Application of biological domain knowledge-based feature selection on gene expression data

Malik Yousef
Zefat Academic College
E-mail: malik.Yousef@gmail.com

The Sixth International Conference on Informatics and Assistive Technologies for Health-Care, Medical Support and Wellbeing
HEALTHINFO 2021
October 03, 2021 to October 07, 2021 - Barcelona, Spain

Short Resume

Prof Malik Y. Yousef is a data scientist, with a focus on bioinformatics with applications to various biomedical/biological problems. He has published more than 70 peer-reviewed articles in top journals and proceedings with over 3600 citations and an H-index of 23 and i10-index of 33 (based on Google scholar).
His international experience includes 3 years as a postdoc at The Wistar Institute, Cancer Center, USA and one year at the University of Pennsylvania. Currently, he is an Associate Professor at the Zefat Academic College in Israel and Pranab K. Sen Distinguished Visiting Professor at the University of North Carolina at Chapel Hill in the Department of Biostatistics at the Gillings School of Global Public Health.

The Topics of Research Interest

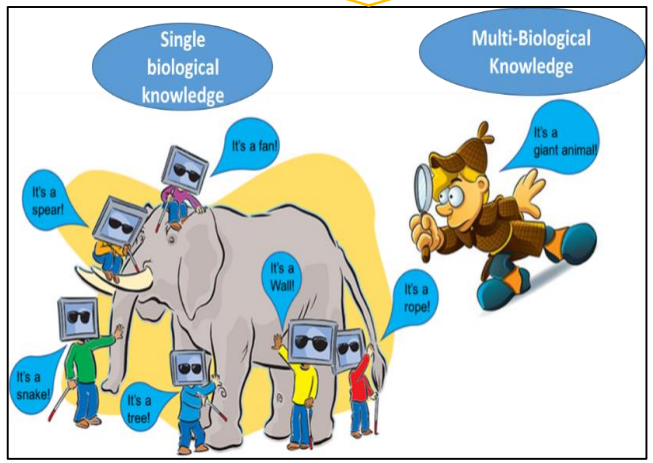
- Biological Domain Knowledge Based Feature Selection, Applied on Gene Expression Data
- Multi-Omics Data Integration
- Applications of Machine Learning in Human Microbiome
- Multi-One class
- Text classification based Topics

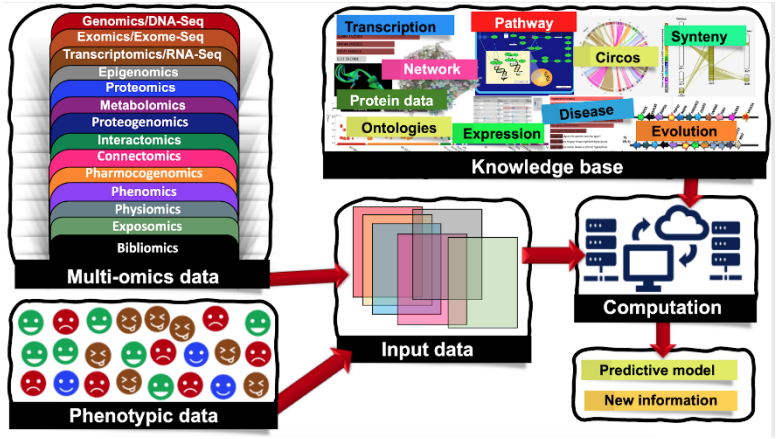
Biological Integrative Model

Review
Application of Biological Domain Knowledge Based Feature Selection on Gene Expression Data
Malik Yousef * , Abhishek Kumar , Burcu Bakir-Gungor

https://www.preprints.org/manuscript/202012.0377/v1

Provide a holistic view of biological processes.





entropy

MDPI

Review

Application of Biological Domain Knowledge Based Feature Selection on Gene Expression Data

Malik Yousef [1,2,*], Abhishek Kumar[3] and Burcu Bakir-Gungor[4]

[1] *Department of Information Systems, Zefat Academic College, Zefat, 13206, Israel*
[2] *Galilee Digital Health Research Center (GDH), Zefat Academic College, Israel*
malik.yousef@gmail.com
[3] Institute of Bioinformatics, International Technology Park, Bangalore, 560066 and Manipal Academy of Higher Education (MAHE), Manipal 576104, Karnataka, India; abhishek@ibioinformatics.org
[4] Department of Computer Engineering, Faculty of Engineering, Abdullah Gul University, Kayseri, Turkey; burcu.gungor@agu.edu.tr
* Correspondence: malik.yousef@gmail.com;

Received: date; Accepted: date; Published: date

Abstract: In the last two decades, there have been massive advancements in high throughput technologies, which resulted in the exponential growth of public repositories of gene expression datasets for various phenotypes. It is possible to unravel biomarkers by comparing the gene expression levels under different conditions, such as disease vs. control, treated vs. not treated, drug A vs. drug B, etc. This problem refers to a well-studied problem in the machine learning domain, i.e., the feature selection problem. In biological data analysis, most of the computational feature selection methodologies were taken from other fields, without considering the nature of the biological data. Thus, integrative approaches that utilize the biological knowledge while performing feature selection are necessary for this kind of data. The main idea behind the integrative gene selection process is to generate a ranked list of genes considering both the statistical metrics that are applied to the gene expression data, and the biological background information which is provided as external datasets.

One of the main goals of this review is to explore the existing methods that integrate different types of information in order to improve the identification of the biomolecular signatures of diseases and the discovery of new potential targets for treatment. These integrative approaches are expected to aid the prediction, diagnosis, and treatment of diseases as well as to enlighten disease state dynamics, mechanisms of their onset and progression. The integration of various types of biological information will necessitate the development of novel techniques for the integration and data analysis. Another aim of this review is to boost the bioinformatics community to develop new approaches for searching and determining significant groups/clusters of features based on one or more biological grouping functions.

Keywords: Feature Selection, Feature Ranking, Grouping, Clustering, Biological Knowledge.

Gene Expression

Gene ID	GSM282855	GSM282856	GSM282857	GSM282858	GSM282859	GSM282860	GSM282861	GSM282862	GSM282863	GSM28293	GSM282938	GSM282940	GSM28294	GSM28294	GSM28294	GSM28294	GSM28294	GSM28294	GSM28294	GSM28295	GSM28295	GSM28295
Class	neg	neg	neg	neg	neg	neg	neg	neg	neg	pos	pos	pos	pos	pos	pos	pos	pos	pos	pos	pos	pos	pos
A_23_P80353	0.039	-0.109	-0.044	-0.033	-0.044	0.027	-0.045	-0.059	-0.043	-0.010	-0.108	-0.140	0.034	-0.027	0.030	-0.030	0.064	-0.003	-0.026	-0.018	-0.081	0.026
APBA2	-0.133	-0.109	-0.136	-0.119	-0.164	-0.185	-0.077	-0.184	-0.108	-0.231	-0.281	-0.118	-0.145	-0.088	-0.156	-0.076	-0.036	-0.116	-0.092	-0.063	-0.185	-0.078
MAP3K6	-0.042	0.042	0.018	0.063	0.072	0.059	0.292	0.101	0.125	0.058	0.038	-0.021	-0.075	-0.051	0.173	-0.018	-0.051	0.077	0.005	0.051	-0.029	-0.133
ZNF121	-0.212	-0.240	-0.339	-0.314	-0.436	-0.050	-0.318	-0.286	-0.016	-0.550	-0.444	-0.313	-0.204	-0.226	-0.454	-0.210	-0.426	-0.210	-0.025	-0.350	0.129	-0.403
Pro25G	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
RET	0.000	-0.115	-0.442	-0.250	-0.[illegible]	-0.467	-0.141	-0.113	-0.430	-0.253	-0.165	-0.267	-0.442	-0.345	0.111	-0.151	0.000	-0.089	-0.028	-0.068	-0.005	-0.068
BX094364	0.035	0.015	-0.054	-0.058	-0.[illegible]	[illegible]	0.037	-0.099	0.027	0.064	0.018	0.040	-0.014	0.053	-0.027	-0.002	0.032	0.061	0.058	0.022	0.058	0.096
C12orf56	0.029	-0.042	0.031	-0.058	-0.[illegible]	[illegible]	0.038	-0.097	-0.027	0.047	-0.031	-0.042	-0.05[illegible]	[illegible]121	0.006	0.038	0.001	0.134	-0.005	0.026	0.068	0.100
E1A_r60_a20	0.000	0.000	0.000	0.000	0.000	[illegible]	0.000	0.000	0.000	0.000	0.000	0.000	0.[illegible]	[illegible]0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
MBD2	0.151	0.211	0.196	0.216	[illegible]	[illegible]220	0.190	0.275	0.281	0.209	0.154	0.265	[illegible]	[illegible]	0.139	0.194	0.177	0.132	0.040	0.107	0.150	0.186
LIFR-AS1	-0.268	-0.479	-0.323	-0.343	[illegible]	-0.434	-0.358	-0.481	-0.445	-0.449	-0.636	-0.363	[illegible]	[illegible]	[illegible]713	-0.597	-0.189	0.045	-0.234	-0.118	-0.248	-0.212
Pro25G	0.000	0.000	0.000	0.000	[illegible]	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.0[illegible]	[illegible]00	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
CMYA5	0.000	-0.033	-0.275	0.0[illegible]	[illegible]	-0.289	-0.054	-0.180	-0.350	-0.154	0.012	-0.335	-0.094	[illegible]	[illegible]	-0.286	0.192	0.123	-0.163	0.116	-0.156	0.087
DEUP1	0.793	0.478	0.149	0.[illegible]	[illegible]	0.751	1.191	0.769	0.299	-0.165	0.040	1.261	0.000	[illegible]	[illegible]	1.361	0.050	0.000	0.000	0.000	0.000	0.000
MIGA2	-0.067	0.036	-0.054	[illegible]	[illegible]	-0.150	-0.072	-0.083	0.003	0.046	-0.027	0.009	0.010	0.[illegible]	[illegible]	[illegible]032	0.115	0.005	0.032	-0.006	0.121	0.016
TMPRSS2	1.273	1.335	1.221	[illegible]	[illegible]3	0.861	0.818	1.240	1.043	0.973	1.083	1.038	1.116	1.03[illegible]	[illegible]	[illegible]	1.031	0.827	0.565	0.430	1.004	0.818
A_32_P58999	0.225	0.246	0.239	[illegible]	[illegible]234	0.080	0.055	0.134	0.149	-0.006	0.131	0.247	0.094	0.133	[illegible]	[illegible]	0.143	0.061	0.014	0.016	0.193	0.088
EIF5B	0.000	0.000	0.000	[illegible]	[illegible]000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	[illegible]	[illegible]	[illegible]000	0.000	0.000	0.000	0.000	0.000
Pro25G	0.000	0.000	0.000	[illegible]	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.00[illegible]	[illegible]	[illegible]00	0.000	0.000	0.000	0.000	0.000
A_32_P188127	0.294	0.044	0.43[illegible]	[illegible]	0.125	0.001	0.122	0.640	0.379	0.378	0.844	0.717	0.492	-0.106	0.223	[illegible]	[illegible].372	0.272	0.107	0.177	0.380	0.092
UCHL5	0.000	0.000	0.000	[illegible]	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	[illegible]	0.000	0.000	0.000	0.000	0.000	0.000
AF230200	0.014	0.053	-0.005	0.05[illegible]	0.006	0.007	0.025	0.018	0.034	0.080	0.042	0.027	0.019	0.010	0.034	0.0[illegible]	0.019	0.065	-0.016	0.020	0.020	0.030
MAPRE1	-0.204	-0.189	-0.357	-0.241	-0.173	-0.141	-0.231	-0.143	-0.147	-0.301	-0.322	-0.133	-0.121	-0.491	-0.191	-0.314	-0.250	-0.295	-0.408	-0.330	-0.182	-0.367
FOXA2	0.228	0.172	0.169	0.161	0.083	0.115	0.186	0.124	0.128	0.161	0.210	0.124	0.116	0.356	0.192	0.186	0.083	0.297	0.157	0.169	0.199	0.177
ICAM1	-0.383	-0.663	-0.661	-0.456	-0.259	0.053	-0.507	-0.405	-0.332	-0.388	-0.479	-0.311	-0.394	-0.485	-0.489	-0.520	-0.289	-0.407	-0.323	-0.271	-0.430	-0.466
Pro25G	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
RFWD3	-0.356	-0.604	-0.614	-0.585	-0.560	-0.362	-0.610	-0.523	-0.239	-0.480	-0.552	-0.364	-0.748	-0.447	-0.323	-0.524	-0.145	-0.300	-0.282	-0.270	-0.429	-0.392
EPYC	0.000	0.000	-0.034	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
TMEM74B	-0.019	-0.073	-0.015	-0.111	-0.062	-0.079	-0.055	-0.130	-0.030	0.042	0.028	-0.050	-0.102	0.048	-0.030	0.062	-0.065	0.021	0.018	0.033	-0.023	0.031
A_32_P159668	-0.065	-0.075	-0.181	-0.158	-0.068	-0.194	-0.389	0.150	-0.150	-0.140	-0.057	-0.047	-0.196	-0.055	-0.126	-0.086	0.250	0.015	-0.012	-0.059	-0.153	-0.027
ZNF337	-0.201	-0.204	-0.210	-0.219	-0.133	-0.023	-0.076	0.018	-0.037	-0.084	-0.206	-0.225	-0.171	-0.159	-0.023	-0.174	-0.313	0.000	-0.057	-0.017	-0.194	-0.070
A_32_P122907	0.125	0.162	0.063	-0.022	0.000	0.079	0.133	0.006	0.021	0.106	0.290	-0.033	-0.093	0.281	0.099	0.078	0.284	0.259	0.211	0.216	0.054	0.065
Pro25G	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
A_23_P327643	-0.358	-0.346	-0.320	-0.674	-0.505	-0.471	-0.251	-0.339	-0.165	-0.513	-0.704	-0.509	-0.802	-0.521	-0.588	-0.509	-0.827	-0.197	-0.356	-0.288	-0.451	-0.177
SIM2	-0.217	-0.144	-0.257	-0.300	-0.332	-0.198	-0.226	-0.224	-0.069	-0.172	-0.346	-0.259	-0.291	-0.431	-0.230	-0.212	-0.288	-0.061	-0.077	-0.137	-0.139	-0.153

