Received: 12 June 2022 | Revised: 29 August 2022 | Accepted: 5 September 2022

DOI: 10.1002/pmic.202200252

Proteomics
Proteomics and Systems Biology

TECHNICAL BRIEF

muSignAI: An algorithm to search for multiple omic signatures with similar predictive performance

Bodhayan Prasad | **Anthony J. Bjourson** | **Priyank Shukla**

Personalised Medicine Centre, School of Medicine, Ulster University, C-TRIC Building, Altnagelvin Area Hospital, Glenshane Road, Londonderry BT47 6SB, UK

Correspondence
Priyank Shukla, Personalised Medicine Centre, School of Medicine, Ulster University, C-TRIC Building, Altnagelvin Area Hospital, Glenshane Road, Londonderry, BT47 6SB, UK.
Email: p.shukla@ulster.ac.uk

Abstract

Multidimensional omic datasets often have correlated features leading to the possibility of discovering multiple biological signatures with similar predictive performance for a phenotype. However, their exploration is limited by low sample size and the exponential nature of the combinatorial search leading to high computational cost. To address these issues, we have developed an algorithm muSignAI (multiple signature algorithm) which selects multiple signatures with similar predictive performance while systematically bypassing the requirement of exploring all the combinations of features. We demonstrated the workflow of this algorithm with an example of proteomics dataset. muSignAI is applicable in various bioinformatics-driven explorations, such as understanding the relationship between multiple biological feature sets and phenotypes, and discovery and development of biomarker panels while providing the opportunity of optimising their development cost with the help of equally good multiple signatures. Source code of muSignAI is freely available at https://github.com/ShuklaLab/muSignAI.

KEYWORDS
algorithm, bioinformatics, biomarkers, Olink, proteomics

In bioinformatics data analysis, sometimes we encounter a small sample size, for example in case of patient recruitment for rare diseases [1, 2]. However, the samples may frequently have a large number of features, such as those involving high throughput omics experiments [3]. Multiple features from such samples are often highly correlated, for example, due to their involvement in associated biological interactions, and hence multiple different combinations of the correlated features can perform similarly in predicting a given phenotype or outcome. This opens-up a possibility of discovering multiple feature sets, that are equally good in predicting the phenotype. These multiple signatures, for example, can help in understanding the relationship between multiple combinations of biological features and phenotypes. They can also help in optimising the cost of biomarker panel development for diagnostic or prognostic applications, by providing more options of signatures with equally good predictive performance. A recent study [4, 5] reported a method to obtain unbiased features in such situations involving low sample size and high feature space. However, the authors did not explore the possibility of recursive search of all possible feature combinations as its complexity is of $O(N^N)$ making it exponentially computer intensive with the increase in features. Taking inspiration from Enroth et al. study [6], we have developed and implemented *muSignAI* algorithm in R, which recursively explores all feature combinations by systematically deleting the selected features one-by-one to facilitate the discovery of multiple signatures that exhibit similar predictive performance (Figure 1).

Initially, 80% (default value) of the data along with all the features is randomly selected and feature selection is performed using generalised linear models (GLMs) (Figure 1, right panel). This process is repeated 100 times (default value) and feature sets showing an area

Abbreviations: FI, feature importance; GLMs, generalised linear models; LASSO, least absolute shrinkage and selection operator; muSignAI, multiple signature algorithm.

