Simpati: Network-based System for Patients' Classification Reveals Disease Specific Pathways Driven by Cohesive Communities

Luca Giudice
A.I. Virtanen Institute for Molecular Sciences
University of Eastern Finland
Kuopio, Finland
luca.giudice@uef.fi

Claudia Mengoni, Rosalba Giugno
Department of Informatics
University of Verona
Verona, Italy
claudia.mengoni@univr.it, rosalba.giugno@univr.it

Abstract—**Patient classifiers should be able to rely on the strength of machine learning methodologies while not losing biological interpretability. So far, most of the developed methods lack in one of the two aspects. We propose Simpati, a pathway-based tool for patient classification, which enables accurate classification focusing on the detection of relevant biological features and patient cohesive communities. The tool makes it possible to classify patients and investigate the features which were mostly representative of each class. It presents ad-hoc algorithms for the processing of patient similarity networks and proposes an effective simulation strategy as a recommender system to predict a patient's class based on graph topology. Its computational performance, classification performance and biological validation were performed on genetic data from different types of cancer and compared favorably with state-of-the-art competitors.**

Pathway-based classification; Network-based propagation; Patient similarity network; Subgroup cohesive algorithm.

I. INTRODUCTION

High-throughput biological data provide valuable information to clinicians for the prognosis and treatment response of patients. They offer quantitative and qualitative evidences to biomedical scientists for developing a study or confirming wet-lab results. Pathway-based analysis is a technique to investigate these data and detect molecular mechanisms related to the patients [1][2]. The pathway space is more robust to noise than the single feature level, summarizes the information of multiple patient’s molecules into the pathway activity (inhibited or activated), reduces the model complexity and maintains predictive accuracy [3][4]. Nowadays, pathway-based analysis is mostly performed through enrichment tools, fundamental methods which provide to clinicians understanding of the cellular functions affected in a patient, so that they can better define a disease phenotype and manually classify patients. Although some attempts have been made to couple pathway enrichment with classification [5], pathway-based classifiers that do not require pathway enrichment (i.e., supervised classifiers able to integrate simple pathway information to classify biological samples), are not yet strongly developed.

Among them there are two classifiers that exploit the idea of pathway. The first is PASNet [6], which incorporates biological pathways in a Deep Neural Network. The neural network is composed by an input gene layer, a pathway layer, a hidden layer that represents hierarchical relationships among biological pathways and an output layer that corresponds to the patient classes. The second is netDx [7] and represents pathways thanks to the Patient Similarity Network (PSN) paradigm. In a PSN, each node is an individual patient and an edge between two patients corresponds to pairwise similarity for a given patient’s feature (e.g., gender, height, gene expression). All the user-provided data are converted into PSNs and molecular data can be converted into networks representing pathways. This made netDx a pioneer classifier able to combine multi-omics and pathway specific features. The decision system of the software relies on GeneMANIA [8], state-of-art gene function predictor, to select the best patient similarity networks and to use them in the classification. netDx revealed to be better than canonical machine learning algorithms and to provide a good level of interpretability based on the network's graphical representation. However, the software requires the user to define a similarity measures for each input data and manually tune hyper-parameters, making the results highly dependent on users choices. Additionally, netDx does not consider the topology of the networks for inferring the relationships between training and testing patients, providing a black box prediction difficult to interpret.

A classifier should be able to benefit both from the interpretability of pathway-based enrichment tools and the strength of machine learning methodologies [9]. We want to stand up to the challenge by proposing the pathway-based classifier Simpati. Our method provides a novel feature-selection strategy for classifiers based on patient similarity networks, implements a subgroup cohesive algorithm for extracting patient communities in PSNs and proposes an effective simulation strategy to predict a patient's class based on graph topology. Plus, the method introduces ad-hoc operations for genetic data to reduce the number of hyper-parameters, similarity measures, or external software that the user has to define or install, it naturally handles outliers and integrates a graphical user interface to allow the visualization of the networks.

This text is structured as follows: in the Methods section the general workflow of the tool is described and different subsections detail the implementation of each step. These include all steps necessary for data preparation, feature selection and prediction, as well as a description of required input data and possible downstream analyses. In the Results section, Simpati performances are compared to those of two state-of-the-art competitors, both in terms of computational requirements, classification performance and biological interpretation. Finally, the Conclusions section remarks the impact of this classifier, its limitations, and its future development.