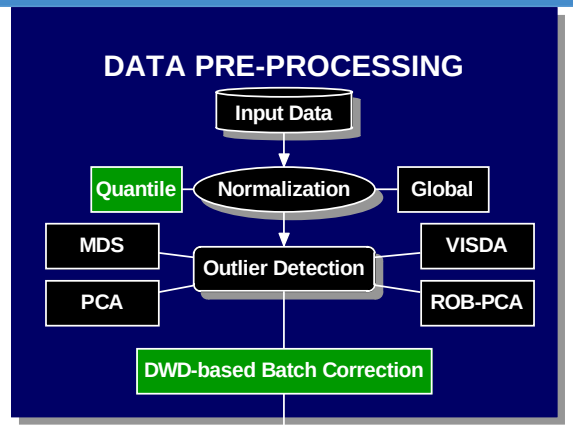
Cancer

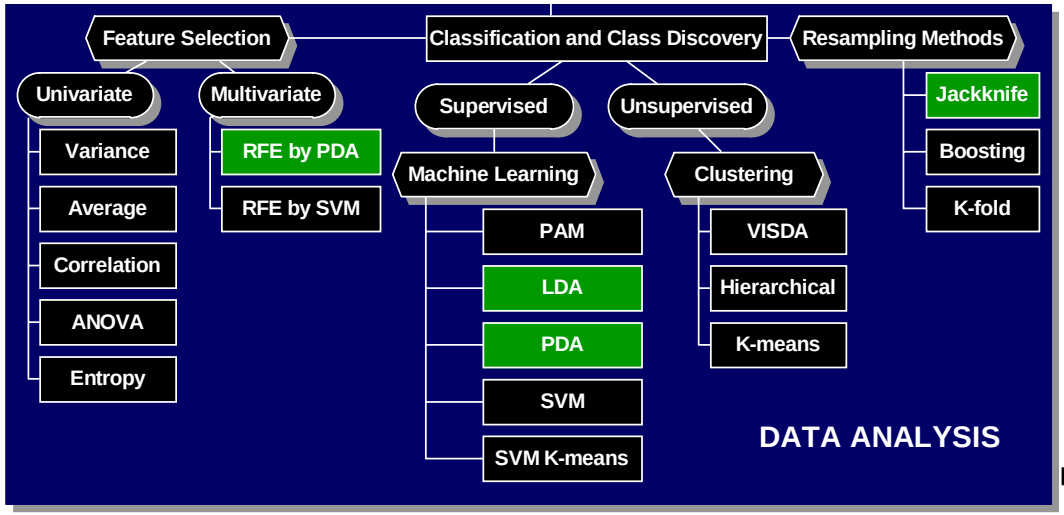
Healthy

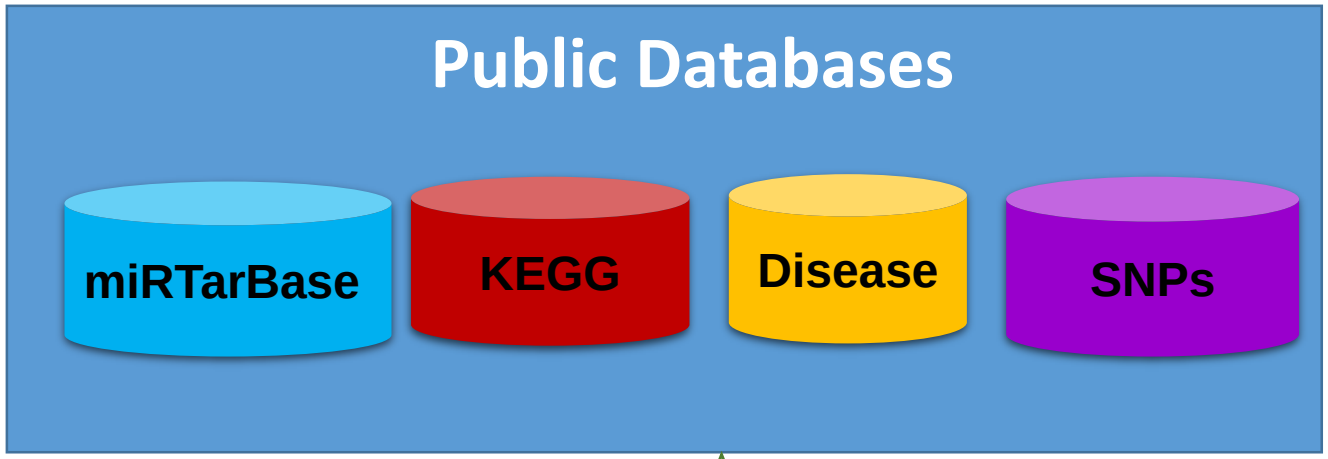
The Merit of Our Approaches

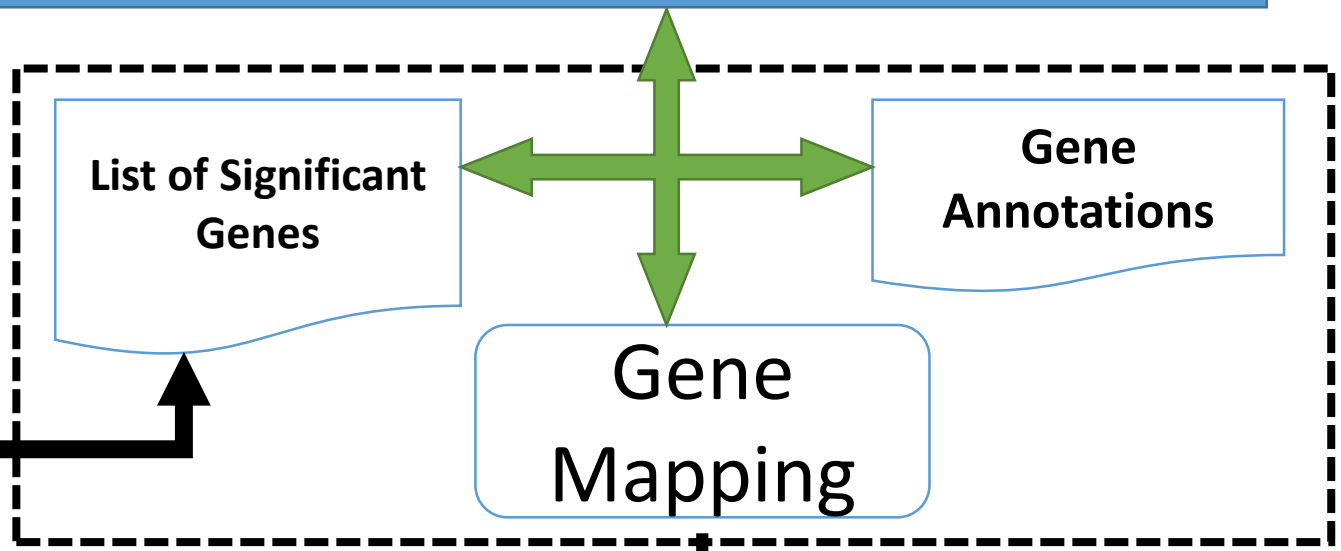
Most of the existing feature selection approaches have been borrowed from the field of computer science and statistics; they do not consider the **Biological Domain Knowledge**

The Traditional Approach









In most of the existing methods, the selection of genes is still based on **expression values alone** and biological knowledge ***is integrated*** at the end of analysis to verify experimental results or to gain biological insights.

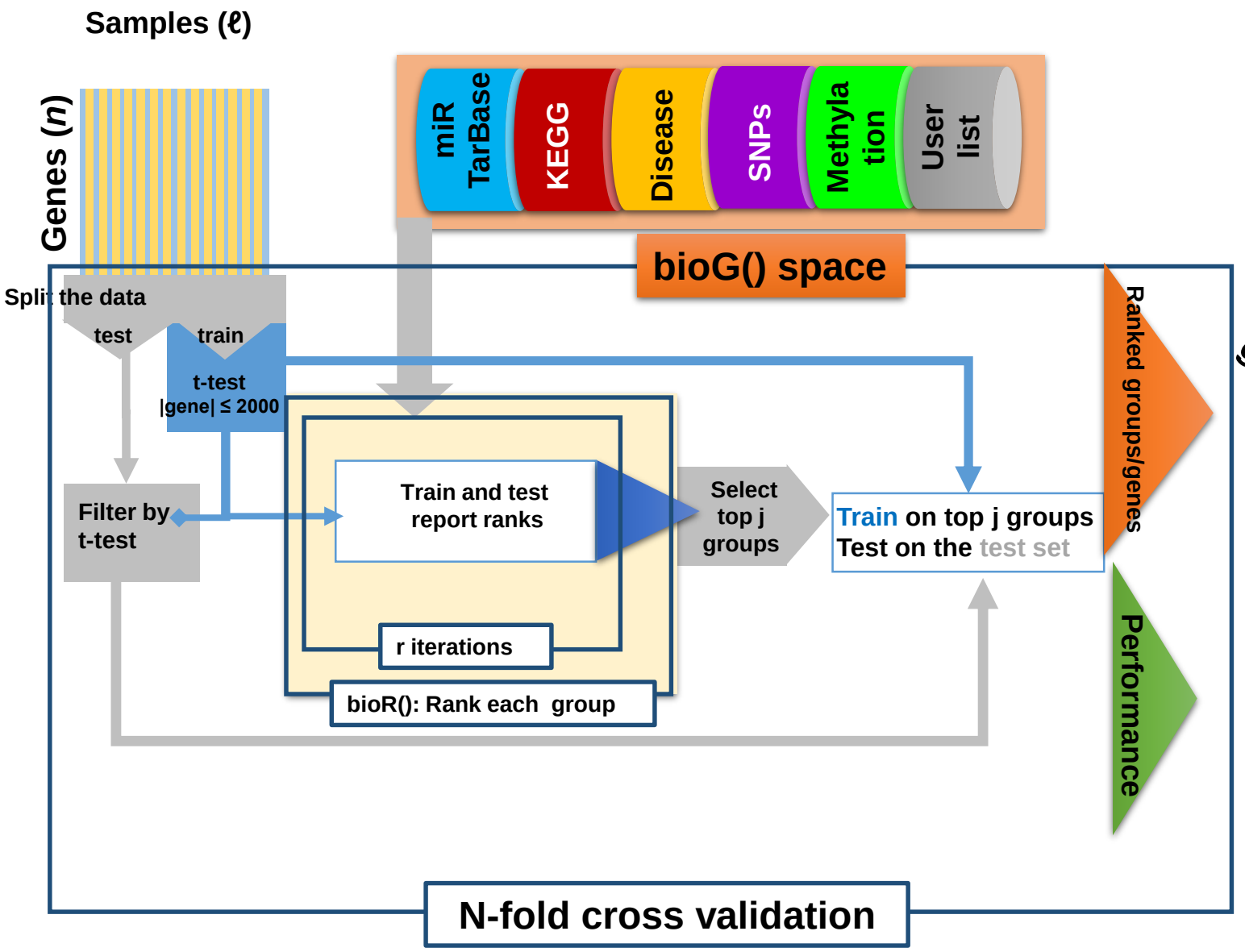
O. H. Fang, N. Mustapha, and Md. N. Sulaiman, "An integrative gene selection with association analysis for microarray data classification," Intell. Data Anal., vol. 18, no. 4, pp. 739–758, Jun. 2014, doi: 10.3233/IDA-140666.

