PriPath: Identifying Dysregulated Pathways from Differential Gene Expression via Grouping, Scoring and Modeling with an Embedded Machine Learning Approach

Malik Yousef (malik.yousef@gmail.com)
Zefat Academic College https://orcid.org/0000-0001-8780-6303
Fatma Ozdemir
Abdullah Gul Universitesi Fen Bilimleri Enstitusu
Amhar Jaaber
Abdullah Gul Universitesi Fen Bilimleri Enstitusu
Jens Allmer
Ruhr West University of Applied Sciences: Hochschule Ruhr West
Burcu Bakir-Gungor
Abdullah Gul Universitesi Fen Bilimleri Enstitusu

Research Article

Keywords: Feature Selection, Feature Scoring, Features Grouping, Clustering, Biological Knowledge Integration, KEGG pathway, Classification, Gene Expression, Enrichment Analysis, Machine learning, Bioinformatics, Data science, Data mining, Genomics

Posted Date: April 19th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1449467/v1

Version of Record: A version of this preprint was published at BMC Bioinformatics on February 23rd, 2023. See the published version at https://doi.org/10.1186/s12859-023-05187-2.

Abstract

Background: Cell homeostasis relies on the concerted actions of several genes; and dysregulated genes lead to disease manifestations. In living organisms, genes or their products do not act alone, but instead act within a large network. Subsets of these networks can be viewed as modules which provide certain functionality in an organism. Kyoto Encyclopedia of Genes and Genomes (KEGG) systematically analyzes gene functions, proteins, molecules, and provides a PATHWAY database. Measurements of gene expression (e.g., RNA-seq data) can be mapped into KEGG pathways in order to determine which modules are affected or dysregulated in a disease. However, genes acting in multiple pathways, and some other inherent issues complicate such analyses. To detect dysregulated pathways, current approaches may only employ gene expression data and neglect some of the existing knowledge stored in KEGG pathways. For a more holistic association between gene expression and pathways, new approaches which take into account more of the compiled information are required.

Results: PriPath is a novel approach that transfers the generic approach of grouping, scoring followed by modeling for the analysis of gene expression with KEGG pathways. In PriPath, we utilize the KEGG pathway as the grouping information and insert this information into a machine learning algorithm for selecting the most significant KEGG pathways. Those groups are utilized to train a machine learning model for the classification task. We have tested PriPath on 13 gene expression datasets of various cancers and other diseases. Our proposed approach successfully assigned biologically and clinically relevant KEGG terms to the differentially expressed genes. We have comparatively evaluated the performance of PriPath against other tools, which are similar in their merit. For each dataset, we manually confirmed the top results of PriPath in literature, and we compared PriPath predictions to the predictions of Reactome and DAVID.

Conclusions: PriPath can thus aid in determining dysregulated pathways, which is applicable to medical diagnostics. In the future, we aim to advance this approach in such a way that it will be possible to perform patient stratification based on gene expression and to identify druggable targets. Thereby, we cover two aspects of precision medicine.

1. Background

Nowadays the healthcare system is facing a shift towards precision medicine. While the diseases are evaluated at the molecular level, the patient stratification becomes possible so that the most suitable medication can be identified for each patient. This approach heavily depends on the molecular level data which are obtained through novel experimental high-throughput methods. For example, next generation sequencing technologies are utilized for whole-genome sequencing, analysis of genome diversity, the discovery of non-coding RNAs and protein-binding sites, metagenomics, epigenomics, and gene-expression profiling [1], [2]. Along this line, the transcriptomic data which adhere to the FAIR (Findable, Accessible, Interoperable, and Reusable) principles is generated at an unpredictable pace. Therefore, large gene expression data sets became publicly available for a diverse set of diseases.

The current bottleneck is in the biomedical data analysis, starting from the preprocessing of the sequencing data, rising up to supporting decision making processes e.g., drug selection. The high dimensionality of the data (large number of genes) combined with a small number of samples makes it difficult to interpret the data. In this respect, the utilization of feature selection is essential for dimensionality reduction and for the selection of the most informative genes. Additionally, more efficient gene selection methods are proposed to achieve the full potential of the growing data, to develop gene-based diagnostic tests and to aid drug discovery. Integrative gene selection incorporates domain knowledge from external biological knowledge assistants to improve the gene selection approaches [3]. For example, Gene Ontology (GO) is one of the resources that is used to integrate biological background into statistical analysis/machine learning (ML) analyses of gene expression.

In order to perform their particular biological functions, genes do not act alone, but instead genes act as a group within several metabolic and signaling pathways. This information can be exploited for feature selection in gene expression data

analysis. In other words, the groupings of the genes in terms of pathways can be incorporated into the feature selection problem to identify gene signatures. It is shown in literature that for gene expression data analysis, the methods which incorporate pathway knowledge usually outperforms their gene-based counterparts where biological domain knowledge or pathway knowledge is not considered [4]. While the traditional methods rely on the identification of statistically significant genes having differential expression between two different phenotypes, pathway knowledge based methods impose further constraints for the prediction task and force training methods to choose more scientifically meaningful genes. Among different pathway databases, KEGG PATHWAY is a commonly utilized external ontology resource [5]. KEGG PATHWAY was recently enriched with the addition of new pathways, the cellular process, and diseases [6], improving its popularity.

The integration of pathway knowledge can vary among different methods. While some of the algorithms treat the pathway as a graph, consider the underlying topology of the pathways, and analyze the connections of genes within this graph; some other methods consider the pathways as gene sets. A comprehensive review of topology-based (TB) vs. non-topology-based (non-TB) pathway analysis methods can be found in [6]. Comparative evaluation of topology-based pathway enrichment analysis methods can also be found in [7], [8]. Non topology based pathway guided gene selection methods treat every gene inside a specific pathway equally and assign equal weights. On the other hand, topology based pathway guided gene selection methods compute the connectivity level of the genes inside a pathway and use this information while weighting the genes. Hence, in topology based pathway guided gene selection methods, the genes having high connectivity in a pathway may be prioritized. There are also functional score based gene selection methods such as [3, 9], in which only the memberships of the genes into pathways are considered to generate an evaluation score. These methods implicitly assume that all genes belonging to a specific pathway co-regulate and co-function. On the other hand, a more structured pathway knowledge is taken into account in topology-based methods, such as in [10]. In terms of predictive accuracy, some studies such as [11, 12], have concluded that pathway-guided gene selection methods do not outperform classic gene-based feature selection methods. This inferiority may be explained by the fact that the pathway knowledge retrieved from those canonical pathway databases/knowledge-bases such as the KEGG [13], Gene Ontology (GO) [14] and Reactome [15] conveys no information or limited meaningful information for a specific dataset or condition/disease. In contrast, the pathways constructed in a “data-driven” way may be more informative for the diseases under investigation and thus preferred over the canonical pathways.

Yousef and others recently proposed grouping, scoring and modeling (G-S-M) based machine learning ideas for integrating biological domain knowledge into gene expression data analysis [16]. They proposed several tools that adopt this approach. For instance, maTE [17] integrates biological knowledge of microRNAs (miRNA) for grouping the genes. cogNet [18] performs KEGG pathway enrichment analysis based on ranked active-subnetworks. mirCorrNet [19] detects groups of miRNA-mRNAs via analyzing the correlation in miRNA and mRNA expression profiles obtained from the same sample. Similarly, miRModuleNet [20] detects miRNA-mRNA regulatory modules to serve as groups while integratively analyzing two -omics datasets. Another G-S-M model based study of Yousef et al [21] utilizes Gene Ontology for grouping the genes. The first study of considering groups or clusters of genes rather than individual genes was also developed by Yousef et al [20, 21]. The above-mentioned tools are different implementations of this idea for different data types. SVM-RCE (Support Vector Machines -Recursive Cluster Elimination) is an example of grouping genes based on their gene expression values and it scores each cluster of genes via incorporating a machine learning algorithm. This approach has received attention from other researchers [24]. Similarly, SVM-RNE [25] is based on gene network detection to serve as groups for scoring by the G-S-M model. SVM-RCE-R is one other example developed along this line. However, there is still room for developing more tools that are based on the G-S-M model and incorporate biological knowledge such as KEGG pathways.

In this paper, we have introduced a novel tool, named PriPath, that is based on the ranking and grouping of biological information based on the G-S-M model. In this study the KEGG pathway is treated as the set of the genes, neglecting the struch of the pathway. PriPath uses a set of KEGG pathways as groups to perform scoring and the classification. PriPath produces performance metrics and a list of dysregulated KEGG pathways. The innovation of our approach is based on its ability to search the space of groups of the KEGG pathways in order to rank and find the most important groups.

We have tested PriPath on 13 gene expression datasets of various cancers and other diseases. The results indicate that we outperform maTE in most cases, and PriPath uses less number of genes than SVM-RCE-R and CogNet. Additionally, for each dataset, we compared PriPath predictions to the predictions of Reactome and DAVID and manually verified the top outcomes in the literature. PriPath could detect biologically and clinically relevant pathways which are affected for the disease under study. PriPath can assist the identification of dysregulated pathways, which is applicable in medical diagnostics. Hence, we tackle an aspect of precision medicine.