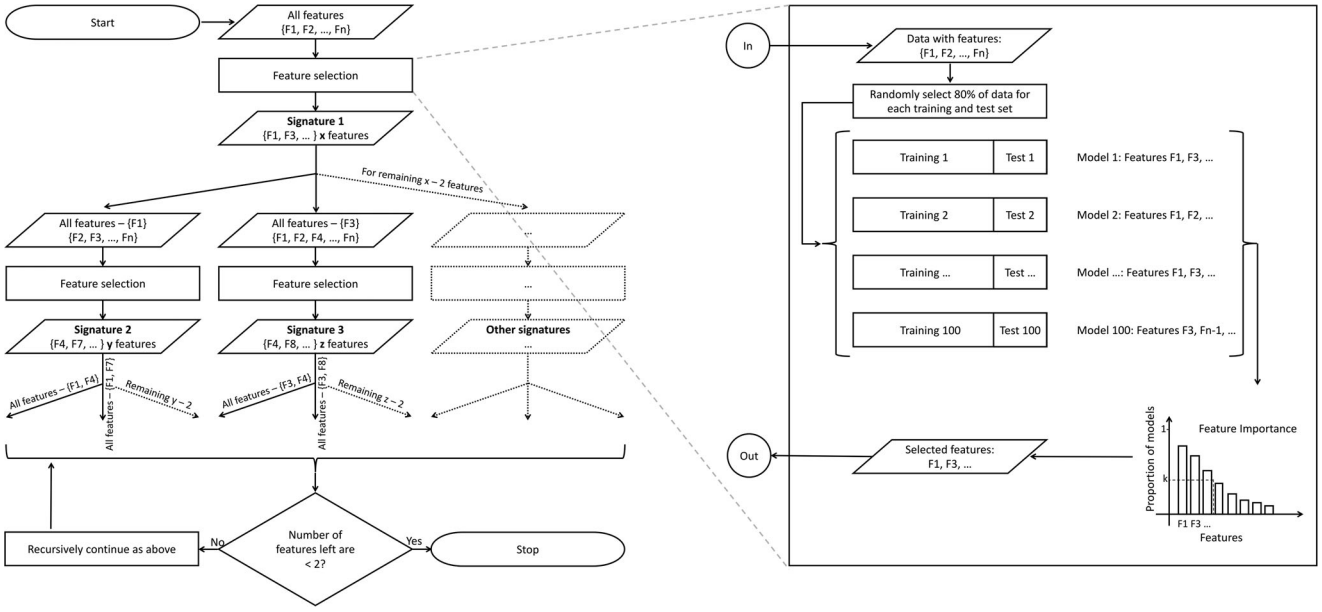

FIGURE 1 Flowchart presentation of the muSignAI algorithm. Oval shapes represent start/stop, circles represent in/out connectors, parallelogram boxes represent input/output, square boxes represent computation process, and rhombus box represent decision process. Signatures retrieved at each pass are presented in bold. The decision of whether the number of features are less than 2 will be taken at every input box. The feature selection box of the main algorithm on the left has been zoomed-out and presented on the right. k is the cut-off for the feature importance (FI) where FI is the proportion of models in which the feature has appeared.

under the receiver operating characteristic curve (AUC) > 0.5 are selected, where > 0.5 ensures better than a random selection. The proportion of models in which the feature has appeared is obtained as a measure of feature importance (FI). A threshold k = 0.9 (default value) is applied on the FI to select the first set of features and its predictive performance is calculated. In the next pass, features from this first feature set are removed one-by-one and the above process of feature selection is recursively repeated on the reduced feature set until the algorithm is left with only two features in each leaf at the bottom of the algorithm tree (Figure 1, left panel). All the feature sets along with their predictive performance are sent as an output of the function. Default values of all the parameters of the algorithm can be changed by the user. A sample evaluation case run with the final output is available in the example folder of the Github repository https://github.com/ShuklaLab/muSignAI, and have been presented and discussed later.

We have used least absolute shrinkage and selection operator (LASSO) for feature selection [7]. However, LASSO saturates with fewer features [8]. This was overcome by partly including Ridge regularisation, resulting in an Elastic Net model. We deployed GLMs to create an intuitive mathematical formulation with a linear combination of feature values. The GLM was an Elastic Net with alpha of 0.9, which implements regression with 90% LASSO and 10% Ridge regularization. The aim was to select non-correlated features, which was achieved by LASSO regularization. The *muSignAI* algorithm is developed in an open-source platform R version 3.6 [9]. The basic data pre-processing was done using the *caret* package [10]. The model was built using the *glmnet* package [8]. ROC was built using the *pROC* package [11]. The algorithm requires a dataframe of features and target variable. The R function *muSignAI()* reads the input data file along with feature space to search