II. Methods

In this section, a general overview of Simpati's workflow is given, then the other subsections detail the specific aspects of implementation of each step. The R package to use Simpati and its graphical interface can be found online [10][11].

A. Overview

Simpati is a binary patient classifier, which exploits the similarity of patients' molecular profiles at the pathway-level. An overview of the method is shown in Figure 1. It takes as input patients' genetic profiles similarly to a gene differential analysis setting where counts have been library normalized and two classes are to be compared. The method has to be provided also with a list of pathways and a gene interaction network. Simpati transforms the profile of each individual patient to take into account the interconnectivity of genes. Each profile is propagated over the interaction network and the transformed data are used in the downstream analysis. Next, Simpati creates, selects and cleans PSNs. For each set of genetic features falling into a pathway, Simpati creates a Pathway-Specific PSN (psPSN), tests if the two patient classes show separability and finds cohesive communities inside each class. A psPSN is retained if it shows a strong intra-similarity between patients of one class, while having at the same time a weak intra-class similarity in the other class and a weak inter-class similarity. Once a network is selected as significant, Simpati removes patients showing an outlier pathway activity as compared to the rest of patients in the same class. Signature pathways are then used to classify patients of unknown class, based on their similarity to labeled patients.

B. Network-based data preparation

The first step is the transformation of patients' biological profiles using a network-based propagation algorithm. Each single-level feature gets a new value based on its a priori

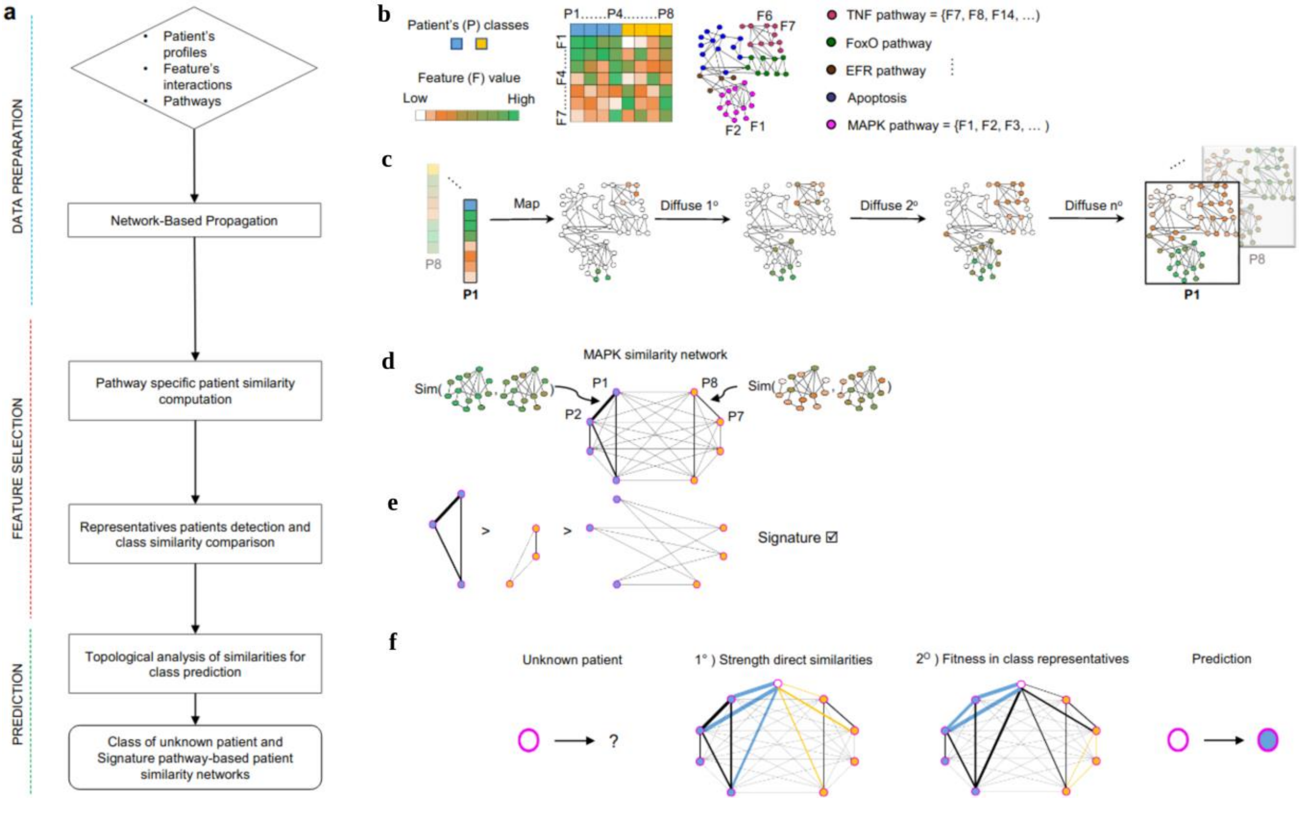
a
DATA PREPARATION
FEATURE SELECTION
PREDICTION
Patient's profiles
Feature's interactions
Pathways
Network-Based Propagation
Pathway specific patient similarity computation
Representatives patients detection and class similarity comparison
Topological analysis of similarities for class prediction
Class of unknown patient and Signature pathway-based patient similarity networks
b
Patient's (P) classes
Feature (F) value
Low
High
P1......P4........P8
F1
F4
F7
F6
F7
F2
F1
TNF pathway = {F7, F8, F14, ...)
FoxO pathway
EFR pathway
Apoptosis
MAPK pathway = {F1, F2, F3, ...)
c
P8
P1
Map
Diffuse 1°
Diffuse 2°
Diffuse n°
P8
P1
d
MAPK similarity network
Sim(
P1
P2
P8
P7
Sim(
e
>
>
Signature ☑
f
Unknown patient
1°) Strength direct similarities
2°) Fitness in class representatives
Prediction
?