The rest of the manuscript is organized as follows. Sections 2 and 3 describe the materials and methods used. In section 4, we evaluate the results by comparing the proposed approach with other tools that employ embedded feature selection and with traditional functional enrichment tools. In section 5, we discuss the top results by providing evidence from the literature. In section 6, we conclude our paper.

2. Materials

2.1.Gene Expression Data

To support algorithm development and testing, 13 human gene expression datasets (shown in Table 1) were downloaded from the Gene Expression Omnibus (GEO) [26] at NCBI. For all datasets, disease (positive) and control (negative) data were acquired. These 13 datasets served to test PriPath and for comparison with other tools. Moreover, we have included these 13 reference datasets since they have been previously utilized for the comparison of gene selection tools.

Table 1
Description of the 13 gene expression datasets used in this study.

GEO Accession	Title	Disease	# of Samples	# of Samples in Classes
GDS1962	Glioma-derived stem cell factor effect on angiogenesis in the brain	Glioma	180	negative = 23 positive = 157
GDS2547	Metastatic prostate cancer (HG-U95C)	Prostate cancer	164	negative = 75 positive = 89
GDS4824	Prostate cancer Analysis of malignant and benign prostate tissues.	Prostate cancer	21	negative = 8 positive = 13
GDS3268	Colon epithelial biopsies of ulcerative colitis patients	Colitis	202	negative = 73 positive = 129
GDS3646	Celiac disease: primary leukocytes	Celiac disease	132	negative = 22 positive = 110
GDS3874	Diabetic children: peripheral blood mononuclear cells (U133A)	Diabetes	117	negative = 24 positive = 93
GDS3875	Diabetic children: peripheral blood mononuclear cells (U133B)	Diabetes	117	negative = 24 positive = 93
GDS5037	Severe asthma: bronchial epithelial cell	Asthma	108	negative = 20 positive = 88
GDS5499	Pulmonary hypertensions: PBMCs	Pulmonary hypertension	140	negative = 41 positive = 99
GDS3837	Non-small cell lung carcinoma in female nonsmokers	Lung cancer	120	negative = 60 positive = 60
GDS4516_4718	Colorectal cancer: laser microdissected tumor tissues Colorectal cancer:humogenized tumor tissues	Colorectal cancer	148	negative = 44 positive = 104
GDS2609	Early onset colorectal cancer: normal-appearing colonic mucosa	Colorectal cancer	22	negative = 10 positive = 12
GDS3794	Rheumatoid arthritis: peripheral blood mononuclear cells	Arthritis	33	negative = 15 positive = 18

2.2. KEGG Data

We have downloaded the KEGG data from Bioconductor using the R programming language [27] at 21.01.2021. The KEGG data contains 32083 entries that represent 331 KEGG pathways [28].

3. Methods

3.1. Algorithm

PriPath algorithm considers KEGG pathways and their target gene expression for two conditions which are control (negative) and disease (positive). Each condition is described by its gene expression values. The aim of the algorithm is to find which KEGG pathways may be dysregulated for the disease under study. Therefore, machine learning is applied to determine which KEGG pathways are associated with gene expression, similar to our previous works [18].

The main flowchart of PriPath is presented in Fig. 1. The main step of the PriPath tool is the grouping phase. The expression of each gene presents a feature in the gene expression dataset. Features are grouped by KEGG pathways. A two-class gene expression dataset consists of a number of covariate samples and genes. The classes could be disease or any experimental condition versus a control. Each KEGG pathway term is composed of a set of genes and the grouping of genes is based on the KEGG pathways.

By iterating over all groups, the KEGG pathways are ranked according to their capacity to classify the two classes based on the test results following training of a random forest (RF) classifier using training and testing data divided in 80:20 ratio (shown in yellow in Fig. 1).

Firstly, the grouping function obtains the associated gene expression data rows from the gene expression dataset, and then RF is used. The ranking function returns the average accuracy by using subsets of the gene expression dataset, where the data is split into training and testing, and repeated iteratively.

The ranking step for each KEGG pathway as the grouping factor, the best j KEGG pathways, are selected and their groups (i.e. their targets) are combined (Fig. 1, yellow part). An RF model is trained with the grouping function given by the best j pathways. Finally, the model is tested, and the performance results are saved.

3.1.1. Classification Approach

We utilized the RF classifier included in the KNIME platform [29]. 80% training data was used to train the classifier and 20% testing data was used to test the classifier. Since the data sets are imbalanced, it can affect the performance of the classifier. We employed under-sampling to address the imbalance between classes. The under-sampling method decreases the size of the abundant class by randomly selecting an equal number of samples with the rare class from the abundant class. This method is performed along with each round of cross-validation. We apply 10-fold Monte Carlo cross validation (MCCV) [30] for model training. The default parameters were used for RF training. The number of levels (tree depth) were not limited, and the number of models was set to 100.

3.1.2. Model Performance Evaluation

In order to compare the performance of the tools, a number of statistical measures such as sensitivity, specificity, and accuracy were calculated [31]. The following formulations were utilized where TP, true positive; FP, false positive; TN, true negative; and FN, false negative:

Sensitivity (SE, Recall) = TP/(TP + FN)

Specificity (SP) = TN/(TN + FP)

Accuracy (ACC) = (TP + TN)/(TP + TN + FP + FN

In addition, the area under the receiver operator characteristic (ROC) curve (AUC) calculates the probability that a classifier will rank a randomly selected positive sample higher than a randomly selected negative sample [32]. The performance

measures are calculated as the average of 10-fold MCCV.

3.2. Implementation

We utilized the Konstanz Information Miner (KNIME) to implement our approach [29]. KNIME workflows include processing nodes and data links (lines/ edges). In the workflow, edges provide data transport from one node to another. For the implementation part of our algorithm, we have decided to utilize the free and open-source platform KNIME since it is simple to use and provides user-friendly graphical representations. Other tools can also be integrated into KNIME easily. For example, an R node can be integrated into the workflow of KNIME. A meta-node is designed as a combination of nodes with a specific task. The workflow of Multi-File PriPath is shown in Fig. 2. The “List Files node” is the first node that uploads a list of the names of the dataset. Next, the “Table Reader” node reads each data and sends it to the PriPath algorithm (Meta-node). The task of the node “Loop End” is to collect all the results.

3.3. Quality Assessment of the Predictions

3.3.1. Closeness to Majority Prediction

For none of the 13 datasets, it is not experimentally verified that which KEGG pathways are dysregulated or not. In the absence of a ground truth, it is hard to assess the qualities of the predictions or to compare the performance among different approaches. Still, it is important to analyze the effectiveness of different methods. To achieve such a comparison, we compiled the predictions generated during the robustness analysis. In this analysis, for each KEGG pathway, we checked whether it has been associated with the disease label of the sample in literature. If it was associated with the disease, the prediction constitutes a true positive result. Otherwise, the prediction is counted as a false positive one.

Additionally, we created a consensus prediction for each dataset using all tools considered here. For any predictor, getting as close to the consensus prediction as possible, becomes a quality measure.

3.3.2. Competing Approaches

In literature, there are other approaches that have been developed with the aim of functionally enriching a set of differentially expressed genes [10]. From these tools, we selected a few that are widely used in literature. It is beyond the scope of this work to compare all existing approaches, but for the quality assessment of the predictions, we have selected two popular functional enrichment tools. Thereby, we aim to avoid introducing a bias in the consensus prediction, and we aim to represent different approaches equally. We have included DAVID (6.8) and Reactome (version 78 - Pathway Browser version 3.7. To compare these tools, 1000 genes were selected with a t-test. Then, selected 1000 genes were submitted to DAVID and Reactome. The top 10 predicted KEGG pathways of these tools were recorded. Finally, we assessed the number of shared pathways among the predictions of these tools.

3.3.2.1. DAVID

Among many other functions of DAVID (https://david.ncifcrf.gov/tools.jsp), it mainly offers gene set enrichment analysis. The tool accepts a set of genes as an input and does not consider quantitative information of each gene. Another notable functionality of DAVID is its ability to translate among identifiers, for example, between Unigene IDs and GenBank IDs. We have provided a filtered gene set to DAVID, which is equal to the combined training and testing data used during the development of the PriPath. DAVID provides interpretation of genome-scale datasets and the transition from data collection to biological meaning [33]. In this study, we applied a t-test to each gene expression data. We used the top 1000 genes which come from the results of the t-test as inputs. Default values are used for DAVID.

3.3.2.2. *Reactome*

Reactome (https://reactome.org/) provides bioinformatics tools for the interpretation, visualization, and analysis of pathway knowledge to assist basic research, modeling, genome analysis, and systems biology. Pathway analysis methods have a

wide field of applications in physiological and biomedical research. The constantly increasing size of the data samples is one of the main problems, from the analysis methods performance point of view [34]. In this study, a t-test was applied to each gene expression data. As an input to Reactome, we have used the top 1000 genes which are obtained as a result of the t-test. The settings of Reactome were used without any change. By using DAVID, the affected KEGG pathways were identified via utilizing the targeted genes within the identified Reactome pathways.