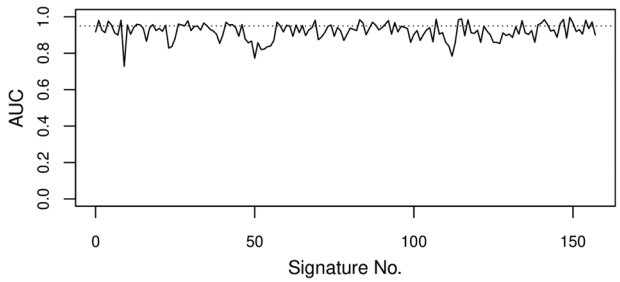

FIGURE 2 AUC values of the first 158 unique signatures. Dotted line shows 0.95 cut-off value for AUC. Forty-seven peaks above the dotted line are the 47 signatures presented in Table 1.

from and target variable. It then outputs multiple signatures along with their performances as a dataframe. The algorithm is provided as a tool on the open-source GitHub repository: https://github.com/ShuklaLab/muSignAI.

For the evaluation of muSignAI algorithm, we have taken the publicly available dataset from Brunner et al. study [12] which includes 77 patient samples and 91 protein features from Olink Proteomics (www.olink.com) CVD-II panel. Samples were grouped as a healthy cohort ($n = 18$) and an atopic dermatitis cohort ($n = 59$). We ran *muSignAI()* with a FI threshold of k = 0.9 on this dataset, which generated 1984 signatures as an output; out of which, 158 were unique. Figure 2 reports the AUC performance of these 158 signatures; out of which, 47 had greater than 0.95 AUC (Table 1). The AUC has been recommended to be used in preference to overall accuracy when evaluating machine learning (ML) algorithms [13, 14], which should equally apply when evaluating different multiple signatures. Other bioinformatics studies also report AUC

TABLE 1 Selected signature proteins (feature sets) based on AUC > 0.95.

Signature No.	Signature proteins	AUC
2	IDUA, IL16, DECR1, SORT1, GT	0.98
5	GT, IL16, PARP-1, STK4, SORT1	0.98
6	IL16, STK4, PARP-1, SORT1	0.95
9	STK4, GT, PARP-1, SCF, SORT1	0.98
11	PARP-1, SORT1, STK4, GT	0.95
14	IL16, PARP-1, GT, SORT1, GH	0.96
15	IL16, GT, PARP-1, TIE2	0.96
19	GT, NEMO, PARP-1, TIE2, IL-4RA, SORT1	0.96
23	TIE2, GT, NEMO, PARP-1	0.95
27	GT, NEMO, PARP-1, TIE2, IL-4RA	0.96
28	IL16, PARP-1, GT, TIE2	0.96
30	IL16, PARP-1, GT, SORT1, TIE2	0.98
35	GT, IL16, PARP-1, SORT1, DCN	0.97
36	IL16, PARP-1, SORT1, GT	0.95
42	STK4, MMP-12, PARP-1, GT, SORT1	0.97
43	GT, PARP-1, STK4, SORT1	0.95
44	GT, IL-4RA, NEMO, PARP-1, SORT1, TIE2	0.96
47	MMP-12, PARP-1, GT,STK4, TIE2	0.96
58	PARP-1, STK4, GT, MMP-12, SORT1	0.97
59	GT, MMP-12, NEMO, PARP-1, TIE2	0.95
61	GT, NEMO, PARP-1, TIE2	0.95
64	GT, STK4, PARP-1, SORT1	0.95
66	IL16, GT, PARP-1, SORT1	0.95
70	GT, STK4, PARP-1, SCF, SORT1	0.98
75	GT, PARP-1, SORT1, STK4	0.95
84	DECR1, GT, IL16, PARP-1, SORT1	0.98
85	DECR1, IL16, PARP-1, SORT1, GH	0.97
88	DECR1, IL16, PARP-1, SORT1	0.97
89	IL16, PARP-1, STK4, SORT1	0.95
92	DECR1, MMP-12, PARP-1, SORT1, GT	0.96
93	DECR1, IL16, GT, SORT1	0.98
95	DECR1, IL16, PARP-1, GT	0.97
108	IDUA, IL16, SORT1, GT	0.99
115	IDUA, IL16, ADM, SORT1	0.98
116	IDUA, IL16, SORT1, PARP-1	0.99
118	IDUA, IL16, SORT1	0.98
135	IDUA, IL16, ADM, DECR1, SORT1	0.98
140	CCL17, ADM, DECR1, IL16, PARP-1	0.96
141	CCL17, DECR1, SORT1, PARP-1	0.97
142	DECR1, IL16, PARP-1, SORT1, GT	0.98
143	CCL17, PARP-1, SORT1, IL16	0.96
147	CCL17, DECR1, SORT1	0.96
148	CCL17, IL16, SORT1, PARP-1, ADM	0.99
150	CCL17, DECR1, ADM, PARP-1, SORT1	1.00
151	CCL17, DECR1, PARP-1, SORT1	0.97
155	CCL17, ADM, DECR1, SORT1	0.98
157	CCL17, ADM, SORT1	0.97