Figure 1. Simpati's workflow: (a) Overview of the main steps. (b) Input: matrix of features by patients, feature interaction network, features grouped by pathways. (c) Each patient's profile is propagated on the interaction network. (d) Within each pathway's subnetwork patient's similarity is computed. Patient's similarities are the edges of Pathway-Specific Patient Similarity Networks (psPSNs). (e) A psPSN is signature if intra-class similarities are stronger that intra-class similarities of the other class and inter-class similarities. (f) Unknown patients are classified based on their similarity to other patients and on how well they resemble class representatives.

information (e.g., gene expression) and on its associations with all the other molecules in the network. At first, the values of the patient's genetic features are mapped to their corresponding nodes in the provided interaction network, then Simpati propagates their values through the interactions. Each node, including the ones without a value, gets a score, which reflects its starting information and the amount given and received from its neighbors. Simpati propagates using the random walk with restart algorithm on the row-normalized network [12]. Propagating a patient's profile starting from the genetic single-value features allows us to obtain a genome-wide profile. This is relevant because the profile can be compared across patients and gives a genome-wide overview of all the genes. Moreover, this is particularly beneficial when we deal with sparse data (e.g., somatic mutation data) where fewer features are identified from the analysis [13].

C. Pathway specific patient similarity

Simpati computes a pairwise similarity between patients for each set of genetic features falling into a specific pathway. In this way, Simpati creates a database of psPSNs reflecting the similarity of patients in each pathway. The nodes of a psPSN are all the patients with known class and the edges are weighted to reflect the pairwise similarity of patients in the features belonging to the pathway.

The approach of measuring similarity on a pathway-level, not only allows to reduce the dimensionality of the features to be compared across patients, but it also creates a feature space, which is more robust to noise compared to single features, while still retaining predictive accuracy [14].

Pathway-specific patient similarity is computed as a linear combination score of three factors. The first one (1) is the Weighted Jaccard and determines how similar the propagated values between two profiles are; the second factor (2) determines how high or low the propagated values are, while the third factor is the opposite of their difference (3). The similarity increases as the two patients have similar values and at the same time high values for the same single-level feature. This is reflected in the final similarity measure, called Trending Matching (4):

$$WJ_p(P_a,P_b) = \frac{\sum_g \min(m_{g,a}, m_{g,b})}{\sum_g \max(m_{g,a}, m_{g,b})} \tag{1}$$

$$MG_p(P_a,P_b) = \frac{\sum_g (m_{g,a}+ m_{g,b})/2}{|p|} \tag{2}$$

$$DIFF_p(P_a,P_b) = 1-|WJ_p(P_a,P_b)-MG_p(P_a,P_b)| \tag{3}$$

$$TM_p(P_a,P_b) = WJ_p(P_a,P_b)+MG_p(P_a,P_b) +DIFF(P_a,P_b) \tag{4}$$

where *p* is a pathway, P_a and P_b are two patients, *g* are all the features | g $\in$ *p*, and *m* is the matrix of features by patients.