4. Results

4.1. Performance Evaluation of PriPath

13 different gene expression data sets were used to test PriPath. Figure 3A and Supplementary Table 1 presents the AUC values for different numbers of clusters for each dataset; and Fig. 3B displays the average number of genes used for each level for each dataset. The average number of genes over 100 iterations that were performed as MCCV is shown in columns #G.

4.2. Comparative Evaluation with other tools that employ embedded feature selection

We compare PriPath with other tools which employ embedded feature selection. To our knowledge, CogNet, maTE, and SVM-RCE are the only examples in this area. We have recorded the AUC values over the top 1–10 groups that were ranked by the scoring stage for each tool, except for SVM-RCE. For SVM-RCE, we have measured the performance starting by 1000 genes and 100 clusters, and decreased 10% at each iteration. We have used the last 10 clusters of SVM-RCE during comparison.

In Fig. 3A, we present the mean AUC values of four tools over the 13 datasets on the 10 clusters/groups; and in Fig. 3B we plot the mean number of genes. As illustrated in Fig. 3A, PriPath performs similarly with maTE, CogNet, and SVM-RCE for different datasets. As illustrated in Fig. 3B, it uses fewer genes than SVM-RCE. Figure 3 implies that on average PriPath outperforms maTE by 1.2% while producing similar results with SVM-RCE. In terms of the mean number of genes, SVM-RCE uses a 16 fold bigger number of genes than PriPath.

4.3. Comparative Evaluation with traditional functional enrichment tools

Additionally, we compared the performance of PriPath with traditional functional enrichment tools such as DAVID, which apply statistical analysis on gene expression datasets and identify overrepresented pathways. Figure 4 summarizes the comparative evaluation of PriPath with traditional functional enrichment tools for the GDS1962 dataset using an UpSetR plot. Results for the other datasets are available in the Supplementary Figs. 1–12. An UpSetR plot comprises two axes and a connected-dot matrix. The vertical rectangles illustrate the number of elements attending in each list combination. The connected-dots matrix shows which combination of lists corresponds to which vertical rectangle. The horizontal bars that correspond to the size of sets indicate the participation of hovered objects (from the vertical rectangles) in the respective lists [35]. As illustrated in Fig. 4, for the GDS1962 dataset, Reactome and DAVID commonly identify seven KEGG pathways in their prediction and they commonly identify one pathway with PriPath. PriPath shares another predicted pathway with Reactome and features eight unique pathways which are not predicted by the other two approaches. When the predictions are analyzed for 13 different datasets, it can be observed from Supplementary Figs. 1–13 that PriPath shares 1–3 predictions with DAVID; 1–4 predictions with Reactome; and features unique pathways not predicted by the other two approaches. In general, Reactome and DAVID share 1–6 pathways. On average PriPath shares 1–2 pathways with either one of these tools. In some cases, PriPath does not share any predictions with neither DAVID nor Reactome.

5. Discussions

In the previous section, we have presented the results of our experiments using PriPath on 13 different datasets; and our comparative performance evaluation with 2 competitor tools. In this section, firstly, we will discuss the biological relevance of our findings. Secondly, the performance of PriPath against its competitors will also be discussed in terms of the robustness analysis and the quality assessment (as explained in detail in the methods section).

Table 2 summarizes the association of the top 3 identified pathways of PriPath with the disease under study. Table 2 also presents whether these top 3 identified pathways were included in the top 3 predictions of DAVID and Reactome. These possible associations can be summarized as following.

Table 2
Association of the top 3 identified pathways of PriPath with the disease under study.

Dataset	Disease under investigation	KEGG Pathway ID	Found in the top 3 predictions of DAVID	Found in the top 3 predictions of Reactome	Pathway Name	Literature support for top 3 identified pathways
GDS1962	Glioma	HSA05165	-	-	Human papillomavirus (HPV) infection	The presence of HCMV and HPV is shown in gliomas. According to one study, HPV infection did not have a significant effect on the prognosis of glioma patients, while another study supports the presence of HPV in gliomas.
		HSA04550	-	-	Signaling pathways regulating pluripotency of stem cells	None
		HSA05131	-	-	Shigellosis	It could occur in cancer patients undergoing chemotherapy.
GDS2547	Prostate cancer	HSA04910	-	-	Insulin signaling pathway	Insulin resistance index is positively correlated with prostate volume in benign prostatic hyperplasia complicated with diabetes patients.
		HSA03010	+	-	Ribosome	Ribosome-targeting drugs may be effective against diverse prostate cancer.
		HSA05171	-	-	Coronavirus disease	Some research results came out pointing to a possible hidden liaison between prostate cancer (PCa) and COVID-19.
GDS2609	Colorectal cancer	HSA04010	-	-	MAPK signaling pathway	Activation signaling pathways including the MAPK pathway enhance Colorectal cancer progression.

Dataset	Disease under investigation	KEGG Pathway ID	Found in the top 3 predictions of DAVID	Found in the top 3 predictions of Reactome	Pathway Name	Literature support for top 3 identified pathways
		HSA04657	-	-	IL-17 signaling pathway	IL-17A inhibitors should be assessed for their therapeutic and preventative potential in human cancers, particularly in colorectal cancer.
		HSA05130	-	-	Pathogenic Escherichia coli infection	Pathogenic E. coli could be a factor in developing colorectal cancer.
GDS3268	Colitis	HSA04151	-	+	PI3K-Akt signaling pathway	Up-regulating the PI3K/Akt-mTOR signaling pathway can trigger cell apoptosis and inflammation in ulcerative colitis.
		HSA05200	-	+	Pathways in cancer	It has long been known that long duration of Ulcerative Colitis is a risk factor for the development of Colitis Associated Cancer.
		HSA05164	-	-	Influenza A	Infection with influenza A could cause hemorrhagic colitis.
GDS3646	Celiac disease	HSA05010	-	-	Alzheimer disease	Several types of dementia such as Alzheimer dementia, vascular dementia, frontotemporal dementia were reported in association with Celiac disease.
		HSA04020	-	-	Calcium signaling pathway	None
		HSA05012	-	-	Parkinson disease	PARK7 plays an important role in the preservation of mucosal integrity in Coeliac disease.

Dataset	Disease under investigation	KEGG Pathway ID	Found in the top 3 predictions of DAVID	Found in the top 3 predictions of Reactome	Pathway Name	Literature support for top 3 identified pathways
GDS3794	Arthritis	HSA04620	-	-	Toll-like receptor signaling pathway	Rheumatoid arthritis (RA) development can be induced by the activation of the Toll-like receptor (TLR) signaling pathway.
		HSA04657	-	-	IL-17 signaling pathway	The IL-17 cytokines have a crucial role in the chronic inflammation of the synovium in Psoriatic arthritis.
		HSA05022	-	-	Pathways of neurodegeneration - multiple diseases	Neurodegenerative disease increases the progress of arthritis.
GDS3837	Lung cancer	HSA04974	+	-	Protein digestion and absorption	None
		HSA04510	-	-	Focal adhesion	FAK is significant in small cell lung cancer biology and targeting its kinase domain may have therapeutic potential.
		HSA04151	+	+	PI3K-Akt signaling pathway	FGF21 may function as a tumor promotor by activating the SIRT1/PI3K/AKT signaling pathway in lung cancer.
GDS3874	Diabetes	HSA05203	-	-	Viral carcinogenesis	None
		HSA04625	-	-	C-type lectin receptor signaling pathway	None
		HSA05166	-	-	Human T-cell leukemia virus 1 infection	None
GDS3875	Diabetes	HSA05168	-	-	Herpes simplex virus 1(HSV-1) infection	HSV-1 infection has an important association with type 2 diabetes.

