on the ordinal disease separation.

Figure 3.

Score plot for the 45 subjects with the 41 union set of proteins.



Conclusions

It can be concluded that the identified best featured 20 proteins have the capability to join forces satisfactorily to discriminate the disease condition in an ordinal manner. Furthermore, 21 more proteins were identified for having expression levels that were effective on their own. Thus, this paper has presented a methodology for analysing high dimensional data which have an ordinal response. Statistically prioritizing critical proteins is always encouraged. Since the identified prime proteins can differentiate between the disease severity satisfactorily, this would be helpful to pinpoint biomarkers and therapeutic targets for future diagnostics.

References

Alberts, B. (2002). Studying Gene Expression and Function - Molecular Biology of the Cell -NCBI Bookshelf. In *Molecular Biology of the Cell. 4th Edition.* https://www.ncbi.nlm.nih.gov/books/NBK26818/.

Ali, A., Rasheed, A., Siddiqui, A. A., Naseer, M., Wasim, S., & Akhtar, W. (2015). Non-Parametric Test for Ordered Medians: The Jonckheere Terpstra Test. *International Journal of Statistics in Medical Research*, 4, 203-207.

Amaratunga, D., Cabrera, J., Cherckas, Y., & Lee, Y.-S. (2012). *Ensemble classifiers* , 235-246. https://doi.org/10.1214/11-imscoll816.

Amaratunga, D., & Cabrera O-Book, J. (2008). Exploration and Analysis of DNA Microarray and Protein Array Data. https://www.wiley.com/en-us/9780470317129.

Ananth, C. V., & Kleinbaum, D. G. (1997). Regression models for ordinal responses: A review of methods and applications. *International Journal of Epidemiology*, 26(6), 1323-1333. https://doi.org/10.1093/ije/26.6.1323.

Archer, K. J., & Williams, A. A. A. (2012). L 1 penalized continuation ratio models for ordinal response prediction using high-dimensional datasets. *Statistics in Medicine*, 31(14), 1464-1474. https://doi.org/10.1002/sim.4484.

Brant, R. (1990). Assessing Proportionality in the Proportional Odds Model for Ordinal Logistic Regression. *Biometrics*, 46(4), 1171. https://doi.org/10.2307/2532457.

Efron, B. (1992). *Bootstrap Methods: Another Look at the Jackknife*. *June 1977*, 569-593. https://doi.org/10.1007/978-1-4612-4380-9_41.

Hesterberg, T. (2011). Bootstrap. *Wiley Interdisciplinary Reviews: Computational Statistics*, 3(6), 497-526. https://doi.org/10.1002/wics.182.

Jafari, M., & Ansari-Pour, N. (2019). Why , When and How to Adjust Your P Values ?. 20(4), 604-607. https://doi.org/10.22074/cellj.2019.5992.

Kingsmore, S. F. (2006). Multiplexed protein measurement: Technologies

and applications of protein and antibody arrays. *Nature Reviews Drug Discovery*, *5*(4), 310-320. https://doi.org/10.1038/nrd2006.

Liu, X. (2010). Ordinal Regression Analysis: Fitting the Continuation Ratio Model to Educational Data Using Stata Design. https://www.researchgate.net/publication/228434347.

Lubomirski, M., D'Andrea, M. R., Belkowski, S. M., Cabrera, J., Dixon, J. M., & Amaratunga, D. (2007). A consolidated approach to analyzing data from high-throughput protein microarrays with an application to immune response profiling in humans. *Journal of Computational Biology*, 14(3), 350-359. https://doi.org/10.1089/cmb.2006.0116.

Rusling, J. F., Kumar, C. V., Gutkind, J. S., & Patel, V. (2010). Measurement of biomarker proteins for point-of-care early detection and monitoring of cancer. *Analyst*, *135*(10), 2496-2511. https://doi.org/10.1039/c0an00204f.

Stoevesandt, O., Taussig, M. J., & He, M. (2009). Protein microarrays: High-throughput tools for proteomics. *Expert Review of Proteomics*, 6(2), 145-157. https://doi.org/10.1586/epr.09.2.

Sundström, S. (2010). Coding in Multiple Regression Analysis: A Review of Popular Coding Techniques. *Uppsala University*, 22. http://www.diva-portal.org/smash/get/diva2:325460/FULLTEXT01.pdf.

Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1), 267-288.

Yang, Q., Zhang, Y., Cui, H., Chen, L., Zhao, Y., Lin, Y., Zhang, M., & Xie, L. (2018). dbDEPC 3.0: the database of differentially expressed proteins in human cancer with multi-level annotation and drug indication. *Database: The Journal of Biological Databases and Curation*, . https://doi.org/10.1093/DATABASE/BAY015.

Zhu, J., & Hastie, T. (2004). Classification of gene microarrays by penalized logistic regression. *Biostatistics*, 5(3), 427-443.