Biological Integrative Approach

Take advantage of **multiple published gene expression datasets**, the integrative analysis of gene expression data has become an effective tool by **aggregating** multiple datasets and increasing the statistical power in identifying a small subset of genes to effectively predict the type of the disease.

Yang, ZY., Liu, XY., Shu, J. et al. Multi-view based integrative analysis of gene expression data for identifying biomarkers. Sci Rep 9, 13504 (2019). https://doi.org/10.1038/s41598-019-49967-4

Our Biological Integrative Approach



Personal Research Grants
Research Grant Application no. 3032/21

Role	Name	Academic Rank	Department	Institute
PI.1	Malik Yousef	Associate Professor	Information System	Zefat Academic College

Research Title

Integrating Biological Patterns And Gene Expression Data Using Machine Learning: Uncovering Gene Patterns Associated With Disease Phenotypes

Aggregations of ranked groups/genes

Ranked Significant Biological Groups and Genes

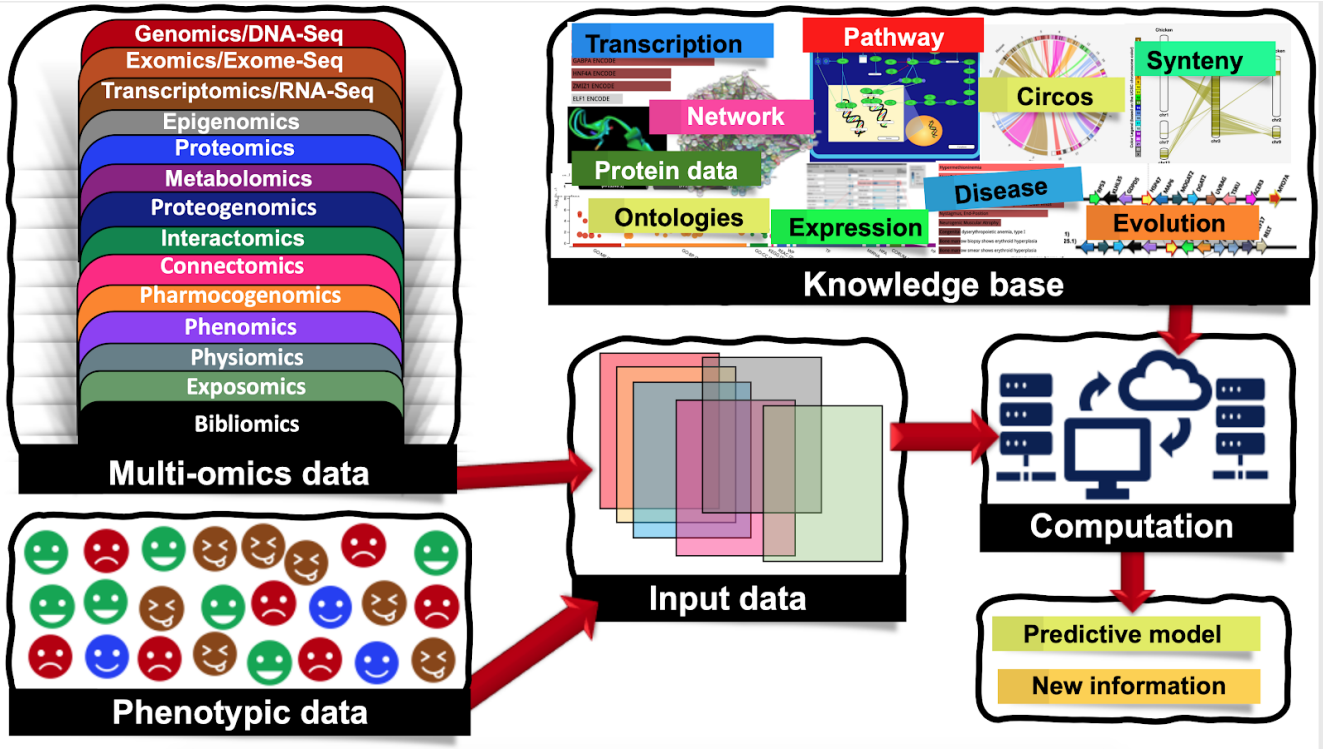
Performance Aggregations

Summary Table over top *j* groups

Integrative Approach

- The rising interest in integrative approach has shifted gene selection from purely data-centric to **incorporating** additional biological knowledge.

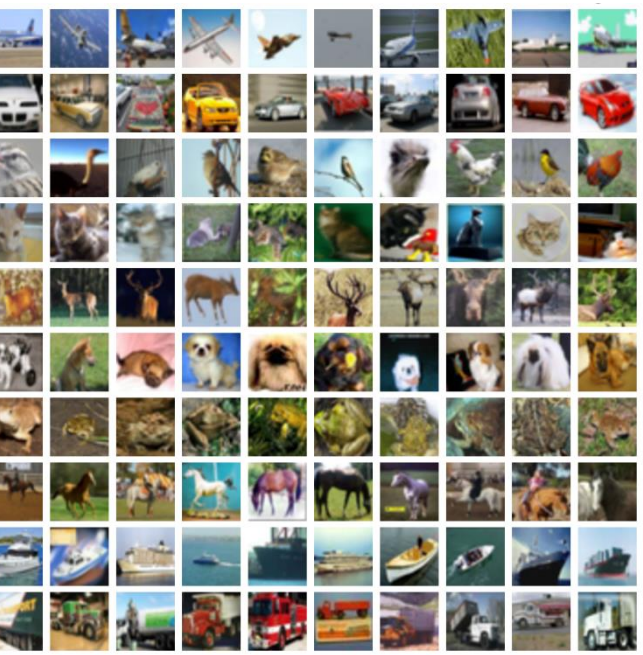
- Integrative gene selection is viewed as a promising approach in microarray data classification that took into consideration the complex relationships among genes



A Review Paper:
Application of Biological Domain Knowledge Based Feature Selection on Gene Expression Data. Entropy
Malik Yousef, Abhishek Kumar and Burcu Bakir-Gungor