Dataset	Disease under investigation	KEGG Pathway ID	Found in the top 3 predictions of DAVID	Found in the top 3 predictions of Reactome	Pathway Name	Literature support for top 3 identified pathways
		HSA04910	-	-	Insulin signaling pathway	This disease, also known as insulin resistance, is generated by the disruption of the insulin signaling pathway.
		HSA05022	-	-	Pathways of neurodegeneration - multiple diseases	One of the conditions which result in neurodegeneration is diabetes.
GDS4516_4718	Colorectal cancer (CRC)	HSA04080	-	-	Neuroactive ligand-receptor interaction	None
		HSA04721	-	-	Synaptic vesicle cycle	None
		HSA04724	-	-	Glutamatergic synapse	Neuroligin1 that is the main component of excitatory glutamatergic synapses complex is verified as a new poor prognostic marker for CRC.
GDS4824	Prostate cancer	HSA04080	-	-	Neuroactive ligand-receptor interaction	None
		HSA05163	-	-	Human cytomegalovirus infection	The activation of the Human cytomegalovirus (HCMV) major immediate early promoter by androgen in the prostate might contribute to oncomodulation in prostate cancers.
		HSA04062	-	-	Chemokine signaling pathway	Chemokines play modulatory roles in prostate cancer metastasis.
GDS5037	Asthma	HSA04530	-	-	Tight junction (TJ)	Asthma may be linked to differential expression of TJ.
		HSA05016	-	-	Huntington disease	None

Dataset	Disease under investigation	KEGG Pathway ID	Found in the top 3 predictions of DAVID	Found in the top 3 predictions of Reactome	Pathway Name	Literature support for top 3 identified pathways
		HSA05022	-	-	Pathways of neurodegeneration - multiple diseases	Asthma, especially when severe, is associated with features of neuroinflammation and neurodegeneration.
GDS5499	Pulmonary Hypertension (PH)	HSA04010	-	-	MAPK signaling pathway	Inhibition of the MAPK axis could prevent vascular remodeling in Pulmonary artery hypertension.
		HSA04621	-	-	NOD-like receptor signaling pathway	NOD-like receptor subfamily C3 may potentially be a diagnosis index and symbolize a prognostic factor for PH patients.
		HSA04390	-	-	Hippo signaling pathway	PAH is ameliorated by suppressing HIPPO signaling pathway.

For the GDS1962 dataset, a study concerning glioma, PriPath top predictions are Human papillomavirus infection (HSA05165), Signaling pathways regulating pluripotency of stem cell (HSA04550), and Shigellosis (HSA05131). The association of viruses and cancer have often been shown and the correlation of HPV and glioma is not an exception [36]. Shigellosis is also an opportunistic infection of the immune-compromised, so it could be seen for cancer patients undergoing chemotherapy. The other tested tools were not able to find these pathways in their top 3 predictions.

For the GDS2547 dataset, a study concerning prostate cancer, PriPath top predictions are insulin signaling pathway (HSA04910), Ribosome (HSA03010), and Coronavirus disease (HSA05171). The connection between diabetes and prostate cancer has been demonstrated in [37]. In the literature, it has been shown that Ribosome-targeting drugs may be effective against diverse prostate cancer [38]. Some research findings point to a possible hidden liaison between prostate cancer (PCa) and COVID-19 [39]. Among the other tested tools, DAVID was able to find associations between the data and only the Ribosome pathway in its top 3 predictions.

For the GDS2609 dataset, a study concerning Colorectal cancer, PriPath top predictions are MAPK signaling pathway (HSA04010), IL-17 signaling pathway (HSA04657), and Pathogenic Escherichia coli infection (HSA05130). The relationship between the MAPK signaling pathway and Colorectal cancer has been illustrated in [40]. The research shows that IL-17A inhibitors have their preventive potential in human cancers, particularly in colorectal cancer [41]. Pathogenic E. coli could also be a factor in developing colorectal cancer [42]. The other tested tools were not able to find these associations between the data and the investigated disease in their top 3 predictions.

For the GDS3268 dataset, a study concerning Colitis, PriPath top predictions are PI3K-Akt signaling pathway (HSA04151), Pathways in cancer (HSA05200), and Influenza A (HSA05164). The association of the PI3K-Akt signaling pathway and Colitis has been shown in [43]. The long duration of Ulcerative Colitis is known as a risk factor for the development of Colitis Associated Cancer [44]. The same study reported that the analysis of the transcriptomic changes in the colonic mucosa of

the long-duration Ulcerative Colitis patients revealed colitis-associated cancer pathways. Infection with influenza A could cause hemorrhagic colitis [45]. Reactome was also able to find the associations between the data and PI3K-Akt signaling pathway and Pathways in cancer in their top 3 predictions.

For the GDS3646 dataset, a study concerning Celiac disease (CD), PriPath top predictions are Alzheimer disease (HSA05010), Calcium signaling pathway (HSA04020), and Parkinson disease (HSA05012). Several types of dementia such as Alzheimer dementia, vascular dementia, frontotemporal dementia were reported in association with CD [46]. Lurie et. al. also reported Alzheimer's type memory impairment in two patients that were both diagnosed with CD after 60 years [47]. The association of Parkinson disease and Celiac has been shown in [48]. The other tested tools were not able to find these pathways in their top 3 predictions for this dataset.

For the GDS3794 dataset, a study concerning Arthritis, PriPath top predictions are Toll-like receptor signaling pathway (HSA04620), IL-17 signaling pathway (HSA04657), and Pathways of neurodegeneration - multiple diseases (HSA05022). Rheumatoid arthritis (RA) is associated with the Toll-like receptor (TLR) signaling pathway [49]. The IL-17 cytokines have an important role in the chronic inflammation of the synovium in Psoriatic arthritis [50]. The research shows that neurodegenerative disease increases the progress of arthritis [51]. The other tested tools were not able to find these pathways in their top 3 predictions for this dataset.

For the GDS3837 dataset, a study concerning Lung cancer, PriPath top predictions are Protein digestion and absorption (HSA04974), Focal adhesion (HSA04510), and PI3K-Akt signaling pathway (HSA04151). The association of Focal adhesion and Lung cancer has often been shown [52]. The dysregulation of the PI3K-Akt signaling pathway is known to have an effect in lung cancer [53]. Both DAVID and Reactome tools were also able to find the associations between the data and PI3K-Akt signaling pathway in their top 3 predictions for this dataset. DAVID was also able to identify the Protein digestion and absorption pathway in its top 3 predictions for this dataset.

For the GDS3874 dataset, a study concerning Diabetes, PriPath top predictions are Viral carcinogenesis (HSA05203), C-type lectin receptor signaling pathway (HSA04625), and Human T-cell leukemia virus 1 infection (HSA05166). There is no study about the association between Diabetes and our predicted pathways for this dataset. The other tested tools did not identify these pathways in their top 3 predictions for this dataset.

For the GDS3875 dataset, which is another study on Diabetes, PriPath top predictions are Herpes simplex virus 1 infection (HSA05168), Insulin signaling pathway (HSA04910), and Pathways of neurodegeneration - multiple diseases (HSA05022). HSV-1 infection has an important association with diabetes, as explained in [54]. This disease, also known as insulin resistance, is generated by the disruption of the insulin signaling pathway [55]. Additionally, one of the conditions which result in neurodegeneration is diabetes [56]. The other tested tools were not able to detect these pathways in their top 3 predictions for the disease under investigation.

For the GDS4516 dataset, a study concerning Colorectal cancer (CRC), PriPath top predictions are Neuroactive ligand-receptor interaction (HSA04080), Synaptic vesicle cycle (HSA04721), and Glutamatergic synapse (HSA04724). The association of the Glutamatergic synapse pathway with Colorectal cancer has been frequently reported, and Glutamatergic synapse is recently verified as a new poor prognostic marker for CRC [57]. The other tested tools were not able to return these pathways in their top 3 predictions for the disease under investigation.

For the GDS4824 dataset, a study concerning Prostate cancer, PriPath top predictions are Neuroactive ligand-receptor interaction (HSA04080), Human cytomegalovirus infection (HSA05163), and Chemokine signaling pathway (HSA04062). The association of Prostate cancer and Human cytomegalovirus infection has been frequently demonstrated [58]. Chemokines play modulatory roles in prostate cancer metastasis [59]. The other tested tools were not able to identify these pathways in their top 3 predictions for the disease under investigation.

For the GDS5037 dataset, a study concerning Asthma, PriPath top predictions are Tight junction (TJ) (HSA04530), Huntington disease (HSA05016), and Pathways of neurodegeneration - multiple diseases (HSA05022). Asthma may be linked to the differential expression of TJ, as reported in [60]. Asthma, especially when it is severe, is associated with features of neuroinflammation and neurodegeneration [61]. The other tested tools were not able to detect these pathways in their top 3 predictions for the disease under study.

For the GDS5499 dataset, a study concerning Pulmonary Hypertension, PriPath top predictions are MAPK signaling pathway (HSA04010), NOD-like receptor signaling pathway (HSA04621), and Hippo signaling pathway (HSA04390). Researchers show that the inhibition of the MAPK axis could prevent vascular remodeling in Pulmonary artery hypertension [62]. NOD-like receptor signaling pathway is previously associated with Pulmonary Hypertension in [63]. PAH has been observed to ameliorate the HIPPO signaling pathway by suppressing it [64]. The other tested tools were not able to return these pathways in their top 3 predictions for the disease under investigation.

In summary, for several datasets, the top 3 predicted pathways of PriPath was associated with the disease under study in literature. Hence, we have shown that PriPath was able to successfully identify those dysregulated pathways for different diseases under investigation.