D. Feature selection and Best Friend Connector algorithm

Simpati evaluates which pathways are signatures for one of the classes. The members of one class must be more similar (strong intra-similarities of one class) than the members of the opposite class (weak intra-similarities of the other class) and the two classes are not similar (weak inter-similarities). In other words, the topology of the psPSN must reflect the presence of a clique of nodes belonging to the same class being more strongly connected than the rest of the patients. Despite this criterion being genetically intuitive, it is not easy to satisfy due to the complex structure of a patient similarity network where each patient is connected to any other member of the classes in comparison. One patient can easily be more similar to the patients of its opposite class in one specific pathway activity and decrease the separability of the groups. To account for this situation and making the feature selection more robust to outliers at the level of the single pathway, we developed an algorithm called Best Friends Connector algorithm (BFC). The latter is a cohesive subgroup detection algorithm implemented specifically for PSNs to find the strongest community of patients from each class in a network. The algorithm relies on the definition of the concepts of first order best friend (1BF), second order best friend (2BF) and outsiders. Given a root node, its 1BFs are its most similar nodes. 2BFs are the nodes that are not among the root's 1BFs but are 1BFs to one of the root's 1BFs. Outsiders do not belong to any of the previous definitions. The algorithm performs the following operations. It first adjusts the weights of the intraclass connections. Precisely, it increases the similarity of two patients when they both have a weak similarity with outsiders and it decreases it in the opposite case. Then, it iteratively considers one patient as root, it assesses the average of the intraclass connection weights of the subgroup composed by his 1BFs and 2BFs. When each patient has been considered, the algorithm retrieves the set of best friends who got the strongest connections. The cardinality of the 1BFs and 2BFs subgroups, as well as the size of the final subgroup, are customizable.

E. Classification

The signature pathways identified by Simpati are used to classify unknown patients. Each of them is compared to already annotated patients and assigned to the same class of who is most similar to. However, the only strength of similarity could be misleading. The unknown patient could have the strongest similarity with outlier members of the class. Therefore, we designed Simpati to consider also how much the unknown patient represents the class.

The patient to be classified undergoes the same preprocessing described for annotated patients: its profile is propagated in the interaction network and its pairwise TM similarity to each annotated patient is computed, so that the unclassified patient becomes itself a node in each signature psPSN. Then, Simpati associates the profile to one of the classes based on the results of two approaches. For the first, it determines the average similarity of the patient to the members of each class. The patient would be assigned the class to which it has the strongest similarity. For the second approach, Simpati pretends that the patient belongs to one class and measures how far it is from being considered an outlier. The patient would be assigned the class in which it is considered less of an outlier with respect to the other members. More specifically, the patient is simulated to belong to one class and the BFC algorithm is run iteratively. At each run, the algorithm is asked to return a smaller number of strongly connected individuals. The iteration stops when the patient does not belong to the best subgroup. A large number of iterations reflects a strong similarity of the patient to the class representatives. Due to this, the patient would be a candidate to be assigned the class in which it survived the highest number of iterations. Simpati assigns the patient to the class that has been predicted by both the approaches. In case, the results are not concordant, then Simpati does not make the prediction and the pathway together with its PSN are removed from the downstream operations. This step is performed for all signature psPSNs, then the patient's definitive class is the one to which the patient has been most frequently assigned.

The classification performance are evaluated with a leave one out cross validation approach, such that iteratively one patient is considered unknown and composes the testing set, while the others are known and are used as training to determine which pathways are signature. The performance on the testing set are computed using area under the receiver operator characteristic curve (auROC) and area under precision recall curve (auPR) metrics.

F. Downstream Analysis

The signature pathways that are used to classify at least one patient are reported in the final output of Simpati and information about which class they were identified to be signature for. To further pinpoint the most relevant pathways and confirm their signature role for a class, an empirical probability value is computed. On each signature psPSN it is tested whether by randomly shuffling the patients between the two classes, the pathway is still predictive of the original signature class.

To improve the interpretability of the results some other information is computed. First, it has been established that signature pathways reflect strong similarity between members of one class. However, Simpati also reports whether the members are similar in having high values (e.g., high gene expression), reported as up-involved signature pathway, or low values (e.g., low gene expression), reported as down-involved signature pathway. Additionally, based on the BFC results, it is reported how many times a patient has been considered an outlier for its class.