entropy

Machine Learning



Original Data Set

Training Set

Test Set

Partition Data

Train Model

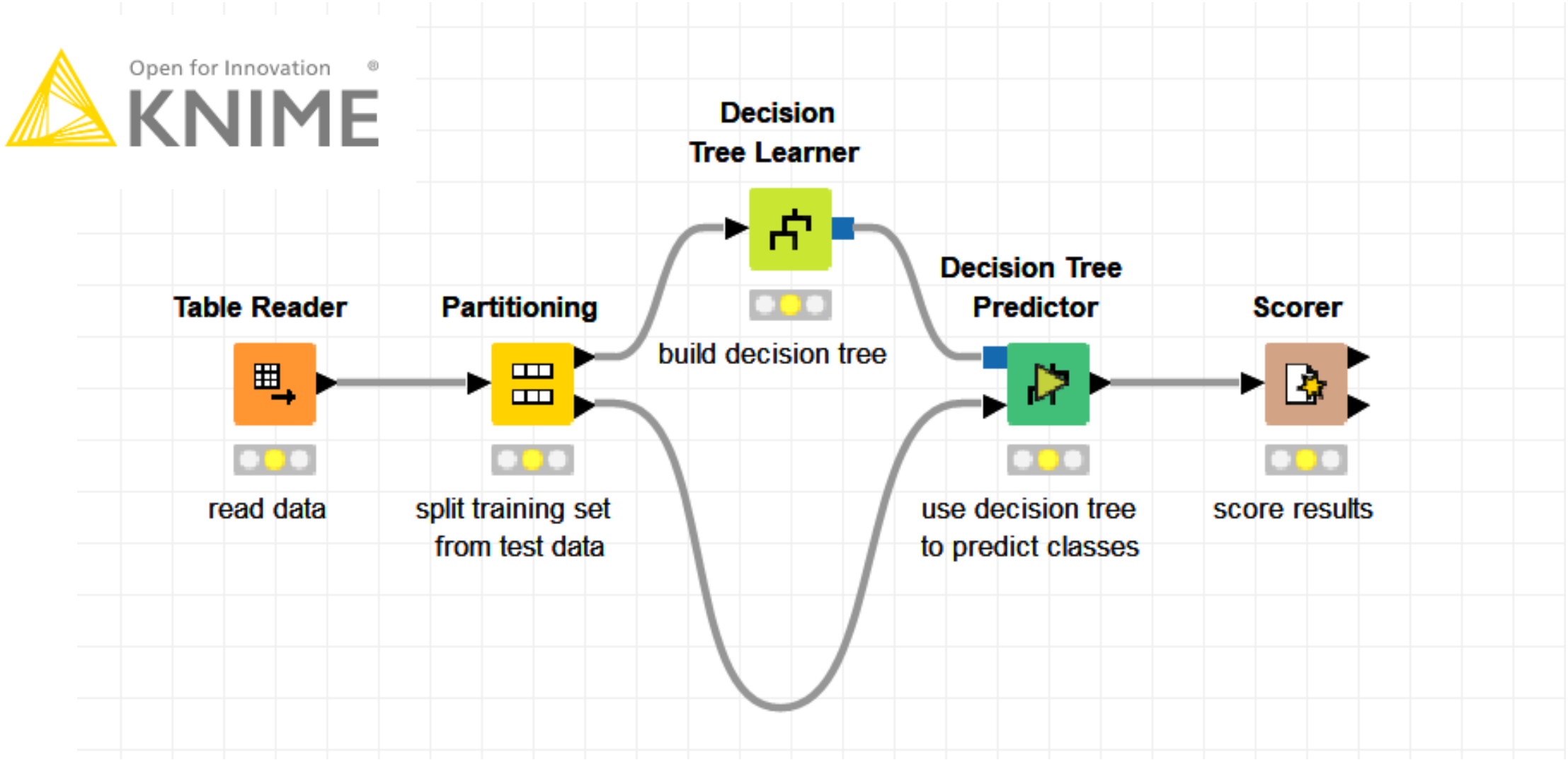
Apply Model

Train and apply models

Score Model

Evaluate

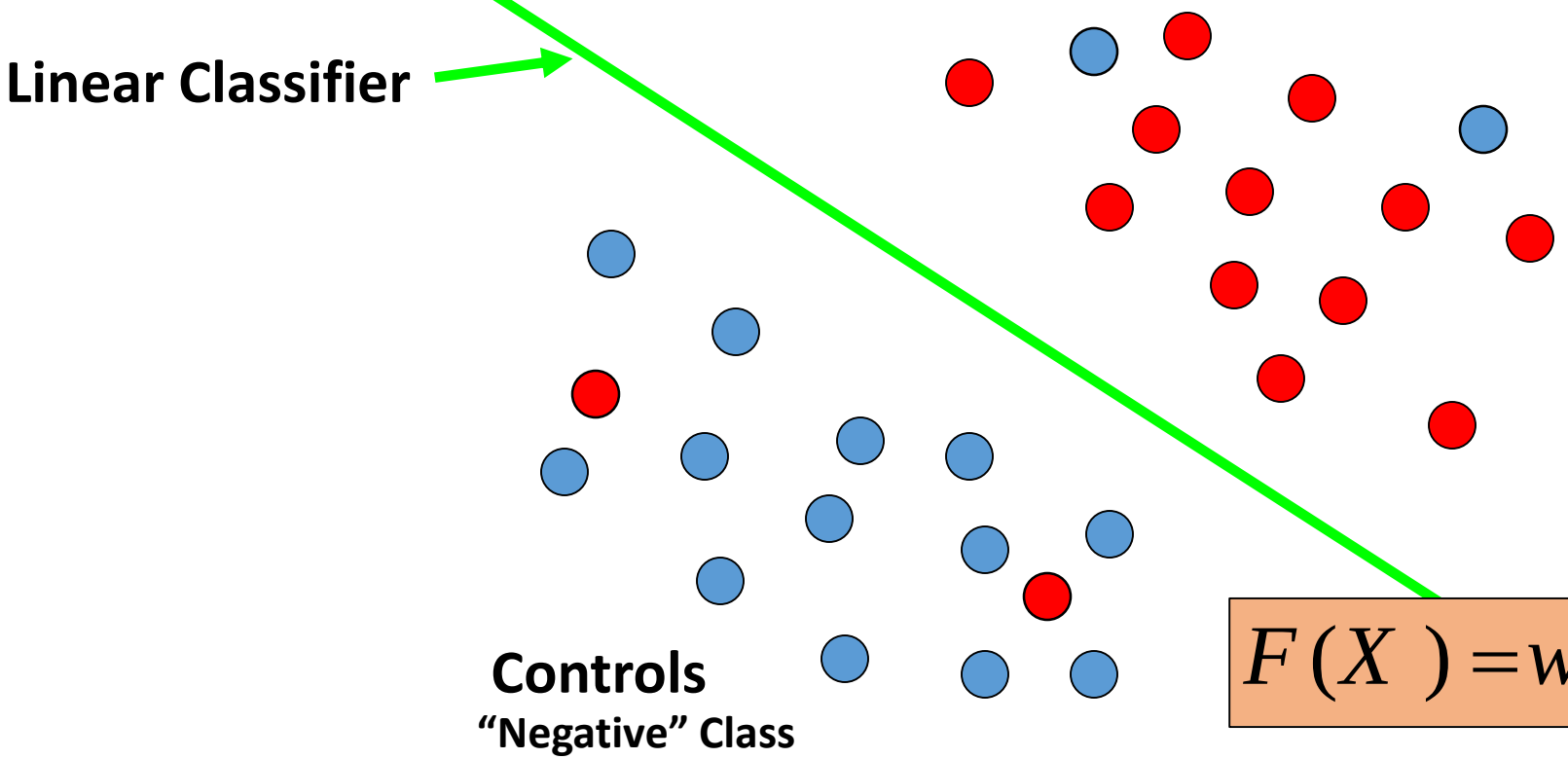
Simple Knime Work Flow



Classification through Machine Learning

Linear Discriminant Analysis

Simple form enables easier interpretation

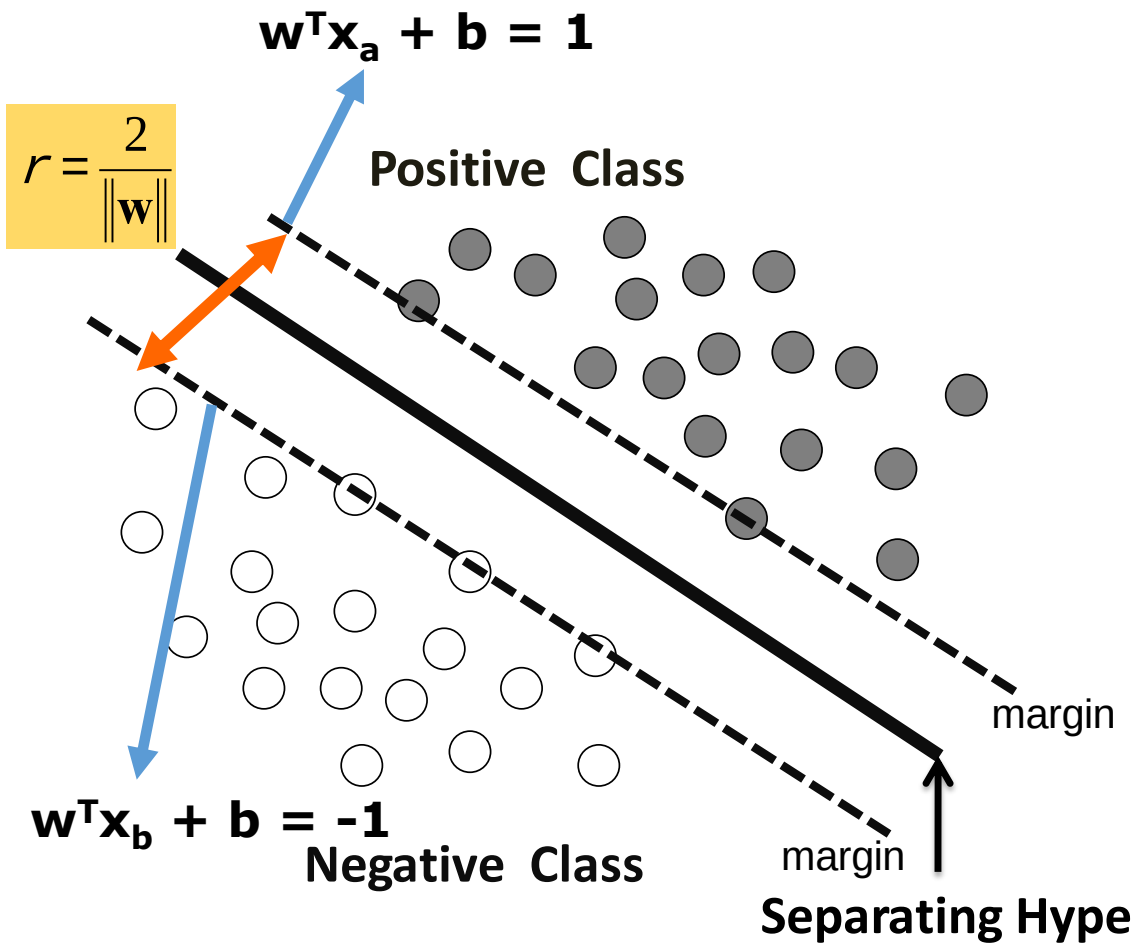


Features with low weights are **weak features-remove from the model**

$$F(X) = w_0 + w_1 x_1 + \ldots + w_n x_n$$

$w_0, w_1, \ldots, w_n$ are the score for each feature/gene

SVM-Support Vector Machine



Quadratic optimization problem

Find $\mathbf{w}$ and b such that

$r = \frac{2}{\|\mathbf{w}\|}$ is maximized; and for all $\{(\mathbf{x_i}, y_i)\}$

$\mathbf{w}^T\mathbf{x_i} + b \geq 1$ if $y_i=1$; $\mathbf{w}^T\mathbf{x_i} + b \leq -1$ if $y_i = -1$

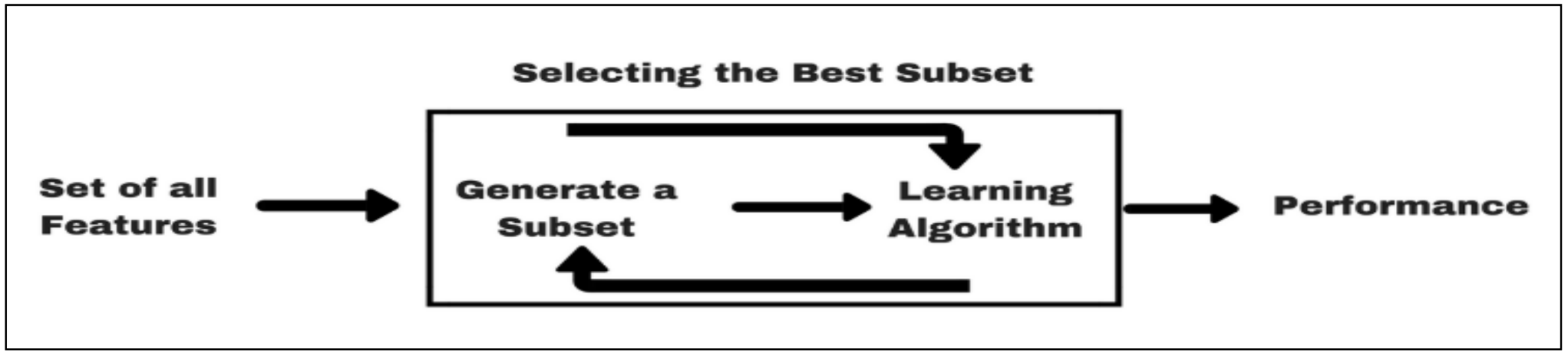
Find $\mathbf{w}$ and b such that

$\Phi(\mathbf{w}) = \frac{1}{2}\mathbf{w}^T\mathbf{w}$ is minimized;

and for all $\{(\mathbf{x_i}, y_i)\}$: $y_i(\mathbf{w}^T\mathbf{x_i} + b) \geq 1$

Feature Selection

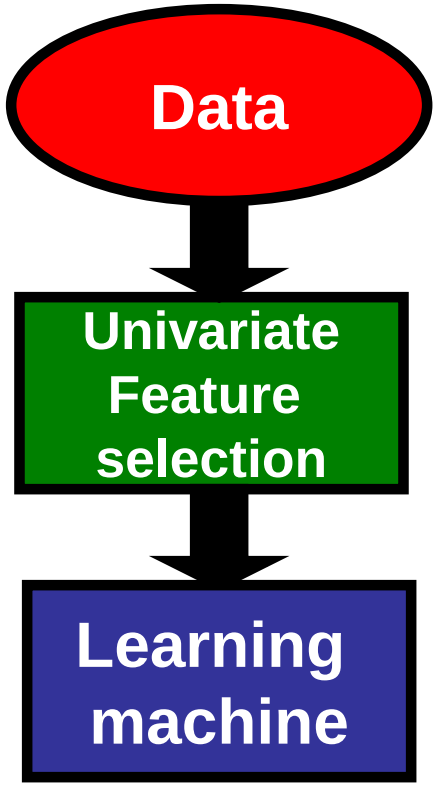
- **Feature Selection is one of the core concepts in machine learning which hugely impacts the performance of your model.** The data features that you use to train your machine learning models have a huge influence on the performance you can achieve.
- Irrelevant or partially relevant features can **negatively** impact model performance.
- Feature Selection is the process where you **automatically** or manually **select** those features which **contribute most** to your prediction variable or output in which you are interested in.



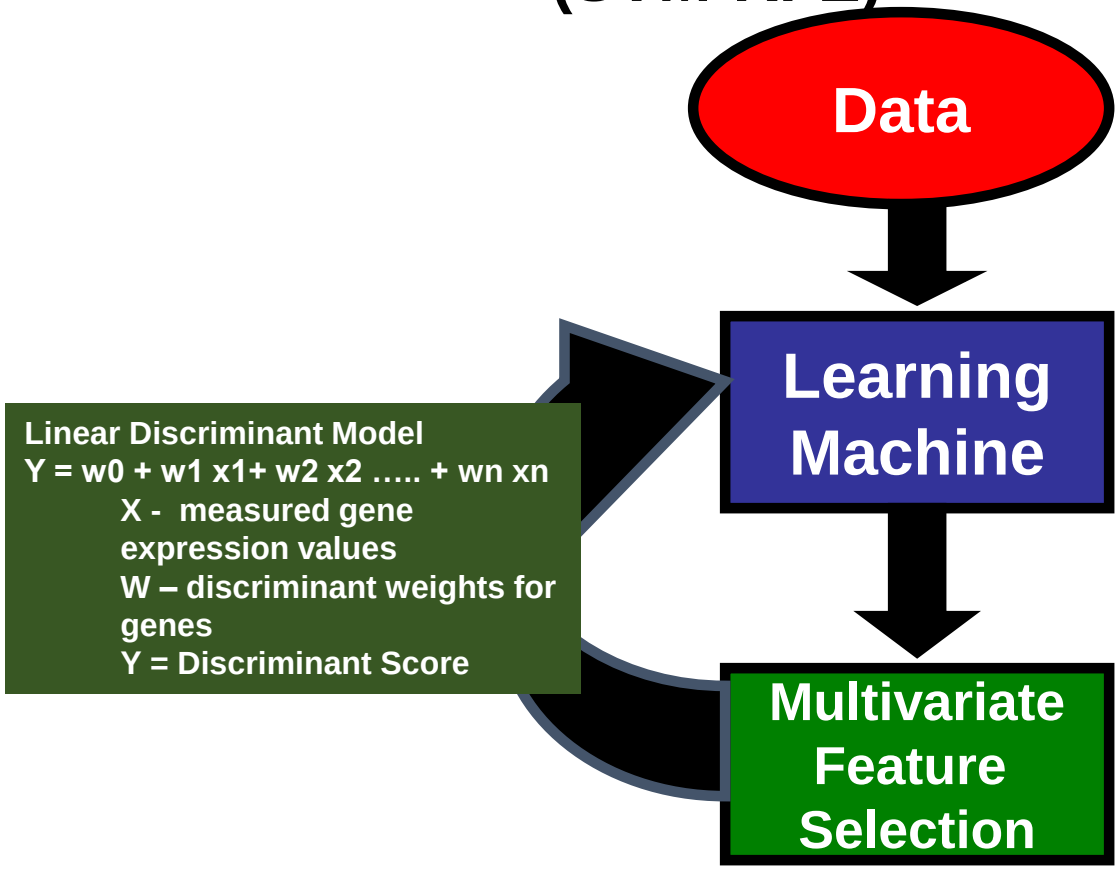
Recursive Feature Elimination

Iteratively discards the least contributing genes

Univariate Filter Approach



Multivariate Wrapper Approach (SVM-RFE)