6. Conclusion

In this study, we have introduced a novel tool named PriPath that is based on the ranking and grouping of biological information. PriPath uses a set of KEGG pathways as a feature to perform classification. PriPath produces performance metrics and a list of important KEGG pathways for the disease under study. PriPath follows a similar approach with maTE and CogNet in terms of ranking a group of genes. PriPath searches for the effectiveness of distinct combinations of the KEGG pathways in terms of classifying the samples (patients). Additionally, instead of ranking each pathway's genes separately, PriPath utilizes a different approach to rank those groups concurrently. CogNet ranks each KEGG pathway by making cross-validation. Here, we integrated CogNet and maTE, where the process of the ranking is implemented on the groups generated by utilizing the biological knowledge in each tool. Furthermore, SVM-RCE utilizes k-means clustering algorithms in order to group the genes into clusters. Therefore, groups are linked with the expression of the genes. Consequently, a biologist may want to discover significant pathways and microRNA targets of these genes together in the same data. In this respect, PriPath, CogNet, maTE, and SVM-RCE integrate biological information into the machine learning algorithm and help wet-lab scientists to understand disease mechanisms at the molecular level and to generate hypotheses.

Abbreviations

ACC: Accuracy

AUC: Area Under Curve

COVID-19: Coronavirus disease

CRC: Colorectal Cancer

DAVID: Database for Annotation, Visualization and Integrated Discovery

FAIR: Findable, Accessible, Interoperable, and Reusable

FAK: Focal adhesion kinase

FGF21: Fibroblast growth factor 21

FN: False Negative

FP: False Positive

GEO: Gene Expression Omnibus

GO: Gene Ontology

G-S-M: Grouping, Scoring and Modeling

HCMV: Human cytomegalovirus

HPV: Human papillomavirus

HSV-1: Herpes simplex virus 1

IL-17: Interleukin-17

KEGG: Kyoto Encyclopedia of Genes and Genomes

KNIME: Konstanz Information Miner

MAPK: Mitogen-Activated Protein Kinase

MCCV: Monte Carlo Cross Validation

miRNA: microRNA

ML: Machine Learning

mRNA: messenger-RNA

mTOR: mammalian target of rapamycin

NCBI: National Center for Biotechnology Information

NOD: Nucleotide-oligomerization domain

non-TB: non-Topology-Based

PAH: Pulmonary arterial hypertension

PARK7: Parkinsonism Associated Deglycase

PCa: Prostate Cancer

PH: Pulmonary hypertension

PI3K-Akt: Phosphatidylinositol 3-kinase

RA: Rheumatoid arthritis

RF: Random Forest

ROC: Receiver Operator Characteristic

SE: Sensitivity

SIRT1: Sirtuin 1

SP: Specificity

SVM-RCE: Support Vector Machines -Recursive Cluster Elimination

SVM-RCE-R: Recursive Cluster Elimination based Rank Function

SVM-RNE: Support Vector Machines with Recursive Network Elimination

TB: Topology-Based

TJ: Tight junction

TLR: Toll-like receptor

TN: True Negative

TP: True Positive

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

The PriPath Knime workflow could be downloaded from:

1. https://github.com/malikyousef/PriPath.git or
2. https://kni.me/s/xdHXGrOP-D2elvvi

All the datasets used in this paper could be downloaded from GEO.

Competing interests

The authors declare that they have no competing interests.

Funding

The work of M.Y. has been supported by the Zefat Academic College. The work of B.B.G. has been supported by the Abdullah Gul University Support Foundation (AGUV).

Authors' contributions

M.Y. conceived the ideas, designed the study, analyzed the results, and he was a major contributor in writing the manuscript. F.O performed the experiments, analyzed the data, and prepared figures and/or tables. She was a major contributor in writing the manuscript. A.J. conducted the experiments, analyzed the data. J.A. analyzed the results, participated in the discussion of the results and writing of the article. B.B.G. analyzed the results, participated in the discussion of the results and writing of the article. All authors read and approved the final manuscript.

Acknowledgements

We would like to thank the anonymous reviewers for their valuable comments and suggestions to improve the quality of the paper.

References

1. Barzon, L.; Lavezzo, E.; Militello, V.; Toppo, S.; Palù, G. Applications of Next-Generation Sequencing Technologies to Diagnostic Virology. Int. J. Mol. Sci. 2011, 12, 7861–7884.

2. Ben-Dor, A.; Shamir, R.; Yakhini, Z. Clustering Gene Expression Patterns. J. Comput. Biol. 1999, 6, 281–297.

3. Dinu, I.; Potter, J.D.; Mueller, T.; Liu, Q.; Adewale, A.J.; Jhangri, G.S.; Einecke, G.; Famulski, K.S.; Halloran, P.; Yasui, Y. Gene-Set Analysis and Reduction. Brief. Bioinform. 2008, 10, 24–34, doi:10.1093/bib/bbn042.

4. Incorporating Pathway Information into Feature Selection towards Better Performed Gene Signatures Available online: https://www.hindawi.com/journals/bmri/2019/2497509/ (accessed on 8 March 2022).

5. Zhang, J.D.; Wiemann, S. KEGGgraph: A Graph Approach to KEGG PATHWAY in R and Bioconductor. Bioinforma. Oxf. Engl. 2009, 25, 1470–1471, doi:10.1093/bioinformatics/btp167.

6. Kanehisa, M.; Araki, M.; Goto, S.; Hattori, M.; Hirakawa, M.; Itoh, M.; Katayama, T.; Kawashima, S.; Okuda, S.; Tokimatsu, T. KEGG for Linking Genomes to Life and the Environment. Nucleic Acids Res. 2007, 36, D480–D484.

7. Ma, J.; Shojaie, A.; Michailidis, G. A Comparative Study of Topology-Based Pathway Enrichment Analysis Methods. BMC Bioinformatics 2019, 20, 546, doi:10.1186/s12859-019-3146-1.

8. A Critical Comparison of Topology-Based Pathway Analysis Methods Available online: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0191154 (accessed on 8 March 2022).

9. Breheny, P. The Group Exponential Lasso for Bi-Level Variable Selection: The Group Exponential Lasso for Bi-Level Variable Selection. Biometrics 2015, 71, 731–740, doi:10.1111/biom.12300.

10. Subramanian, A.; Tamayo, P.; Mootha, V.K.; Mukherjee, S.; Ebert, B.L.; Gillette, M.A.; Paulovich, A.; Pomeroy, S.L.; Golub, T.R.; Lander, E.S. Gene Set Enrichment Analysis: A Knowledge-Based Approach for Interpreting Genome-Wide Expression Profiles. Proc. Natl. Acad. Sci. 2005, 102, 15545–15550.

11. Cun, Y.; Fröhlich, H. Prognostic Gene Signatures for Patient Stratification in Breast Cancer - Accuracy, Stability and Interpretability of Gene Selection Approaches Using Prior Knowledge on Protein-Protein Interactions. BMC Bioinformatics 2012, 13, 69, doi:10.1186/1471-2105-13-69.

12. Staiger, C.; Cadot, S.; Kooter, R.; Dittrich, M.; Müller, T.; Klau, G.W.; Wessels, L.F.A. A Critical Evaluation of Network and Pathway-Based Classifiers for Outcome Prediction in Breast Cancer. PLoS ONE 2012, 7, e34796, doi:10.1371/journal.pone.0034796.

13. Kanehisa, M. The KEGG Database; 2002; Vol. 247;.

14. Consortium, T.G.O. Gene Ontology: Tool for the Unification of Biology. The Gene Ontology Consortium; 2000; Vol. 25;.

15. Jassal, B.; Matthews, L.; Viteri, G.; Gong, C.; Lorente, P.; Fabregat, A.; Sidiropoulos, K.; Cook, J.; Gillespie, M.; Haw, R.; et al. The Reactome Pathway Knowledgebase. Nucleic Acids Res. 2019, gkz1031, doi:10.1093/nar/gkz1031.

16. Yousef, M.; Kumar, A.; Bakir-Gungor, B. Application of Biological Domain Knowledge Based Feature Selection on Gene Expression Data. Entropy 2020, 23, 2, doi:10.3390/e23010002.

17. Yousef, M.; Abdallah, L.; Allmer, J. MaTE: Discovering Expressed Interactions between MicroRNAs and Their Targets. Bioinformatics 2019, 35, 4020–4028.

18. Yousef, M.; Ülgen, E.; Sezerman, O.U. CogNet: Classification of Gene Expression Data Based on Ranked Active-Subnetwork-Oriented KEGG Pathway Enrichment Analysis. PeerJ Comput. Sci. 2021, 7, e336.

19. Yousef, M.; Goy, G.; Mitra, R.; Eischen, C.M.; Jabeer, A.; Bakir-Gungor, B. MiRcorrNet: Machine Learning-Based Integration of MiRNA and MRNA Expression Profiles, Combined with Feature Grouping and Ranking. PeerJ 2021, 9, e11458.

20. Yousef, M.; Goy, G.; Bakir-Gungor, B. MiRModuleNet: Detecting MiRNA-MRNA Regulatory Modules. Rev.

21. Yousef, M.; Sayıcı, A.; Bakir-Gungor, B. Integrating Gene Ontology Based Grouping and Ranking into the Machine Learning Algorithm for Gene Expression Data Analysis. In Proceedings of the International Conference on Database and Expert Systems Applications; Springer, 2021; pp. 205–214.