When the features of the profiles provided as input to Simpati are genes and the classification aims to determine association to a disease, it is possible to validate the biological relevance of the identified pathways within Simpati. Queries to the gene-disease associations database (DisGeNet) [15] and to the Human Protein Atlas [16] allows detecting whether the features returned are already known to be associated with the disease being tested.

To obtain a graphical representation of the psPSNs of interest, Simpati offers a graphical interface, which allows to obtain a compact representation of the networks. Patients are grouped based on their similarity so that, instead of plotting all nodes, only some representatives are depicted, making the interpretation of the figure much more feasible.

G. Data preparation for testing

Simpati performances were tested by classifying patients from five cancer types, extracted from The Cancer Genome Atlas (TCGA) using the R packages curatedTCGAData (v1.1.38) [17] and TCGAutils [18]. Two types of biological omics were tested for each cancer type, gene expression from RNAseq data and somatic mutations. The classes assigned to the patients were based on disease stage progression binarized into Early (stage I and II) or Late (stage III and IV). Data preparation for the RNAseq followed the workflow defined by Law et al. [19], while somatic mutation data have been converted into a binary matrix, where a value equal to one was indicating a mutated gene in a patient and zero otherwise. Finally, the six datasets were composed of the following number of samples: 14 Liver hepatocellular carcinoma (LIHC) (7 Early, 7 Late), 21 Stomach adenocarcinoma (STAD) (8 Early, 13 Late), 37 Kidney renal clear cell carcinoma (KIRC) (24 Early, 13 Late), 45 Bladder Urothelial Carcinoma (BLCA) (8 Early, 37 Late), 75 Lung squamous cell carcinoma (LUSC) (60 Early, 13 Late) and 152 Esophageal carcinoma (ESCA) (91 Early, 61 Late) patients.

Pathways were collected from the major databases MSigDB [20] and GO [21] and KEGG [22], while a Biogrid network (v4.2.191) [23] was used to model the biological feature's interactions.

III. RESULTS

Simpati classification results and computational performance were compared to those obtained with netDx (v1.2.0 14-10-2020) for both gene expression and somatic mutations on the prepared TCGA datasets and with PASNet only for gene expression, as this tool does not handle sparse data. Additionally, a biological validation of the pathways retrieved was performed on Simpati and netDx. An online repository is available with a tutorial on how to replicate the results [24].

The classification comparison was performed on the metrics supported by both netDx and PASNet, the auROC and the auPR. These were obtained from a 10-fold cross-validation approach in netDx and a stratified 5-fold cross-validation repeated 10 timed in PASNet, based on the authors' vignette, while for Simpati it was obtained through the leave one out cross validation approach. Figure 2 shows how Simpati performs better than the competitors in both the measures and the biological omics. Simpati also proves

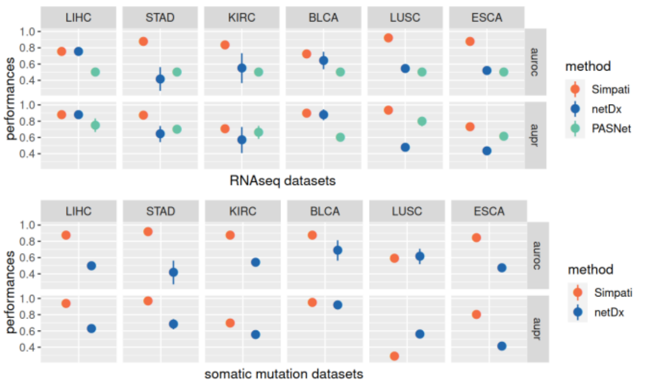

Figure 2. Classification performance comparison between methods. The top box shows performance on the RNAseq datasets, the bottom box on the somatic mutation datasets.