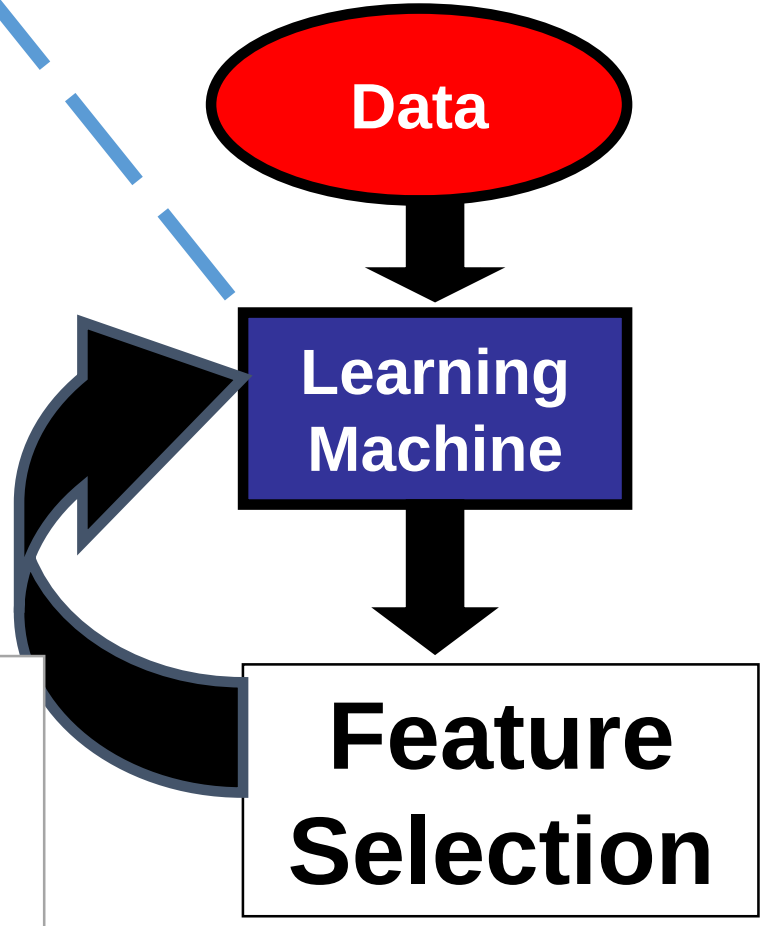


Recursive Feature Elimination (RFE)

Features with low value of weight W_i are weak features. $\mathbf{W_i}$ are ranks

$$F(X) = w_0 + w_1 x_1 + \ldots + w_n x_n$$

Recursive feature elimination (RFE) is a feature selection method that **fits a model** and **removes the weakest feature** (or features) until the specified number of features is reached.



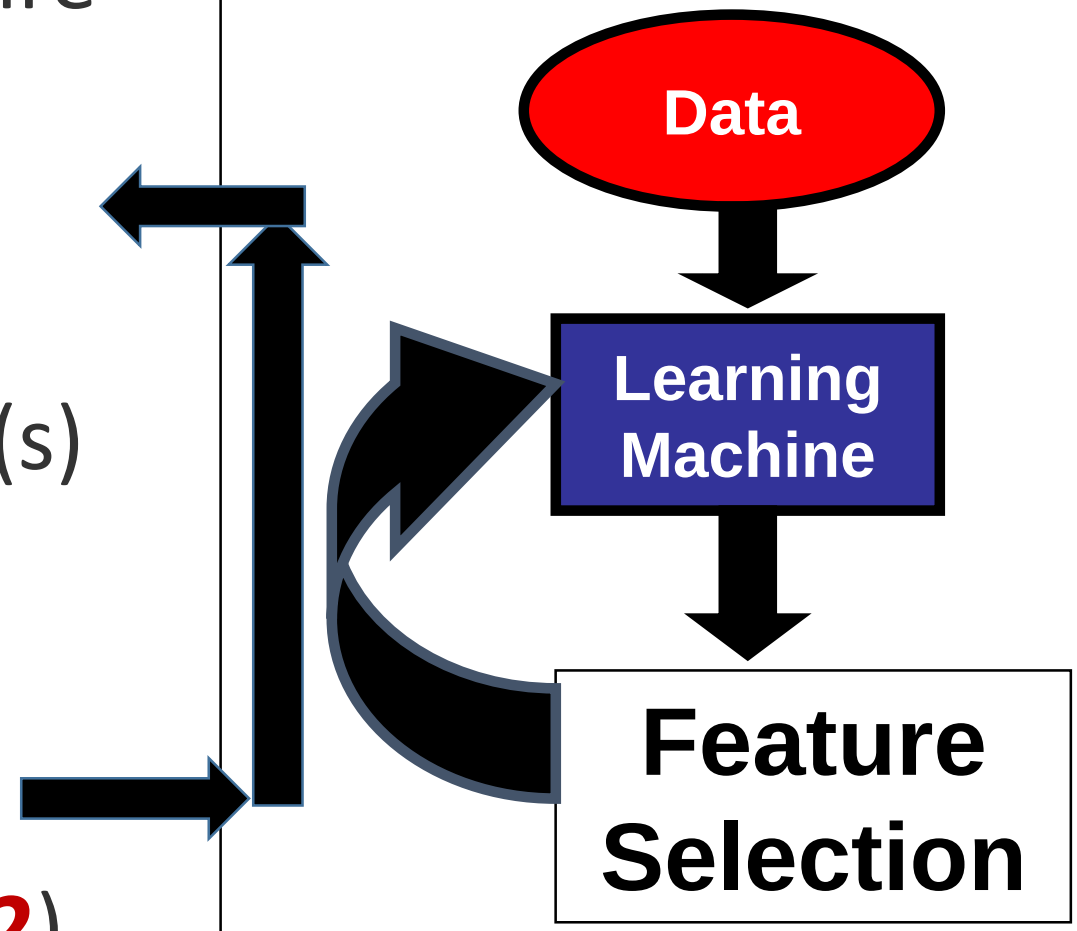
Isabelle Guyon

Gene Selection for Cancer Classification using Support Vector Machines, Machine Learning volume 46, pages389–422(2002)

Recursive Feature Elimination

1. This technique begins by building a model on the entire set of features.
2. Computing an "**importance score**" for each feature.
3. The least important feature(s) are then **removed**.
4. The model is **re-built**, and importance scores are computed again(**Go to step 2**)

$$F(X) = w_0 + w_1 x_1 + \ldots + w_n x_n$$



The Generic Approach

Data X = Two class gene expression data with *n* genes $g_1, g_2, \ldots, g_n$

G()-Grouping Function

G()- Grouping Component:
Create a list of groups that each group contains its set of genes
$G(g_1, g_2, \ldots, g_n) = \{grp_1, grp_2, \ldots, grp_m\}$

R()- Ranking Component :
Ranking each group grp_i by their ability of separating the two classes.
Ranking Function: Use machine learning algorithm with cross validation repeated *r* times. The rank is the average value

Training data

Build classifier from the top *j* groups

Applying the model

Testing data

For example, Mootha et al. (2003) showed that the oxidative phosphorylation genes are significantly associated with **diabetes as a group**, while none of the genes in the pathway showed significant **change individually**

Integrating Biological Knowledge with Gene Expression Profiles for Survival Prediction of Cancer
Xi Chen and Lily Wang

Create Sub Data for each Group

S miRNA	S Concatenate(Target Gene)
ISA-LET-7A-3P	CCND1, CCND2, E2F2, CCND2, E2F2, CCND1
ISA-LET-7A-5P	CDK6, NKIRAS2, ITGB3, NF2, NRAS, KRAS, KRAS, PRDM1, PRDM1, FOXA1, NR1I2, VDR, RAVER2,
ISA-LET-7B-3P	CASP3, CASP3, CASP3
ISA-LET-7B-5P	CDC34, RDH10, IGF2BP1, HMGA1, MTPN, RPIA, ACTG1, HMGA2, HMGA2, HMGA2, CDC25A, CDK6
ISA-LET-7C-3P	MTPN, MTPN
ISA-LET-7C-5P	CDC25A, TGFBR1, HMGA2, HMGA2, HMGA2, MYC, BCL2L1, BCL2L1, MPL, AGO1, IGF1R, ITGB3, IL
ISA-LET-7D-5P	HMGA2, HMGA2, HMGA2, APP, DICER1, SLC11A2, IL13, MPL, AGO1, HMGA2, DICER1, SLC11A2, A
ISA-LET-7E-3P	COPS8, GPS1, COPS6, COPS8, GPS1, COPS6
ISA-LET-7E-5P	HMGA2, SMC1A, WNT1, CCND1, MPL, MYCN, MYCN, AGO1, IGF1R, MMP9, IGF1, LIN28A, ARID3A,
ISA-LET-7F-5P	KLK10, KLK10, KLK6, PRDM1, IL13, MPL, CYP19A1, CCND1, MYH9, SOCS3, ELF4, DYRK2, CCL7, AG
ISA-LET-7G-3P	CCL2, CCL5, CCL2, CCL5, CCL2, CCL5

Sub data represent genes belongs to HAS-LET-7A-5P

Row ID	S class	D RRM2#1	D NRAS#1	D HAS2#1	D UHRF1#1
GSM97926	pos	0.248	0.148	0.142	0.188
GSM97806	pos	0.367	0.763	0.273	0.328
GSM97925	pos	0.103	0.48	0.443	0.115
GSM97885	pos	0.557	0.476	0.339	0.327
GSM97950	pos	0.096	0.439	0.108	0.262
GSM97902	pos	0.081	0.113	0.054	0.022
GSM97877	pos	0.188	0.193	0.121	0.073
GSM97960	pos	0.066	0.141	0.034	0.25
GSM97841	pos	0.02	0.086	0.033	0.032
GSM97862	pos	0.072	0.033	0.066	0.319
GSM97935	pos	0.624	0.392	0.253	0.53
GSM97905	pos	0.208	0.412	0.231	0.312
GSM97864	pos	0.003	0.062	0.003	0.013
GSM97837	pos	0.048	0.401	0.242	0.109
GSM97822	pos	0.852	0.741	0.656	0.374
GSM97940	pos	0.71	0.833	0.181	0.171
GSM97844	pos	0.004	0.15	0.042	0.022
GSM97874	pos	0.098	0.358	0.199	0.16
GSM97883	pos	0.349	0.265	0.153	0.202
GSM97924	pos	0.01	0.191	0.04	0.029
GSM97903	pos	0.375	0.173	0.245	0.306
GSM97923	pos	0.175	0.613	0.384	0.212
GSM97835	pos	0.042	0.075	0.032	0.081

Sub data represent genes belongs to HAS-MIR-103A-3P

Row ID	S class	D CAV1#19	D CDK2#19
GSM97860	pos	0.035	0.137
GSM97962	pos	0.03	0.092
GSM97836	pos	0.103	0.269
GSM97931	pos	0.118	0.138
GSM97858	pos	0.04	0.047
GSM97854	pos	0.024	0.034
GSM97818	pos	0.074	0.22
GSM97947	pos	0.064	0.177
GSM97852	pos	0.18	0.172
GSM97796	pos	0.165	0.348
GSM97845	pos	0.164	0.188
GSM97829	pos	0.047	0.157
GSM97800	neg	0.035	0.006
GSM97803	neg	0.016	0.023
GSM97805	neg	0.065	0.052
GSM97807	neg	0.031	0.032

Sub data represent genes belongs to a specific miRNA

S class	D AKT1#...	D CXCR4...	D PTBP1...	D SIRT1#...
pos	0.045	0.406	0.431	0.185
pos	0.025	0.072	0.075	0.508
pos	0.054	0.176	0.273	0.558
pos	0.027	0.079	0.334	0.529
pos	0.049	0.513	0.514	0.363
pos	0.032	0.406	0.157	0.611
pos	0.027	0.114	0.082	0.477
pos	0.019	0	0.065	0.279
pos	0.037	0.405	0.474	0.301
pos	0.049	0.058	0.377	0.609
pos	0.027	0.238	0.443	0.194
pos	0.026	0.798	0.52	0.139
pos	0.045	0.203	0.358	0.157
pos	0.033	0.519	0.412	0.328
neg	0.023	0.05	0.057	0.448
neg	0.015	0.049	0.067	0.54
neg	0.017	0.067	0.08	0.518
neg	0.015	0.051	0.09	0.418
neg	0.019	0.074	0.091	0.44
neg	0.015	0.08	0.07	0.466
neg	0.022	0.047	0.051	0.46
neg	0.018	0.066	0.194	0.55
neg	0.018	0.076	0.036	0.433
neg	0.018	0.051	0.045	0.47
neg	0.025	0.07	0.022	0.499
neg	0.023	0.051	0.041	0.369
neg	0.015	0.107	0.194	0.61
neg	0.027	0.056	0.048	0.473
neg	0.02	0.041	0.037	0.355
neg	0.028	0.164	0.182	0.44
neg	0.02	0.041	0.009	0.429
neg	0.021	0.058	0.031	0.464
neg	0.021	0.026	0.069	0.59