22. Yousef, M.; Bakir-Gungor, B.; Jabeer, A.; Goy, G.; Qureshi, R.; Showe, L.C. Recursive Cluster Elimination Based Rank Function (SVM-RCE-R) Implemented in KNIME. F1000Research 2020, 9.

23. Yousef, M.; Jabeer, A.; Bakir-Gungor, B. SVM-RCE-R-OPT: Optimization of Scoring Function for SVM-RCE-R. In Database and Expert Systems Applications - DEXA 2021 Workshops; Kotsis, G., Tjoa, A.M., Khalil, I., Moser, B., Mashkoor, A., Sametinger, J., Fensel, A., Martinez-Gil, J., Fischer, L., Czech, G., Sobieczky, F., Khan, S., Eds.; Communications in Computer and Information Science; Springer International Publishing: Cham, 2021; Vol. 1479, pp. 215–224 ISBN 978-3-030-87100-0.

24. Yousef, M.; Jung, S.; Showe, L.C.; Showe, M.K. Recursive Cluster Elimination (RCE) for Classification and Feature Selection from Gene Expression Data. BMC Bioinformatics 2007, 8, 1–12.

25. Yousef, M.; Ketany, M.; Manevitz, L.; Showe, L.C.; Showe, M.K. Classification and Biomarker Identification Using Gene Network Modules and Support Vector Machines. BMC Bioinformatics 2009, 10, 1–7.

26. Home - GEO - NCBI Available online: https://www.ncbi.nlm.nih.gov/geo/ (accessed on 14 February 2022).

27. R: The R Project for Statistical Computing Available online: https://www.r-project.org/ (accessed on 14 February 2022).

28. KEGG PATHWAY Database Available online: https://www.genome.jp/kegg/pathway.html (accessed on 14 February 2022).

29. Dietz, C.; Berthold, M.R. KNIME for Open-Source Bioimage Analysis: A Tutorial. Focus Bio-Image Inform. 2016, 179–197.

30. Xu, Q.-S.; Liang, Y.-Z. Monte Carlo Cross Validation. Chemom. Intell. Lab. Syst. 2001, 56, 1–11, doi:10.1016/S0169-7439(00)00122-2.

31. Zhu, W.; Zeng, N.; Wang, N. Sensitivity, Specificity, Accuracy, Associated Confidence Interval and ROC Analysis with Practical SAS Implementations. NESUG Proc. Health Care Life Sci. Baltim. Md. 2010, 19, 67.

32. Floch, J.-P.L.; Escuyer, P.; Baudin, E.; Baudon, D.; Perlemuter, L. Blood Glucose Area under the Curve: Methodological Aspects. Diabetes Care 1990, 13, 172–175.

33. Dennis, G.; Sherman, B.T.; Hosack, D.A.; Yang, J.; Gao, W.; Lane, H.C.; Lempicki, R.A. DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome Biol. 2003, 4, 1–11.

34. Fabregat, A.; Sidiropoulos, K.; Viteri, G.; Forner, O.; Marin-Garcia, P.; Arnau, V.; D'Eustachio, P.; Stein, L.; Hermjakob, H. Reactome Pathway Analysis: A High-Performance in-Memory Approach. BMC Bioinformatics 2017, 18, 1–9.

35. Thanati, F.; Karatzas, E.; Baltoumas, F.A.; Stravopodis, D.J.; Eliopoulos, A.G.; Pavlopoulos, G.A. FLAME: A Web Tool for Functional and Literature Enrichment Analysis of Multiple Gene Lists. Biology 2021, 10, 665, doi:10.3390/biology10070665.

36. Limam, S.; Missaoui, N.; Hmissa, S.; Yacoubi, M.T.; Krifa, H.; Mokni, M.; Selmi, B. Investigation of Human Cytomegalovirus and Human Papillomavirus in Glioma. Cancer Invest. 2020, 38, 394–405, doi:10.1080/07357907.2020.1793352.

37. Yang, T.; Zhou, Y.; Wang, H.; Chen, S.; Shen, M.; Hu, Y.; Wang, T.; Liu, J.; Jiang, Z.; Wang, Z.; et al. Insulin Exacerbated High Glucose-Induced Epithelial-Mesenchymal Transition in Prostatic Epithelial Cells BPH-1 and Prostate Cancer Cells PC-3 via MEK/ERK Signaling Pathway. Exp. Cell Res. 2020, 394, 112145, doi:10.1016/j.yexcr.2020.112145.

38. Fenner, A. Prostate Cancer: Targeting the Ribosome in Advanced Disease. Nat. Rev. Urol. 2016, 13, 562, doi:10.1038/nrurol.2016.162.

39. Bhowmick, N.A.; Oft, J.; Dorff, T.; Pal, S.; Agarwal, N.; Figlin, R.A.; Posadas, E.M.; Freedland, S.J.; Gong, J. COVID-19 and Androgen-Targeted Therapy for Prostate Cancer Patients. Endocr. Relat. Cancer 2020, 27, R281–R292, doi:10.1530/ERC-20-0165.

40. Sun, H.; Ou, B.; Zhao, S.; Liu, X.; Song, L.; Liu, X.; Wang, R.; Peng, Z. USP11 Promotes Growth and Metastasis of Colorectal Cancer via PPP1CA-Mediated Activation of ERK/MAPK Signaling Pathway. EBioMedicine 2019, 48, 236–247, doi:10.1016/j.ebiom.2019.08.061.

41. Dmitrieva-Posocco, O.; Dzutsev, A.; Posocco, D.F.; Hou, V.; Yuan, W.; Thovarai, V.; Mufazalov, I.A.; Gunzer, M.; Shilovskiy, I.P.; Khaitov, M.R.; et al. Cell-Type-Specific Responses to Interleukin-1 Control Microbial Invasion and Tumor-Elicited Inflammation in Colorectal Cancer. Immunity 2019, 50, 166-180.e7, doi:10.1016/j.immuni.2018.11.015.

42. Bonnet, M.; Buc, E.; Sauvanet, P.; Darcha, C.; Dubois, D.; Pereira, B.; Déchelotte, P.; Bonnet, R.; Pezet, D.; Darfeuille-Michaud, A. Colonization of the Human Gut by E. Coli and Colorectal Cancer Risk. Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res. 2014, 20, 859–867, doi:10.1158/1078-0432.CCR-13-1343.

43. Yan, S.; Hui, Y.; Li, J.; Xu, X.; Li, Q.; Wei, H. Glutamine Relieves Oxidative Stress through PI3K/Akt Signaling Pathway in DSS-Induced Ulcerative Colitis Mice. Iran. J. Basic Med. Sci. 2020, 23, 1124–1129, doi:10.22038/ijbms.2020.39815.9436.

44. Low, E.N.D.; Mokhtar, N.M.; Wong, Z.; Raja Ali, R.A. Colonic Mucosal Transcriptomic Changes in Patients with Long-Duration Ulcerative Colitis Revealed Colitis-Associated Cancer Pathways. J. Crohns Colitis 2019, 13, 755–763, doi:10.1093/ecco-jcc/jjz002.

45. Okayama, S.; Arakawa, S.; Ogawa, K.; Makino, T. A Case of Hemorrhagic Colitis after Influenza A Infection. J. Microbiol. Immunol. Infect. Wei Mian Yu Gan Ran Za Zhi 2011, 44, 480–483, doi:10.1016/j.jmii.2011.04.003.

46. Makhlouf, S.; Messelmani, M.; Zaouali, J.; Mrissa, R. Cognitive Impairment in Celiac Disease and Non-Celiac Gluten Sensitivity: Review of Literature on the Main Cognitive Impairments, the Imaging and the Effect of Gluten Free Diet. Acta Neurol. Belg. 2018, 118, 21–27, doi:10.1007/s13760-017-0870-z.

47. Rashtak, S.; Murray, J.A. Celiac Disease in the Elderly. Gastroenterol. Clin. North Am. 2009, 38, 433–446, doi:10.1016/j.gtc.2009.06.005.

48. Veres-Székely, A.; Bernáth, M.; Pap, D.; Rokonay, R.; Szebeni, B.; Takács, I.M.; Lippai, R.; Cseh, Á.; Szabó, A.J.; Vannay, Á. PARK7 Diminishes Oxidative Stress-Induced Mucosal Damage in Celiac Disease. Oxid. Med. Cell. Longev. 2020, 2020, 4787202, doi:10.1155/2020/4787202.

49. Li, X.; Xu, T.; Wang, Y.; Huang, C.; Li, J. Toll-like Receptor-4 Signaling: A New Potential Therapeutic Pathway for Rheumatoid Arthritis. Rheumatol. Int. 2014, 34, 1613–1614, doi:10.1007/s00296-013-2890-1.

50. Gravallese, E.M.; Schett, G. Effects of the IL-23-IL-17 Pathway on Bone in Spondyloarthritis. Nat. Rev. Rheumatol. 2018, 14, 631–640, doi:10.1038/s41584-018-0091-8.