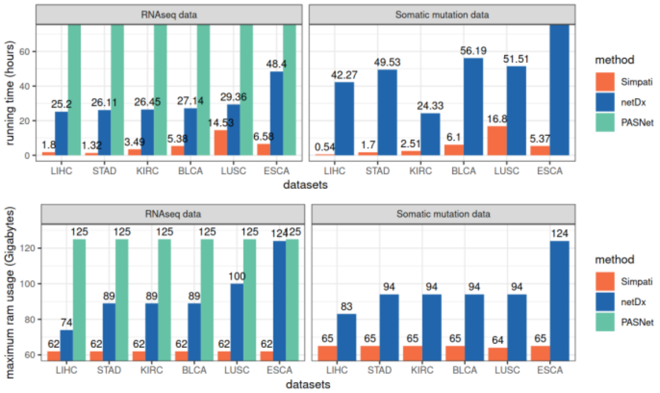

Figure 3. Computational performance comparison between methods. The top box shows running time in hours, the bottom box shows RAM usage in Gigabytes.

to be more reliable in each dataset with a standard error equal to zero due to its leave-one out cross-validation approach.

The patient similarity network paradigm used by Simpati and netDx brings many advantages both in the feature selection, in the classification phase and in the overall interpretability of the software. However, these pros come with a price, which is the software scalability already introduced as a challenge by Pai et al. [5]. A PSN is a complete graph that the methods build with all the patients and for every pathway. This means that an increment in the number of patients and in the number of annotated pathways lead the methods to require more computational resources. netDx and Simpati faced this point with different approaches. netDx is implemented in R and Java, uses the disk to save temporary files and applies a sparsification of the PSNs to decrease the number of edges and so the amount of information associated with them. Simpati is implemented completely in R, natively supports parallel computing and handles all the data of the workflow as sparse matrices or vectors. The RAM usage and the running time required to classify the TCGA datasets were monitored with the same hardware settings for all tools (32-Core Processor, 251 Gigabyte System memory). Simpati compared favorably in the usage of the resources, as reflected in Figure 3. On average across the datasets, Simpati it's ~ 16 times faster than netDx and requires ~ 1.5 times less Gb of RAM. Both netDx and Simpati outperformed PASNet performance.

Both Simpati and netDx provide the most relevant pathways they detect during the workflow. These pathways should help characterize patient's classes and improve the interpretability of the method. For this reason, Simpati integrates into its workflow a biological validation step exploiting DisGeNet and the Human Protein Atlas. For each dataset, a set of key words describing the disease are defined, then the percentage of key words associated with the pathway in DisGeNet at least once are reported. Additionally, Simpati reports the percentage of features in each pathway which are associated with the cancer type in the Human Protein Atlas. In order to compare the biological validity of the methods, these values were computed for netDx and Simpati signature pathways and only the most biologically relevant pathways were kept. Two criteria for retaining relevant pathways were tested: pathways having at least one key word associated in DisGeNet and pathways having more than 90% of features associated with the cancer type in the Human Protein Atlas. The number of pathways satisfying these constraints were compared and results are shown in Figure 4.

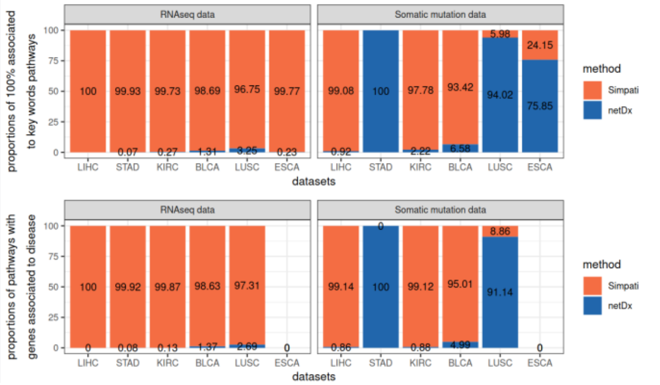

Figure 4. Biological validation comparison. The top box shows the relative proportion of signature pathways associated with relevant dataset key-words between the two methods and the bottom box shows the relative proportion of signature pathways associated with disease-type between the two methods.

This analysis highlights how Simpati is able to select biologically significant pathways directly associated with the patients it classifies and it performs better than the competitor.

IV. Conclusions

Simpati is a pathway-based classifier of patient classes for genetic data. It is the first classifier employing novel ad-hoc algorithms for PSNs to detect pathway-specific similarities. The tool is strongly centered around providing a good interpretability, as it provides signature pathways to unveil the altered biological mechanisms of a disease phenotype. Thanks to a propagation algorithm that considers the interconnected nature of the cell's molecules, Simpati can classify dense, sparse, and nonhomogeneous genetic data. Future work will be focused on the development of strategies for the integration of multiple omics and on improving scalability for larger datasets.