Rank Each Group=Sub Data

Groups=Sub Datasets

grp_1

Row ID	class	RRM2#1	NRAS#1	HAS2#1	UHRF1#1
GSM97926	pos	0.248	0.148	0.142	0.188
GSM97806	pos	0.367	0.763	0.273	0.328
GSM97925	pos	0.103	0.48	0.443	0.115
GSM97885	pos	0.557	0.476	0.339	0.327
GSM97950	pos	0.096	0.439	0.108	0.262
GSM97902	pos	0.081	0.113	0.054	0.022
GSM97877	pos	0.188	0.193	0.121	0.073
GSM97960	pos	0.066	0.141	0.034	0.25
GSM97841	pos	0.02	0.086	0.033	0.032
GSM97862	pos	0.072	0.033	0.066	0.319
GSM97935	pos	0.624	0.392	0.253	0.53
GSM97905	pos	0.208	0.412	0.231	0.312
GSM97864	pos	0.003	0.062	0.003	0.013
GSM97837	pos	0.048	0.401	0.242	0.109
GSM97822	pos	0.852	0.741	0.656	0.374
GSM97940	pos	0.71	0.833	0.181	0.171
GSM97844	pos	0.004	0.15	0.042	0.022
GSM97874	pos	0.098	0.358	0.199	0.16
GSM97883	pos	0.349	0.265	0.153	0.202
GSM97924	pos	0.01	0.191	0.04	0.029
GSM97903	pos	0.375	0.173	0.245	0.306
GSM97923	pos	0.175	0.613	0.384	0.212
GSM97835	pos	0.042	0.075	0.032	0.081

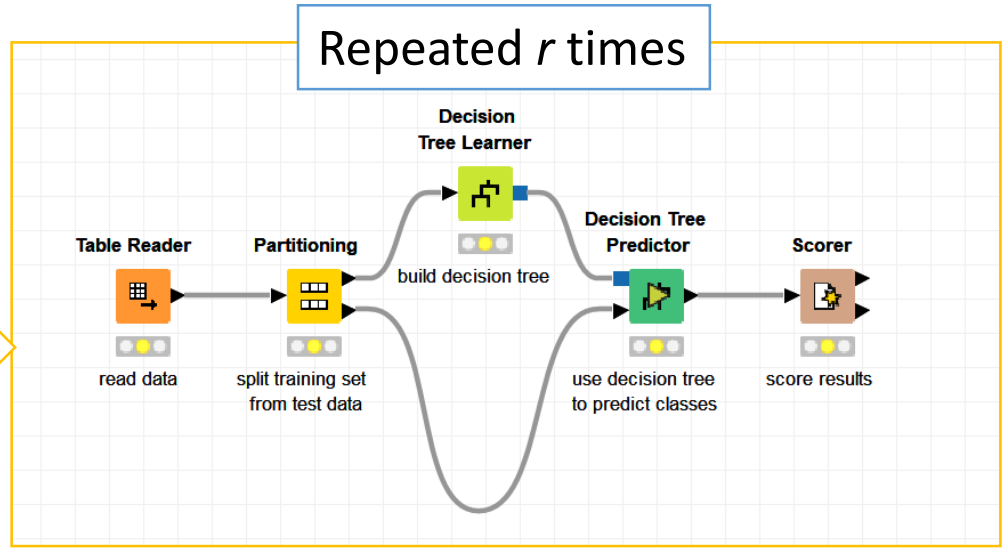
grp_2

Row ID	class	CAV1#19	CDK2#19
GSM97860	pos	0.035	0.157
GSM97962	pos	0.03	0.092
GSM97836	pos	0.103	0.269
GSM97931	pos	0.118	0.138
GSM97858	pos	0.04	0.047
GSM97854	pos	0.024	0.034
GSM97818	pos	0.074	0.22
GSM97947	pos	0.064	0.177
GSM97852	pos	0.18	0.172
GSM97796	pos	0.165	0.348
GSM97845	pos	0.164	0.188
GSM97829	pos	0.047	0.157
GSM97800	neg	0.035	0.006
GSM97803	neg	0.016	0.023
GSM97805	neg	0.065	0.052
GSM97807	neg	0.031	0.032

grp_m

class	AKT1#...	CXCR4...	PTBP1...	SIRT1#...
pos	0.045	0.406	0.431	0.185
pos	0.025	0.072	0.075	0.508
pos	0.054	0.176	0.273	0.558
pos	0.027	0.079	0.334	0.529
pos	0.049	0.513	0.514	0.363
pos	0.032	0.406	0.157	0.611
pos	0.027	0.114	0.082	0.477
pos	0.019	0	0.065	0.279
pos	0.037	0.405	0.474	0.301
pos	0.049	0.058	0.377	0.609
pos	0.027	0.238	0.443	0.194
pos	0.026	0.798	0.52	0.139
pos	0.045	0.203	0.358	0.157
pos	0.033	0.519	0.412	0.328
neg	0.023	0.05	0.057	0.448
neg	0.015	0.049	0.067	0.54
neg	0.017	0.067	0.08	0.518
neg	0.015	0.051	0.09	0.418
neg	0.019	0.074	0.091	0.44
neg	0.015	0.08	0.07	0.466
neg	0.022	0.047	0.051	0.46
neg	0.018	0.066	0.194	0.55
neg	0.018	0.076	0.036	0.433
neg	0.018	0.051	0.045	0.47
neg	0.025	0.07	0.022	0.499
neg	0.023	0.051	0.041	0.369
neg	0.015	0.107	0.194	0.61
neg	0.027	0.056	0.048	0.473
neg	0.02	0.041	0.037	0.355
neg	0.028	0.164	0.182	0.44
neg	0.02	0.041	0.009	0.429
neg	0.021	0.058	0.031	0.464
neg	0.021	0.026	0.069	0.59

Repeated *r* times



Ranks or score each group individually ?

$$W_1\,Grp_1\;W_2\,grp_2\;\ldots\;W_m\,grp_m$$

$w_0, w_1, \ldots, w_m$ are the score for each **group**

We consider all genes together

$$F(X)=w_0+w_1x_1+\ldots+w_nx_n$$

$w_0, w_1, \ldots, w_n$ are the score for each feature/gene

The Generic Approach

Ranking Algorithm - $R(X_s, M, f, r)$

$\mathbf{X_s}$: any subset of the input gene expression data X, the features are gene expression values

M $\{m_1, m_2, \ldots, m_p\}$ is a list of groups produced by k-means.

f *is a scalar ( )*: split into train and test data

r: repeated times (iteration)

res={} for aggregation the scores for each m_i

Generate Rank for each m_i, *Rank(m_i):*

For each m_i in M

sm_i=0;

Perform *r* times (here r=5) steps 1-5:

1. Perform stratified random sampling to split X_s into train X_t and test X_v data sets according to *f* (here 80:20)
2. **Remove all genes (features) from X_t and X_v which are not in the group m_i**
3. Train classifier on X_t using SVM
4. *t* = Test classifier on X_v –calculate performance
5. $sm_i = sm_i + t$;

Score(m_i)= sm_i */r* ; Aggregate performance

res= $\bigcup_{i=1}^{p} Score(m_i)$

Output

Return res (res = {Rank(m_1),Rank(m_2),…,Rank(m_p)})

The data represented Just by the genes belongs to the specific group

Score/Rank Each Group=Sub Data

$W_1\ Grp_1\ W_2\ grp_2 \ldots\ W_m\ grp_m$

?

$w_0, w_1, \ldots, w_m$ are the score for each **group**

Ranks or score each group individually!

In this way, we do not consider the relationship between the groups that could be important in term of computational and **biological perspectives**

We consider all genes together

$$F(X) = w_0 + w_1 x_1 + \ldots + w_n x_n$$

$w_0, w_1, \ldots, w_n$ are the score for each feature/gene

To find the optimal solution one needs to consider all combination of two or three groups (or even more), which is a computationally intensive approach

There are two steps for building prediction models based on pathways: (1) for each gene set, select genes associated with outcome and **summarize information from these selected genes by estimating the underlying latent variable, which are the "super genes;"** and (2) construct prediction model using relevant super genes as predictors. In this paper, we study pathway-based versus gene-based prediction models using Supervised PCA (Bair and Tibshirani, 2004; Bair et al., 2006) and Lasso because of their simplicity and popularity, but the proposed strategy can be easily adapted to other prediction models as well.

Rank Groups Simultaneously

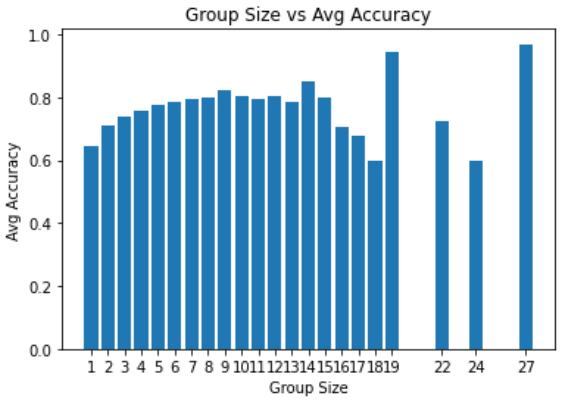
Let $c_1, c_2, \ldots, c_k$ groups

1. For each c_i choose a *representative gene*
2. Apply Linear SVM on the ***representative*** genes
3. Assign a score to each group as the absolute value of the correspond *representative*

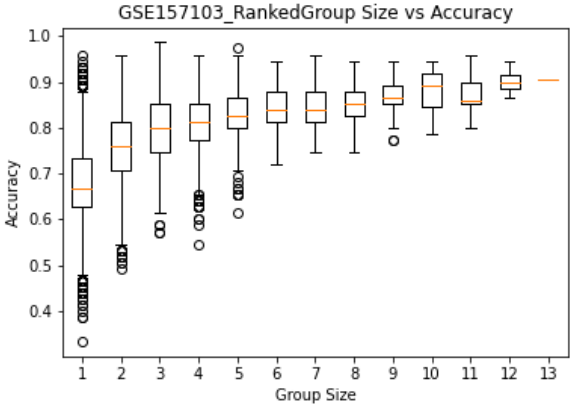
Chen X, Wang L. Integrating biological knowledge with gene expression profiles for survival prediction of cancer. *J Comput Biol*. 2009;16(2):265-278. doi:10.1089/cmb.2008.12TT

The Score/Rank for each Group

Group	Mean(Accuracy)	Mean(Sensitivity)	Mean(Specifity)	Mean(Recall)	Mean(Precision)	Mean(F-measure)	Mean(Cohen's kappa)
HSA-LET-7A-5P	0.88	0.87	0.90	0.87	0.96	0.90	0.74
HSA-LET-7B-5P	0.54	0.58	0.45	0.58	0.69	0.62	0.04
HSA-LET-7C-5P	0.57	0.64	0.40	0.64	0.70	0.67	0.05
HSA-MIR-1-3P	0.91	0.93	0.85	0.93	0.93	0.93	0.79
HSA-MIR-1-5P	0.69	0.76	0.55	0.76	0.79	0.77	0.29
HSA-MIR-100-5P	0.74	0.78	0.65	0.78	0.84	0.80	0.40
HSA-MIR-101-3P	0.92	0.96	0.85	0.96	0.94	0.94	0.82
HSA-MIR-101-5P	0.92	0.93	0.90	0.93	0.96	0.94	0.82
HSA-MIR-103A-3P	0.75	0.76	0.75	0.76	0.88	0.81	0.46
HSA-MIR-105-5P	0.91	0.98	0.75	0.98	0.90	0.94	0.77
HSA-MIR-106A-5P	0.86	0.87	0.85	0.87	0.93	0.90	0.69
HSA-MIR-106B-3P	0.75	0.78	0.70	0.78	0.85	0.81	0.45
HSA-MIR-106B-5P	0.89	0.96	0.75	0.96	0.91	0.93	0.73
HSA-MIR-107	0.71	0.76	0.60	0.76	0.83	0.78	0.32
HSA-MIR-10B-3P	0.88	0.93	0.75	0.93	0.90	0.91	0.70
HSA-MIR-10B-5P	0.82	0.87	0.70	0.87	0.87	0.87	0.57
HSA-MIR-1207-5P	0.66	0.76	0.45	0.76	0.76	0.75	0.21
HSA-MIR-122-5P	0.54	0.67	0.25	0.67	0.67	0.66	-0.09
HSA-MIR-1236-3P	0.75	0.80	0.65	0.80	0.84	0.82	0.43
HSA-MIR-124-3P	0.92	0.91	0.95	0.91	0.98	0.94	0.83
HSA-MIR-1245A	0.71	0.71	0.70	0.71	0.84	0.75	0.39
HSA-MIR-1247-5P	0.89	0.91	0.85	0.91	0.94	0.92	0.76
HSA-MIR-125A-3P	0.75	0.82	0.60	0.82	0.82	0.82	0.43
HSA-MIR-125A-5P	0.88	0.87	0.90	0.87	0.95	0.91	0.73
HSA-MIR-125B-5P	0.85	0.82	0.90	0.82	0.96	0.88	0.67
HSA-MIR-126-3P	0.91	0.93	0.85	0.93	0.94	0.93	0.79
HSA-MIR-126-5P	0.78	0.80	0.75	0.80	0.88	0.83	0.53
HSA-MIR-127-3P	0.75	0.78	0.70	0.78	0.86	0.81	0.45
HSA-MIR-128-3P	0.89	0.91	0.85	0.91	0.94	0.92	0.75



Sena B. Yengec-Tasdemir
AGU University



Burcu Bakir-Gungor
Assisitan Professot
Department of Computer Engineering
Abdullah Gul University

Jens Allmer
Professor for Medical Informatics and Bioinformatics at Hochschule Ruhr West
University of Applied Sciences

Different methods for Performing Biological Grouping

Group genes based on **biological knowledge**. The genes that have similar functions are correlated. Use clustering methods such as K-Means.