51. Lang, S.C.; Harre, U.; Purohit, P.; Dietel, K.; Kienhöfer, D.; Hahn, J.; Baum, W.; Herrmann, M.; Schett, G.; Mielenz, D. Neurodegeneration Enhances the Development of Arthritis. J. Immunol. Baltim. Md 1950 2017, 198, 2394–2402, doi:10.4049/jimmunol.1601472.

52. Aboubakar Nana, F.; Lecocq, M.; Ladjemi, M.Z.; Detry, B.; Dupasquier, S.; Feron, O.; Massion, P.P.; Sibille, Y.; Pilette, C.; Ocak, S. Therapeutic Potential of Focal Adhesion Kinase Inhibition in Small Cell Lung Cancer. Mol. Cancer Ther. 2019, 18, 17–27, doi:10.1158/1535-7163.MCT-18-0328.

53. Yu, X.; Li, Y.; Jiang, G.; Fang, J.; You, Z.; Shao, G.; Zhang, Z.; Jiao, A.; Peng, X. FGF21 Promotes Non-Small Cell Lung Cancer Progression by SIRT1/PI3K/AKT Signaling. Life Sci. 2021, 269, 118875, doi:10.1016/j.lfs.2020.118875.

54. Sun, Y.; Pei, W.; Wu, Y.; Yang, Y. An Association of Herpes Simplex Virus Type 1 Infection with Type 2 Diabetes. Diabetes Care 2005, 28, 435–436, doi:10.2337/diacare.28.2.435.

55. Chakraborty, C.; Doss, C.G.P.; Bandyopadhyay, S.; Agoramoorthy, G. Influence of MiRNA in Insulin Signaling Pathway and Insulin Resistance: Micro-Molecules with a Major Role in Type-2 Diabetes. Wiley Interdiscip. Rev. RNA 2014, 5, 697–712, doi:10.1002/wrna.1240.

56. Kang, K.; Xu, P.; Wang, M.; Chunyu, J.; Sun, X.; Ren, G.; Xiao, W.; Li, D. FGF21 Attenuates Neurodegeneration through Modulating Neuroinflammation and Oxidant-Stress. Biomed. Pharmacother. Biomedecine Pharmacother. 2020, 129, 110439, doi:10.1016/j.biopha.2020.110439.

57. Yu, Q.; Wang, X.; Yang, Y.; Chi, P.; Huang, J.; Qiu, S.; Zheng, X.; Chen, X. Upregulated NLGN1 Predicts Poor Survival in Colorectal Cancer. BMC Cancer 2021, 21, 884, doi:10.1186/s12885-021-08621-x.

58. Moon, J.-S.; Lee, M.-Y.; Park, S.W.; Han, W.K.; Hong, S.-W.; Ahn, J.-H.; Kim, K.-S. Androgen-Dependent Activation of Human Cytomegalovirus Major Immediate-Early Promoter in Prostate Cancer Cells. The Prostate 2008, 68, 1450–1460, doi:10.1002/pros.20817.

59. Adekoya, T.O.; Richardson, R.M. Cytokines and Chemokines as Mediators of Prostate Cancer Metastasis. Int. J. Mol. Sci. 2020, 21, E4449, doi:10.3390/ijms21124449.

60. Chen, X.; Corry, D.B.; Li, E. Mechanisms of Allergy and Adult Asthma. Curr. Opin. Allergy Clin. Immunol. 2020, 20, 36–42, doi:10.1097/ACI.0000000000000601.

61. Rosenkranz, M.A.; Dean, D.C.; Bendlin, B.B.; Jarjour, N.N.; Esnault, S.; Zetterberg, H.; Heslegrave, A.; Evans, M.D.; Davidson, R.J.; Busse, W.W. Neuroimaging and Biomarker Evidence of Neurodegeneration in Asthma. J. Allergy Clin. Immunol. 2022, 149, 589-598.e6, doi:10.1016/j.jaci.2021.09.010.

62. Yan, S.; Wang, Y.; Liu, P.; Chen, A.; Chen, M.; Yao, D.; Xu, X.; Wang, L.; Huang, X. Baicalin Attenuates Hypoxia-Induced Pulmonary Arterial Hypertension to Improve Hypoxic Cor Pulmonale by Reducing the Activity of the P38 MAPK Signaling Pathway and MMP-9. Evid.-Based Complement. Altern. Med. ECAM 2016, 2016, 2546402, doi:10.1155/2016/2546402.

63. Zha, L.-H.; Zhou, J.; Li, T.-Z.; Luo, H.; He, J.-N.; Zhao, L.; Yu, Z.-X. NLRC3: A Novel Noninvasive Biomarker for Pulmonary Hypertension Diagnosis. Aging Dis. 2018, 9, 843–851, doi:10.14336/AD.2017.1102.

64. Zuo, W.; Liu, N.; Zeng, Y.; Xiao, Z.; Wu, K.; Yang, F.; Li, B.; Song, Q.; Xiao, Y.; Liu, Q. Luteolin Ameliorates Experimental Pulmonary Arterial Hypertension via Suppressing Hippo-YAP/PI3K/AKT Signaling Pathway. Front. Pharmacol. 2021, 12, 663551, doi:10.3389/fphar.2021.663551.

Figures

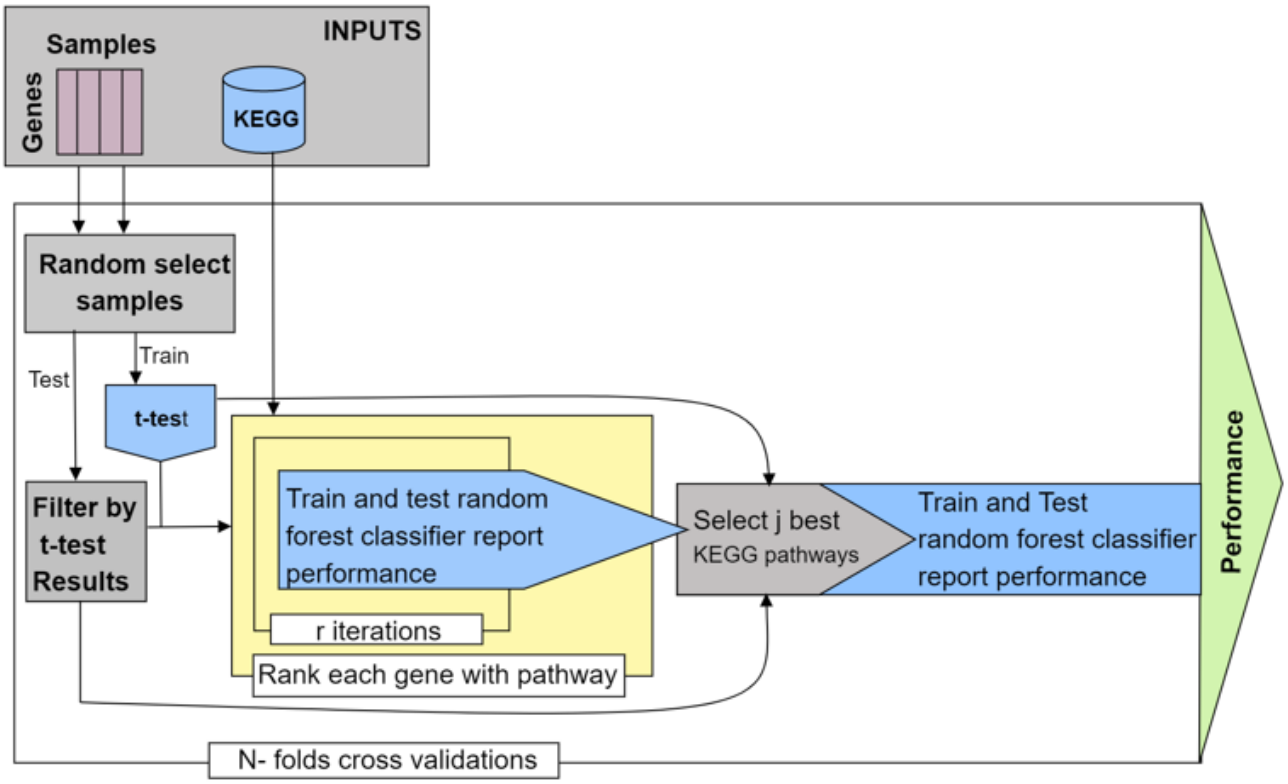

Figure 1

PriPath workflow. Two main steps of the workflow create models for each KEGG pathway, then merge multiple pathways into one model and finally train a classifier utilizing these KEGG pathways. Input data is samples including the two classes. Genes are depicted by the vertical bars. KEGG shows the target data from the KEGG database. Loops are symbolized by rectangles with a tag (e.g. N-fold cross-validation). The t-test computations are based on the training data.

