Bioinformatics, 38(6), 2022, 1773–1775
https://doi.org/10.1093/bioinformatics/btab862
Advance Access Publication Date: 25 December 2021
Applications Note

Data and text mining

MAINE: a web tool for multi-omics feature selection and rule-based data exploration

Aleksandra Gruca [1,*], Joanna Henzel [1], Iwona Kostorz[2], Tomasz Stęclik[2], Łukasz Wróbel [1] and Marek Sikora [1,2,*]

[1]Department of Computer Networks and Systems, Silesian University of Technology, 44-100 Gliwice, Poland and [2]Łukasiewicz Research Network – Institute of Innovative Technologies EMAG, 40-189 Katowice, Poland

*To whom correspondence should be addressed.
Associate Editor: Jonathan Wren

Received on March 15, 2021; revised on September 22, 2021; editorial decision on December 9, 2021; accepted on December 22, 2021

Abstract

Summary: Patient multi-omics datasets are often characterized by a high dimensionality; however, usually only a small fraction of the features is informative, that is change in their value is directly related to the disease outcome or patient survival. In medical sciences, in addition to a robust feature selection procedure, the ability to discover human-readable patterns in the analyzed data is also desirable. To address this need, we created MAINE—Multi-omics Analysis and Exploration. The unique functionality of MAINE is the ability to discover multidimensional dependencies between the selected multi-omics features and event outcome prediction as well as patient survival probability. Learned patterns are visualized in the form of interpretable decision/survival trees and rules.
Availability and implementation: MAINE is freely available at maine.ibemag.pl as an online web application.
Contact: aleksandra.gruca@polsl.pl or marek.sikora@polsl.pl
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Multi-omics data are characterized by a high dimensionality while, at the same time, only a small fraction of the features describing these data are informative. Therefore, the first step of the multi-omics data analysis usually consists of selecting the most relevant set of features describing the available data. However, in medical sciences, it is crucial that a data analysis method is also able to discover human-readable patterns hidden in the analyzed data. To address this need, we developed MAINE—a web application for feature selection and explanatory analysis of multi-omics data. MAINE provides illustrative reports describing multidimensional dependencies between features. Our approach is based on the observation that the most interesting features are related to event outcome prediction (classification) or patient survival probability (survival analyses), and those dependencies can be represented in a form of trees and rules (Burkart and Huber, 2021; Ishwaran *et al.*, 2008; Sikora *et al.*, 2019).

With the growing number of multi-omics datasets, there is an urgent need to develop applications that provide the users with an intuitive interface, making the analysis available also to domain experts who do not have programming skills. Recently, several software tools have been developed providing a wide spectrum of different methods and approaches both for feature selection as well as data visualization and explanation. Those tools are available either in a form of a stand-alone software packages or as web services. Typically, stand-alone software packages provide the user with extended functionality and wrapper functions build on top of existing libraries. However, using such tools require programming skills, as well as the access to computers with high computing power when analyzing larger datasets. MixOmics (Rohart *et al.*, 2017) is the example of the stand-alone R-package for multi-omics data analysis. It provides a wrapper function for a set of statistical methodologies to analyze high-throughput data as well as a package for data visualization such as relevance networks, clustered image maps and circle plots. The tool, however, does not allow to perform survival analysis. IntLIM (Siddiqui *et al.*, 2018) is a tool that integrates metabolomics and transcriptomics data and is also available as an R-package with shiny-based GUI . Here, gene-metabolite associations that are specific to a particular phenotype are uncovered by linear modeling approach. Another example of a stand-alone application is PROMO (Netanely *et al.*, 2019), a Windows application with a fully interactive graphical user interface that runs over the MATALB environment This tool provides the user with a set of standard methods for genomics cancer data analysis starting from data prepossessing and visualization trough clustering, decision trees generation and survival analysis to biomarker discovery. However, most of those methods are dedicated only to the single type of data and their multi-omics functionality is limited to feature correlation analysis in two selected omics or clustering the samples based on several omic matrices simultaneously. The example of a web-based platform for multi-omics data analysis is PaintOmics 3 (Hernández-