ACKNOWLEDGMENTS

L.G developed the method, implemented and tested the software, wrote the text, and produced the images. C.M. tested the software and prepared the text. R. G. coordinated the project and revised the text.

This work was supported by the European Union's Horizon 2020 research and innovation programme under grant agreement 814978 (TUBE), and by the JPND Personalized Medicine for Neurodegenerative Diseases project 334799 (ADAIR).

REFERENCES

[1] L. Jin et al., "Pathway-based analysis tools for complex disease: a review", Genomics Proteomics Bioinformatics, no. 12, pp.210-220, 2014.

[2] Y. Drier, M. Sheffer, and E. Domany, "Pathway-based personalized analysis of cancer", PNAS, no.110, pp.6388-6393, 2013.

[3] M. P. Segura-Lepe, H. C. Keun, and T. M. D. Ebbels, "Predictive modelling using pathway scores: robustness and significance of pathway collections", BMC Bioinformatics, no.20, pp.543, 2019.

[4] E. Lee, H. Y. Chuang, J. W. Kim, T. Ideker, and D. Lee, "Inferring pathway activity toward precise disease classification", PLoS Comput Biol, 2008.

[5] M. Yousef, E. Ülgen, and O. U. Sezerman, "CogNet: classification of gene expression data based on ranked active-subnetwork-oriented KEGG pathway enrichment analysis". PeerJ Computer Science no7, pp336, 2021.

[6] J. Hao, Y. Kim, T. K. Kim, and M. Kang, "PASNet: pathway-associated sparse deep neural network for prognosis prediction from high-throughput data", BMC Bioinformatics, no.19 pp.510, 2018.

[7] S. Pai, et al., "netDx: interpretable patient classification using integrated patient similarity networks", Mol Syst Biol no.15, 2019.

[8] S. Mostafavi, D. Ray, D. Warde-Farley, C. Grouios, and Q. Morris, "GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function", Genome Biology, vol.9, 2018.

[9] F. Fabris, D. Palmer, J. P. de Magalhães, and A. A. Freitas "Comparing enrichment analysis and machine learning for identifying gene properties that discriminate between gene classes", Brief Bioinform, no.21, pp.803–14, 2020.

[10] Simpati R package. [Online]. Available at: https://github.com/InfOmics/Simpati.

[11] Simpati GUI. [Online]. Available at: https://github.com/LucaGiudice/propaGUIation.

[12] D. H. Le, "Random walk with restart: A powerful network propagation algorithm in Bioinformatics field", 4th NAFOSTED Conference on Information and Computer Science, p. 242–247, 2017.

[13] M. Hofree, J. P. Shen, H. Carter, A. Gross, and T. Ideker, "Network-based stratification of tumor mutations", Nat Methods, no.10, pp.1108–1115, 2013.

[14] M. P. Segura-Lepe, H. C. Keun, and T. M. D. Ebbels, "Predictive modelling using pathway scores: robustness and significance of pathway collections", BMC Bioinformatics no.20, pp.543, 2019.

[15] J. Piñero et al., "DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants", Nucleic Acids Res, no.45, pp.D833–9. , 2017.

[16] M. Uhlen et al., "A pathology atlas of the human cancer transcriptome", Science vol.357, 2017.

[17] Multiomic Integration of Public Oncology Databases in Bioconductor - PubMed n.d. https://pubmed.ncbi.nlm.nih.gov/33119407/ [retrieved May, 2021].

[18] M. Ramos, L. Schiffer, S. Davis, and L. Waldron. "TCGAutils: TCGA utility functions for data management", R package version 1.10.1, 2021.

[19] C. W. Law et al., "RNA-seq analysis is easy as 1-2-3 with limma, Glimma and edgeR", F1000Res vol.5, 2015

[20] A. Subramanian et al., "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles", Proc Natl Acad Sci USA, no.102, pp.15545–50, 2005.

[21] M. Ashburner et al., "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium." Nat Genet, no.25, pp.25–29, 2000.

[22] M. Kanehisa M., and S. Goto. "KEGG: kyoto encyclopedia of genes and genomes", Nucleic Acids Res no.28, pp.27–30, 2000.

[23] R. Oughtred et al., "The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions", Protein Sci, no.30, pp.187–200, 2021.

[24] Simpati's Supplementary for results replication. [Online]. Available at: https://github.com/LucaGiudice/supplementary-Simpati.