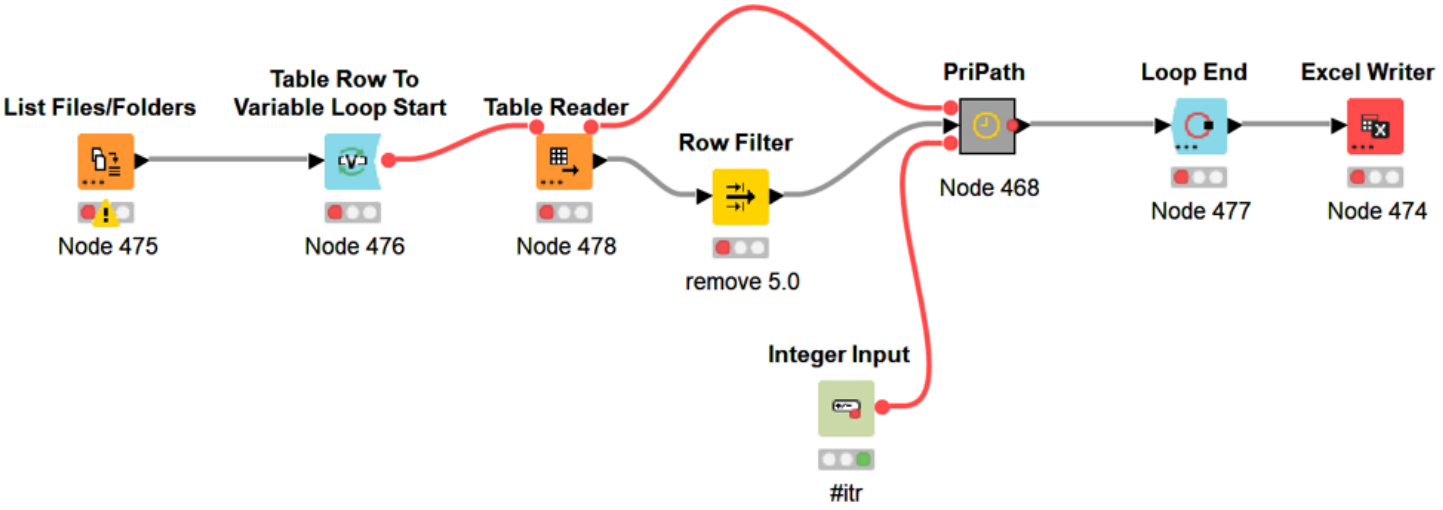

Figure 2

Multi-File PriPath workflow that can be applied on multiple datasets. Workflow control contains programming structures such as branching and loops (shown in blue), input nodes (shown as orange boxes), and collecting results (green box). In the blue and yellow boxes, the main processing is performed. Meta-nodes, shown in gray, have sub-workflows to raise the modularity and for better readability. The green dots under the nodes demonstrate that the process has been successfully performed.

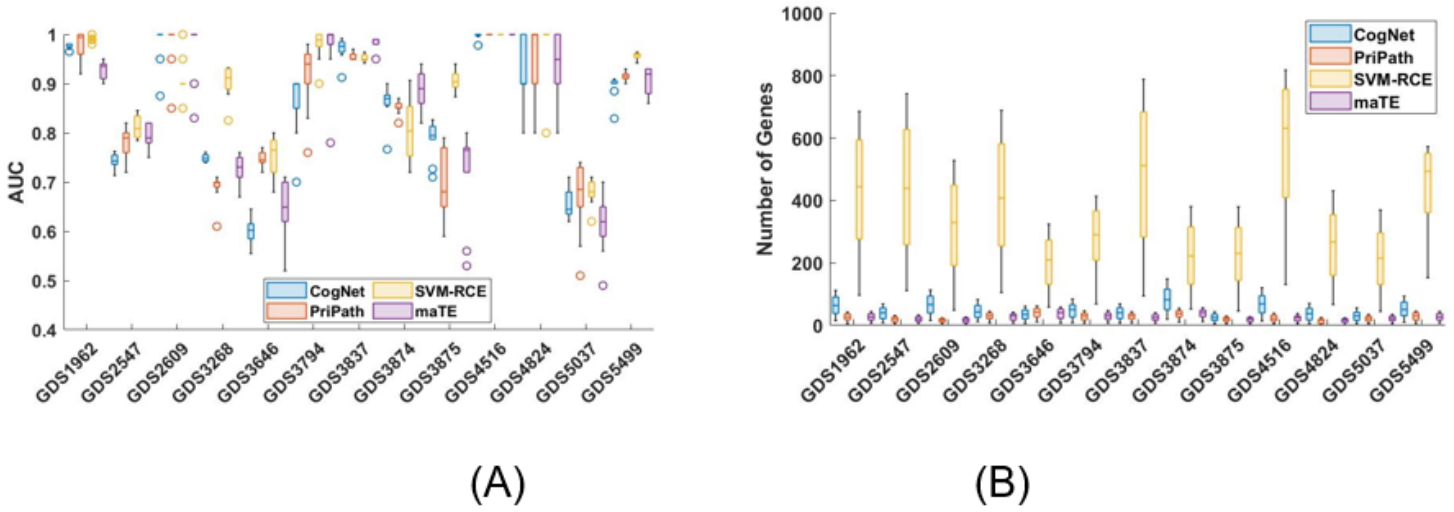

Figure 3

Performance evaluations of PriPath, CogNet, maTE, SVM-RCE. (A) AUC values; and (B) number of genes of 4 competitor tools for 13 different datasets, for 10 cluster levels.

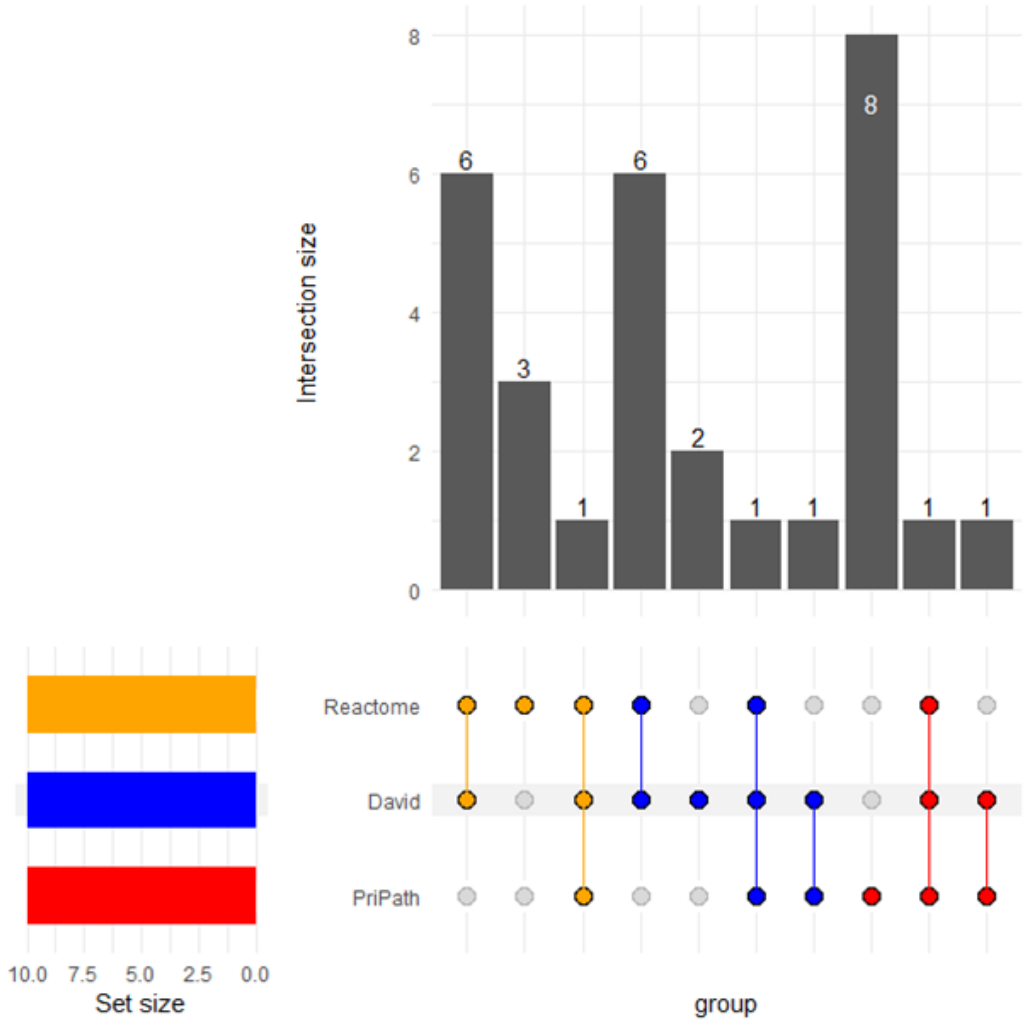

Figure 4

Comparative Evaluation of PriPath with functional enrichment analysis provided by Reactome and DAVID for the GDS1962 dataset. Results for the other datasets are available in the Supplementary Figures 1-12.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Appendix.docx