de Diego *et al.*, 2018) that provides the user with the methods for feature matching, pathway enrichment analysis and pathway-based results visualization. MultiSLIDE (Ghosh *et al.*, 2021) is a web-based interactive tool that allows to identify molecular signatures by statistical analysis and simultaneous visualization of molecular features in heatmaps of multi-omics datasets. MiBiOmics (Zoppi *et al.*, 2021) is another example of a web-based application for multi-omics data filtration, normalization and transformation. The main functionality of this tool is data exploration based on PCA and PCoA plots, and network inference based on Weighted Gene Correlation Network Analysis. Mergeomics (Ding *et al.*, 2021) is a tool which, after filtering omics marker redundancies, uses Marker Set Enrichment Analysis to summarizes enrichment of disease/trait omics markers in sets of functionally related genes and Meta-MSEA to integrate of multiple datasets of the same omics type or multiple omics types.

MAINE provides a statistical and machine learning-based frameworks for explaining multidimensional dependencies between selected multi-omics attributes and event outcome or patient survival time based on decision/survival trees and decision/survival rules. In our classification and survival reports, we focus on explaining the relation between selected attribute values and outcome prediction/survival time. Our approach differs from other tools by providing not only a list of relevant features but also by creating subgroups of patients that are similar according to the criteria presented in a tree node or rule premise. The subgroups of patients can be then described by attributes and their values, and characterized by their outcome or survival time. Tree generation methods are based on divide-and-conquer (DnC) approach and rule generation methods are based on separate-and-conquer (SnC) approach. Both approaches allow to create subgroups of similar patients described by interpretable rules; however, the discovered patterns might differ due to different algorithmic approaches. The DnC strategy does not allow examples to be covered by many rules, whereas the SnC approach lacks this limitation. In addition rules generated from decision trees may contain redundant features, unlike SnC-based rules where each rule is induced separately. Moreover, our service provides the unique method for feature selection. The method is based on Rough Set Theory (RST) and Monte Carlo-based Approximate Relative Reducts (MCARR) feature selection (Riza *et al.*, 2014). Within the RST-MCARR approach a minimal set of features is selected that ensures the same discernability between examples from different decision classes as the whole (non-reduced) feature set.

2 Materials and methods

MAINE web application allows the user to submit patient data that include multi-omics experimental measurements obtained with different high-throughput platforms. Currently, accepted data types are probe measurements from methylation assays, RNA-Seq expression data and copy number variance (CNV). Feature selection can be performed based on statistical approach typical for downstream methylation/expression data analyses or based on the RST-MCARR approach. Three workflows for data analysis are available: (i) attribute selection and normalization are done separately for each multi-omics data type and are based on statistical approach for methylation and RNA-Seq data, and the RST-MCARR algorithm for CNV data, (ii) attribute selection is done separately for each data type and is based on RST-MCARR, (iii) all attributes are first combined into a single table and then attribute selection is performed using RST-MCARR. Detailed explanatory charts of the workflows are presented in Supplementary Section S1. After the feature selection process, the results can be downloaded by the user for further analyses or as a basis to generate explanatory classification or survival reports.

The second part of the MAINE application provides the user with explanatory reports showing important features and discovered patterns. To obtain the highly interpretable results, we provide decision trees and rules for classification reports and survival trees, and rules for survival reports. Both the trees and the rules enable to divide observations (patients) into subgroups with different outcome/survivability characteristics. Both methods allow not only to identify the variables that have significant impact on the outcome prediction/survival time, but also enable to model nonlinear dependencies and interactions between the variables.