as one of the major performance metrics when applying ML [15–17]. Above case run took 20.05 h on a PowerEdge R740XD server. However, since the muSignAl algorithm implements a recursive function, the computational run-time may vary depending on the dataset and computational resources. For example, if most of the feature variables present in the dataset are predictive (i.e., significantly associated with a phenotype), the algorithm will identify more signatures and hence will take longer. Similarly, if most of the feature variables present in the dataset are non-predictive, the algorithm will stop quickly.

Thus, muSignAl algorithm can discover multiple omic signatures with similar predictive performance. It will be useful in analysing multi-dimensional omic datasets especially those with low sample sizes often encountered for example in studies of rare diseases. It will be applicable in various bioinformatics driven explorations, such as understanding the relationship between multiple combinations of biological features and phenotypes, and discovery and development of biomarker panels while providing the opportunity of optimising their development cost with the help of equally good multiple signatures.

ACKNOWLEDGEMENTS

B.P. would like to acknowledge the funding support of Vice-Chancellor's Research Scholarship (VCRS), Ulster University. A.J.B. would like to acknowledge funding support by a programme grant jointly from the European Union (EU) Regional Development Fund (ERDF) EU Sustainable Competitiveness Programme for Northern Ireland, the Northern Ireland Public Health Agency (HSC R&D) and Ulster University. P.S. would like to acknowledge funding support from Innovate UK NxNW ICURe programme and UKRI NCSi4P programme 'Optimal cellular assays for SARS-CoV-2 T-cell, B-cell and innate immunity'. A.J.B. and P.S. would like to acknowledge funding support by a programme grant jointly from Science Foundation Ireland (SFI), Republic of Ireland and Department for the Economy (DfE), Northern Ireland, UK, 'COVRES: Understanding the host-virus response in patients with mild versus serious disease'. A.J.B. and P.S. would like to acknowledge funding support by Northern Ireland Public Health Agency (HSC R&D Division), 'COVRES2: Identifying temporal immune responses associated with Covid-19 severity'.

CONFLICT OF INTEREST

The authors have declared no conflict of interest.

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analysed in this study.

ORCID

Bodhayan Prasad https://orcid.org/0000-0002-7383-2460
Priyank Shukla https://orcid.org/0000-0002-4985-9305