SVM-RCE
SVM-RCE-R
SVM-RNE

Group genes based on other Biological Database such as KEGG-PATHWASY, Gene Ontology , mirTarBase(microRNA) , Diseases,…

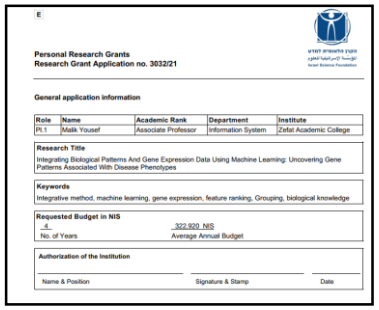
maTE
CogNet
PredDisGeNetML
GeNetOntology
GeNetKEGG
miRNADisNet

MicroBiomeNet

Group genes based on two or more omics data such as mRNA and miRNA expressions

miRcorrnet
miRModuleNet

Our Biological Integrative Approach


Personal Research Grants
Research Grant Application no. 3032/21

General application information

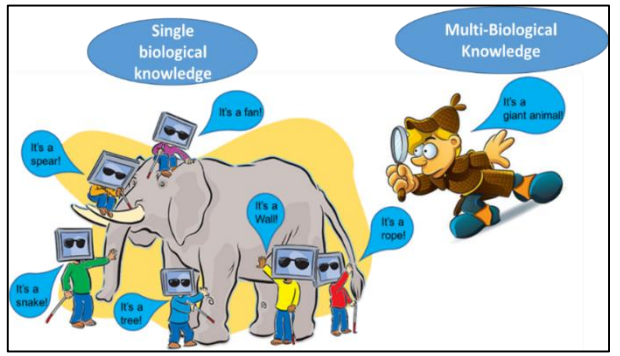
Role	Name	Academic Rank	Department	Institute
PI.1	Malik Yousef	Associate Professor	Information System	Zefat Academic College

Research Title
Integrating Biological Patterns And Gene Expression Data Using Machine Learning: Uncovering Gene Patterns Associated With Disease Phenotypes

Keywords
Integrative method, machine learning, gene expression, feature ranking, Grouping, biological knowledge

Requested Budget in NIS
4 No. of Years
322,920 NIS Average Annual Budget

Authorization of the Institution
Name & Position Signature & Stamp Date



Samples (ℓ)

Genes (*n*)

miR TarBase | KEGG | Disease | SNPs | Methylation | User list

bioG() space

Split the data

test

train

t-test
|gene| ≤ 2000

Filter by t-test

Train and test report ranks

r iterations

bioR(): Rank each group

Select top j groups

Train on top j groups
Test on the test set

N-fold cross validation

Ranked groups/genes

Performance

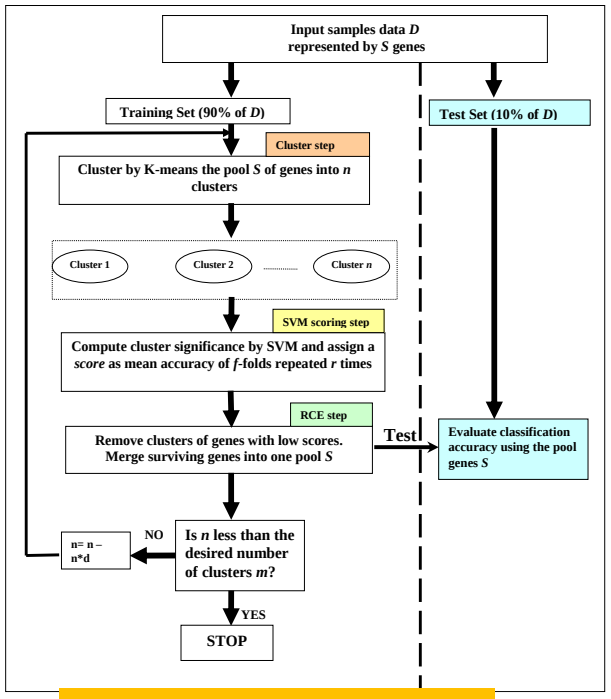
Aggregations of ranked groups/genes

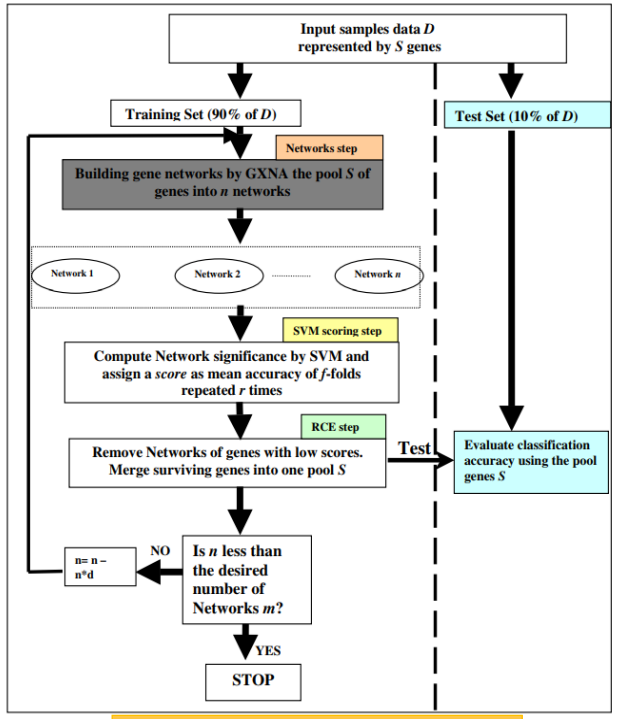
Performance Aggregations

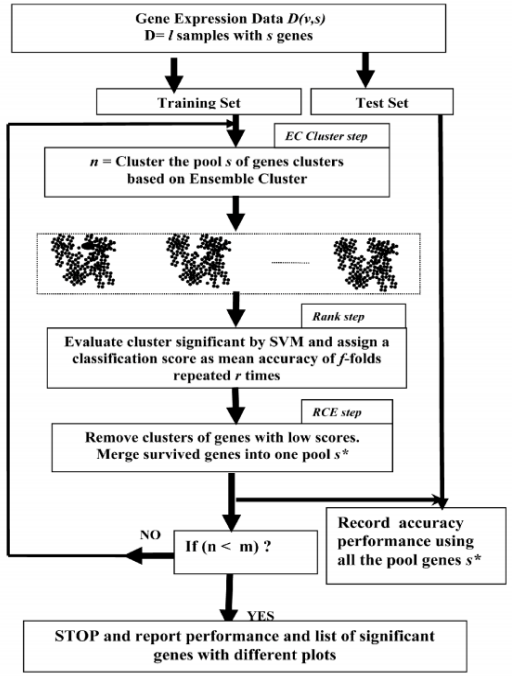
Ranked Significant Biological Groups and Genes