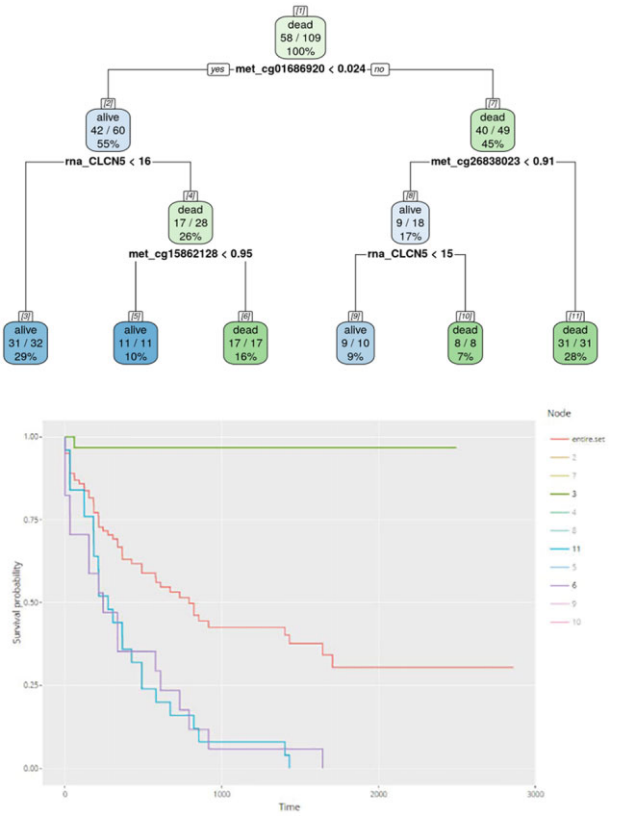

Fig. 1. Decision tree. Each node provides: the name of the outcome value indicating the majority class assigned to that node; ratio of patients with the node outcome to all patients assigned to that node; the information about the percentage of the cases assigned to the node. Survival curves drawn for the selected (3, 6, 11) leaves of the decision tree. Red line represents the survival curve calculated for the entire dataset. Each curve shows the probability of staying alive for a certain amount of time after the treatment

Each classification report contains decision trees generated with two different methods: rpart (Therneau *et al.*, 2019) and ctrees (Hothorn *et al.*, 2006), a set of decision rules (Gudyś *et al.*, 2020), and a list of the most important features. If patient data include survival information, it can be used in the classification report to assign the Kaplan–Meier survival function to the discovered subgroups of observations. This approach allows not only to highlight the importance of particular features but also to understand relations between conditional features and decision attribute. Survival report provides information on which features are related to the occurence of an event of interest.. This relation can be represented either in a form of survival trees or survival rules.

3 Results and discussion

Here, we briefly provide an illustrative example of data analysis with the use of MAINE obtained with the (iii) workflow. Analyzed dataset was derived from the TCGA study of Acute Myeloid Leukemia (AML; Cancer Genome Atlas Research Network, 2013). In our example, we focused on experimental data from RNA expression, DNA-methylation profiling and CNV. After filtering for the samples that had measurements for all three experimental platforms, we obtained data for 109 patients, including 51 patients with survival status *alive* and 58 patients with survival status *dead*.

Input dataset included: 40 571 RNA-Seq, 321 500 methylation and 19 482 CNV features. After the selection process, the number of

features was significantly reduced to 6. The number of selected features depends on the RST-MCARR parameter settings and selection can be made less restrictive by tuning the method parameters. The selected features were probes from methylation dataset: cg01686920 (targeting CRNDE, IRX5), cg26838023 (targeting LINC00028, REM1), cg20895586 (targeting RP11-126F18.2), cg15862128 (targeting MIR1193, MIR494), one gene from RNA-Seq dataset: CLCN5 and one gene from CNV dataset: ACTN2. By analyzing the literature, we notice that those genes frequently take part in pathways and processes related to tumor development or suppression. For example, among related pathways of ACTN2 is RET signaling pathway and various AML subtypes are dependent on expression of the RET receptor tyrosine kinase (Rudat *et al.*, 2016). Another example is MIR1193 which suppresses the proliferation and invasion of human T-cell leukemia cells through directly targeting TM9SF3 (Shen *et al.*, 2017). Interestingly, among selected genes we can also find CLCN5 which, based on current literature reports, is not related to leukemia prognostic. Since we can see this gene frequently appear both in discovered trees and rules it can be a potential target for future research for new prognostic markers in AML.

Figure 1 presents the decision tree and the survival curves obtained as the results of our analyses. Highlighted curves are calculated on the basis of the examples assigned to the leaves labeled as 3, 6, 7 and for the entire dataset. The biological description of the results and the full classification report is provided in Supplementary Sections S2 and S3.

The biological analysis of the results obtained for the TCGA Lung Adenocarcinoma study and full classification report is provided in Supplementary Sections S2 and S3.