REFERENCES

1. Frésard, L., Smail, C., Ferraro, N. M., Teran, N. A., Li, X., Smith, K. S., Bonner, D., Kernohan, K. D., Marwaha, S., Zappala, Z., Balliu, B., Davis, J. R., Liu, B., Prybol, C. J., Kohler, J. N., Zastrow, D. B., Reuter, C. M., Fisk, D. G., Grove, M. E., … Montgomery, S. B. (2019). Identification of rare-disease genes using blood transcriptome sequencing and large control cohorts. *Nature Medicine*, *25*(6), 911–919.
2. Papuc, S. M., Abela, L., Steindl, K., Begemann, A., Simmons, T. L., Schmitt, B., Zweier, M., Oneda, B., Socher, E., Crowther, L. M., Wohlrab, G., Gogoll, L., Poms, M., Seiler, M., Papik, M., Baldinger, R., Baumer, A., Asadollahi, R., Kroell-Seger, J., … Rauch, A. (2019). The role of recessive inheritance in early-onset epileptic encephalopathies: A combined whole-exome sequencing and copy number study. *European Journal of Human Genetics: EJHG*, *27*(3), 408–421.
3. Koh, Y., Park, I., Sun, C. H., Lee, S., Yun, H., Park, C. K., Park, S. H., Park, J. K., & Lee, S. H. (2015). Detection of a distinctive genomic signature in rhabdoid glioblastoma, a rare disease entity identified by whole exome sequencing and whole transcriptome sequencing. *Translational Oncology*, *8*(4), 279–287.
4. Shi, L., & Brunius, C. (2018). A brief tutorial on MUVR: Multivariate methods with unbiased variable selection in R. https://gitlab.com/CarlBrunius/MUVR/-/blob/20210719_01/Tutorial/MUVR_Tutorial.pdf
5. Shi, L., Westerhuis, J. A., Rosén, J., Landberg, R., & Brunius, C. (2019). Variable selection and validation in multivariate modelling. *Bioinformatics (Oxford, England)*, *35*(6), 972–980.
6. Enroth, S., Berggrund, M., Lycke, M., Broberg, J., Lundberg, M., Assarsson, E., Olovsson, M., Stå¥Lberg, K., Sundfeldt, K., & Gyllensten, U. (2019). High throughput proteomics identifies a high-accuracy 11 plasma protein biomarker signature for ovarian cancer. *Communications Biology*, *2*, 221.
7. Fonti, V., & Belitser, E. (2017). Feature selection using lasso. *VU Amsterdam Research Paper in Business Analytics*, *30*, 1–25.
8. Friedman, J., Hastire, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, *33*(1), 1–22.
9. Core Team, R. (2019). *R: A language and environment for statistical computing*. R foundation for statistical computing. https://www.R-project.org/
10. Kuhn, M. (2021). *Caret: Classification and regression training*. R package version 6.0-90. https://CRAN.R-project.org/package=caret
11. Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J. C., & Müller, M. (2011). pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, *12*, 77.
12. Brunner, P. M., Suárez-Fariñas, M., He, H., Malik, K., Wen, H. C., Gonzalez, J., Chan, T. C. C., Estrada, Y., Zheng, X., Khattri, S., Dattola, A., Krueger, J. G., & Guttman-Yassky, E. (2017). The atopic dermatitis blood signature is characterized by increases in inflammatory and cardiovascular risk proteins. *Scientific Reports*, *7*(1), 8707.
13. Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, *30*(7), 1145–1159.
14. Larrañaga, P., Calvo, B., Santana, R., Bielza, C., Galdiano, J., Inza, I., Lozano, J. A., Armañanzas, R., Santafé, G., Pérez, A., & Robles, V. (2006). Machine learning in bioinformatics. *Briefings in Bioinformatics*, *7*(1), 86–112.
15. Hung, T., Le, N., Le, N., Van Tuan, L., Nguyen, T. P., Thi, C., & Kang, J. H. (2022). An AI-based prediction model for drug-drug interactions in osteoporosis and Paget's diseases from SMILES. *Molecular Informatics*, *41*(6), 2100264.
16. Le, N., & Ho, Q. T. (2022). Deep transformers and convolutional neural network in identifying DNA N6-methyladenine sites in cross-species genomes. *Methods (San Diego, Calif.)*, *204*, 199–206.
17. Mustafić, S., Brkić, S., Prnjavorac, B., Sinanović, A., Porobić Jahić, H., & Salkić, S. (2018). Diagnostic and prognostic value of procalcitonin in patients with sepsis. Medicinski glasnik: Official publication of the medical association of Zenica-Doboj Canton. *Bosnia and Herzegovina*, *15*(2), 93–100.

How to cite this article: Prasad, B., Bjourson, A. J., & Shukla, P. (2023). muSignAI: An algorithm to search for multiple omic signatures with similar predictive performance. *Proteomics*, *23*, e2200252. https://doi.org/10.1002/pmic.202200252