4 Conclusions

In this work we presented MAINE, a web server that provides two main functionalities: multi-omics feature selection and classification/survival explanatory reports generation. It is the first web application enabling RST-MCARR-based feature selection of multi-omics data. MAINE not only provides a list of selected features but also supports the user in understanding multidimensional dependencies hidden in data by explaining how feature values are related to the event outcome (classification reports) or survival probability (survival reports).

Funding

This work was partially funded by the Polish National Centre for Research and Development [Grant No. STRATEGMED3/304586/5/NCBR/2017]; the Statutory Research Fund of Łukasiewicz Research Network—Institute of Innovative Technologies EMAG; and the Young Researchers funds of Department of Computer Networks and Systems, Faculty of Automatic Control, Electronics and Computer Science, Silesian University of Technology, Gliwice, Poland [Project No. 02/120/BKM21/0012].

Conflict of Interest: none declared.

Data availability statement

The data underlying this article are available in the Genomic Data Commons Data Portal at https://portal.gdc.cancer.gov and on the MAINE website at maine.ibemag.pl/#exemplaries.

References

Burkart,N. and Huber,M.F. (2021) A survey on the explainability of supervised machine learning. *J. Artif. Intell. Res.*, **70**, 245–317.

Cancer Genome Atlas Research Network (2013) Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *N. Engl. J. Med.*, **368**, 2059–2074.

Ding,J. *et al.* (2021) Mergeomics 2.0: a web server for multi-omics data integration to elucidate disease networks and predict therapeutics. *Nucleic Acids Res.*, **49**, W375–W387.

Ghosh,S. *et al.* (2021) MultiSLIDE is a web server for exploring connected elements of biological pathways in multi-omics data. *Nat. Commun.*, **12**, 2279.

Gudyś,A. *et al.* (2020) RuleKit: a comprehensive suite for rule-based learning. *Knowl. Based Syst.*, **194**, 105480.

Hernández-de Diego,R. *et al.* (2018) PaintOmics 3: a web resource for the pathway analysis and visualization of multi-omics data. *Nucleic Acids Res.*, **46**, W503–W509.

Hothorn,T. *et al.* (2006) Unbiased recursive partitioning: a conditional inference framework. *J. Comput. Graph. Stat.*, **15**, 651–674.

Ishwaran,H. *et al.* (2008) Random survival forests. *Ann. Appl. Stat.*, **2**, 841–860.

Netanely,D. *et al.*. (2019) PROMO: an interactive tool for analyzing clinically-labeled multi-omic cancer datasets. BMC Bioinformatics, **20**, 732 10.1186/s12859-019-3142-531878868PMC:

Riza,L.S. *et al.* (2014) Implementing algorithms of rough set theory and fuzzy rough set theory in the R package "roughsets". *Inform. Sci.*, **287**, 68–89.

Rohart,F. *et al.* (2017) MixOmics: an R package for 'omics feature selection and multiple data integration'. *PLoS Comput. Biol.*, **13**, e1005752.

Rudat,S. *et al.* (2016) The RET receptor tyrosine kinase promotes acute myeloid leukemia through protection of FLT3-ITD mutants from autophagic degradation. *Blood*, **128**, 2849.

Shen,L. *et al.* (2017) miR-1193 suppresses the proliferation and invasion of human T-cell leukemia cells through directly targeting the transmembrane 9 superfamily 3 (TM9SF3). *Oncol. Res.*, **25**, 1643–1651.

Siddiqui,J.K. *et al.* (2018) IntLIM: integration using linear models of metabolomics and gene expression data. *BMC Bioinform.*, **19**, 81.

Sikora,M. *et al.* (2019) Guider: a guided separate-and-conquer rule learning in classification, regression, and survival settings. *Knowl. Based Syst.*, **173**, 1–14.

Therneau,T. *et al.* (2019) Package 'rpart'. https://cran.r-project.org/web/packages/rpart/rpart.pdf (26 October 2020, date last accessed).

Zoppi,J. *et al.* (2021) MiBiOmics: an interactive web application for multi-omics data exploration and integration. *BMC Bioinform.*, **22**, 6.