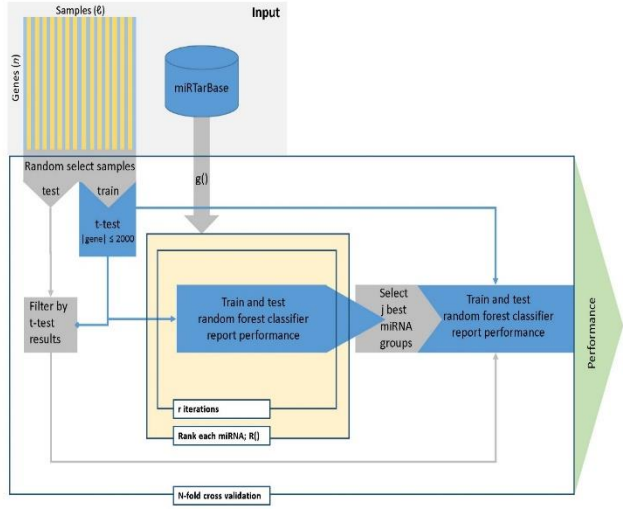
Summary Table over top *j* groups

Domain Knowledge Based Feature Selection









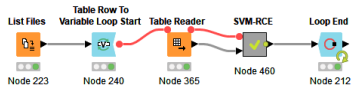
SVM-RCE 2007

SVM-RNE 2009

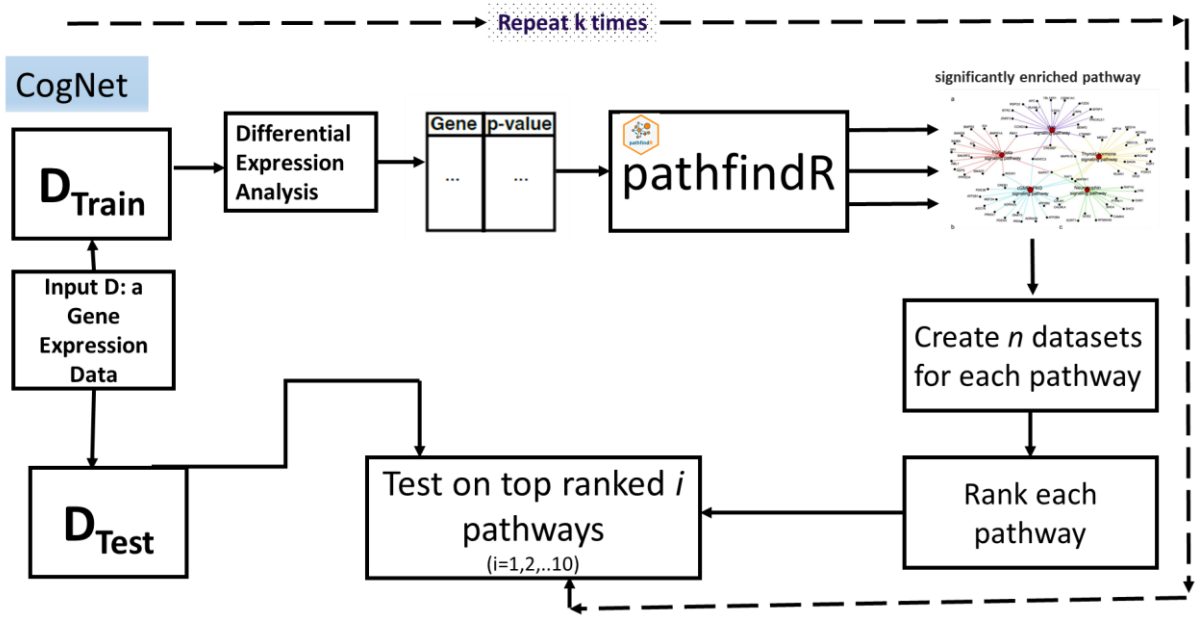
RCE-EC 2017

maTE 2019

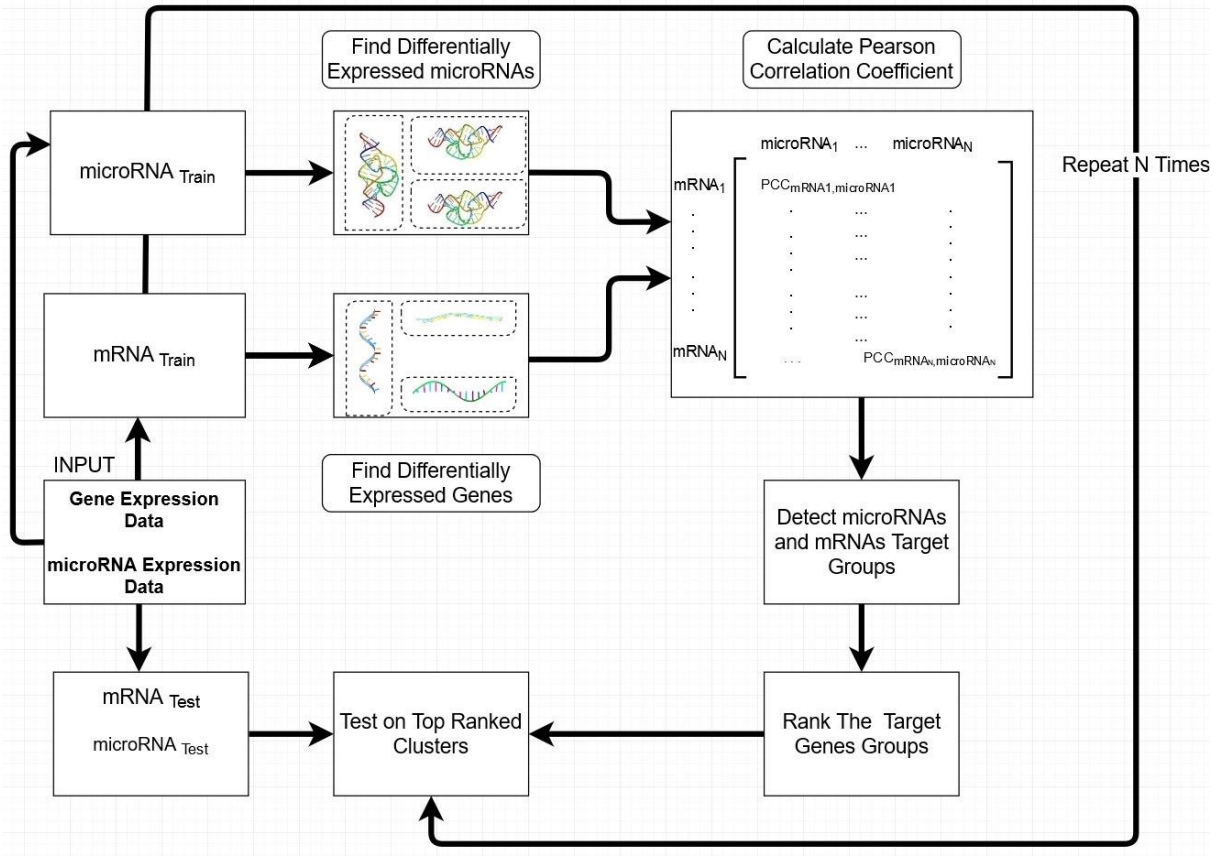
Update version SVM-RCE-R based Ranks Functio 2020



Domain Knowledge Based Feature Selection

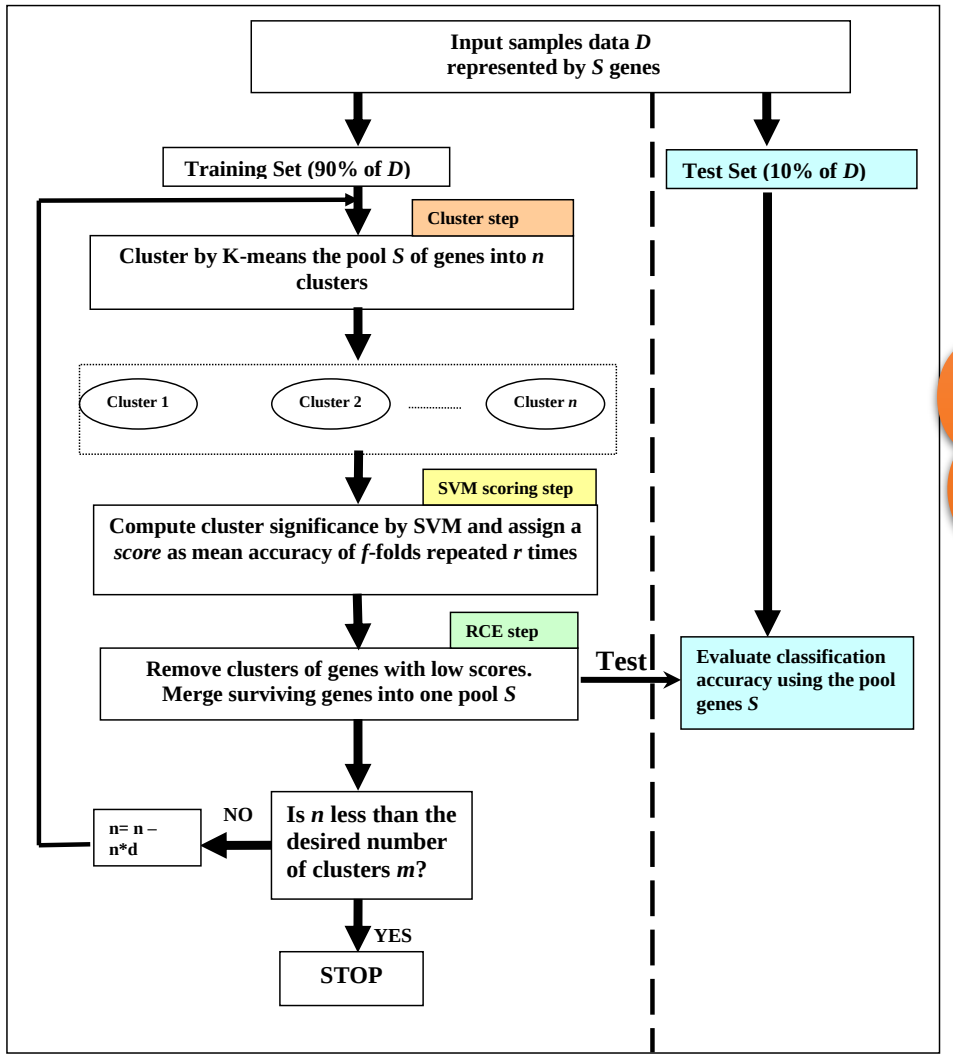


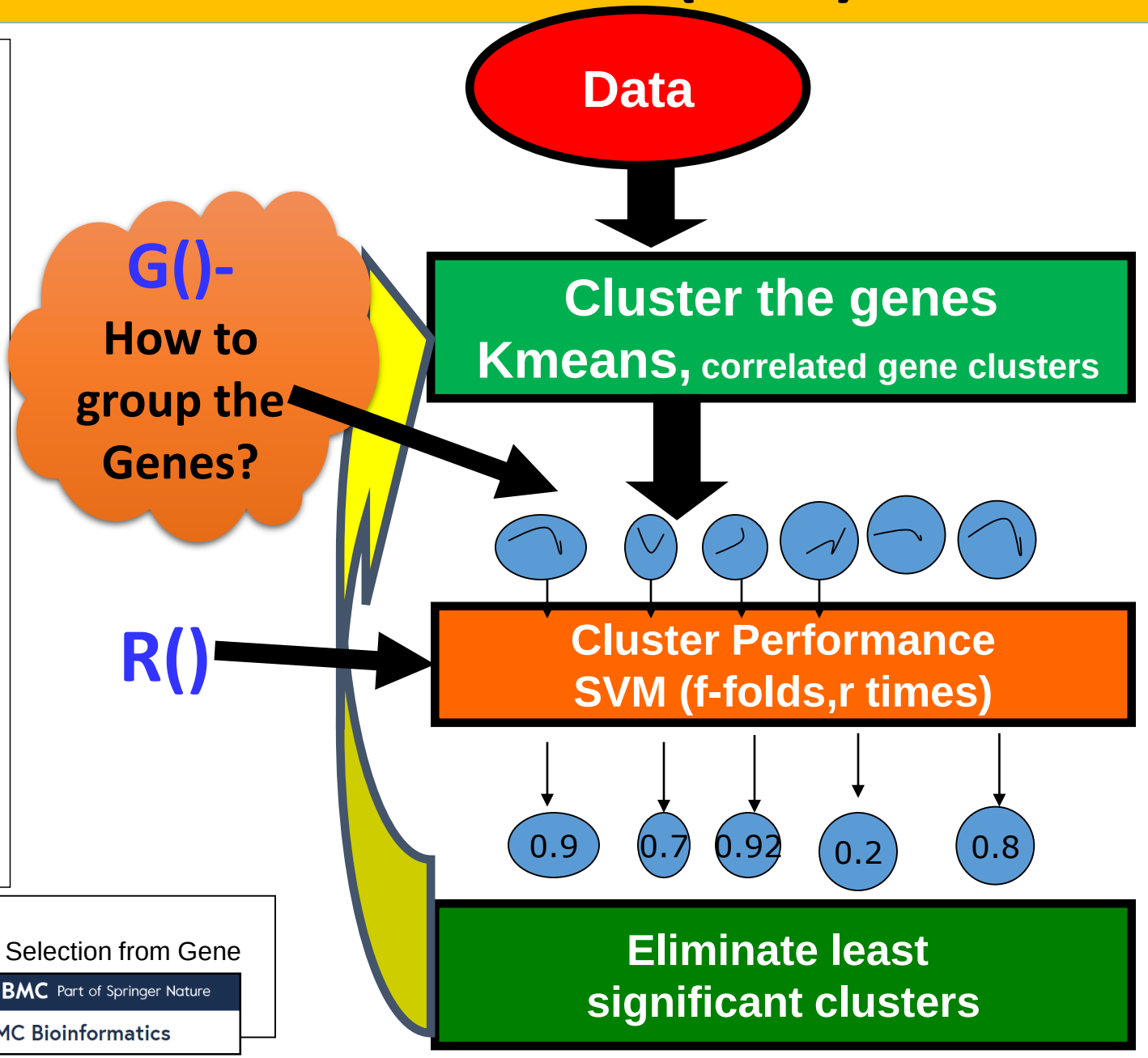
CogNet 2020



miRcorrNet 2020

Recursive Cluster Elimination (RCE)





Malik Yousef, Louise C. Showe and Michael K. Showe,
Recursive Cluster Elimination (RCE) for Classification and Feature Selection from Gene Expression Data.
BMC Bioinformatics 8:144 (2007) [Highly Accessed]

BMC Part of Springer Nature
BMC Bioinformatics

Output of SVM-RCE

#Clusters	#Genes (Mean)	Accuracy (Mean)	Sensitivity (Mean)	Specifity (Mean)	Recall (Mean)	Precision (Mean)	F-measure (Mean)	Area Under Curve (Mean)	Cohen's kappa (Mean)
90	978.5	0.94	0.94	0.95	0.94	0.98	0.96	0.98	0.87
72	955.8	0.94	0.94	0.95	0.94	0.98	0.96	0.98	0.87
54	929.7	0.94	0.94	0.95	0.94	0.98	0.96	0.98	0.87
35	910.7	0.94	0.94	0.95	0.94	0.98	0.96	0.98	0.87
18	857.3	0.94	0.94	0.95	0.94	0.98	0.96	0.98	0.87
13	814.7	0.94	0.94	0.95	0.94	0.98	0.96	0.97	0.87
12	775.6	0.94	0.94	0.95	0.94	0.98	0.96	0.97	0.87
11	739.3	0.94	0.93	0.94	0.93	0.98	0.95	0.97	0.85
10	707.5	0.95	0.95	0.95	0.95	0.98	0.96	0.98	0.88
9	662.5	0.94	0.94	0.95	0.94	0.98	0.96	0.98	0.87
8	619.7	0.94	0.94	0.95	0.94	0.98	0.96	0.98	0.87
7	518.0	0.94	0.93	0.94	0.93	0.98	0.95	0.98	0.85
6	483.0	0.95	0.95	0.95	0.95	0.98	0.96	0.98	0.88
5	389.8	0.94	0.94	0.95	0.94	0.98	0.96	0.98	0.87
4	326.2	0.94	0.94	0.95	0.94	0.98	0.96	0.98	0.87
3	254.4	0.94	0.94	0.95	0.94	0.98	0.96	0.98	0.87
2	149.3	0.94	0.94	0.95	0.94	0.98	0.96	0.98	0.87
1	80.0	0.94	0.94	0.95	0.94	0.98	0.96	0.96	0.87

Gene	#Frequence	Score
THPO	787	4.14
ZBTB33	787	4.14
C1orf68	779	4.10
221116_at	749	3.94
LOC100130331	744	3.92
DPY19L1P1	729	3.84
LAMA4	724	3.81
LOC220077	716	3.77
TBX10	711	3.74
C10orf10	709	3.73
KLHL4	704	3.71
SLC16A6	698	3.67
RIPK4	696	3.66
TIMP4	693	3.65
PPP1R3A	692	3.64
FAIM2	691	3.64
MSC	686	3.61
MUC4	685	3.61
RGS1	682	3.59
ZNF343	678	3.57
KIAA1661	674	3.55
ZNF718	671	3.53
PLA2G5	664	3.49
EZR	663	3.49
SLC7A8	663	3.49
ZBTB14	661	3.48
RASAL1	655	3.45
FBLN1	647	3.41

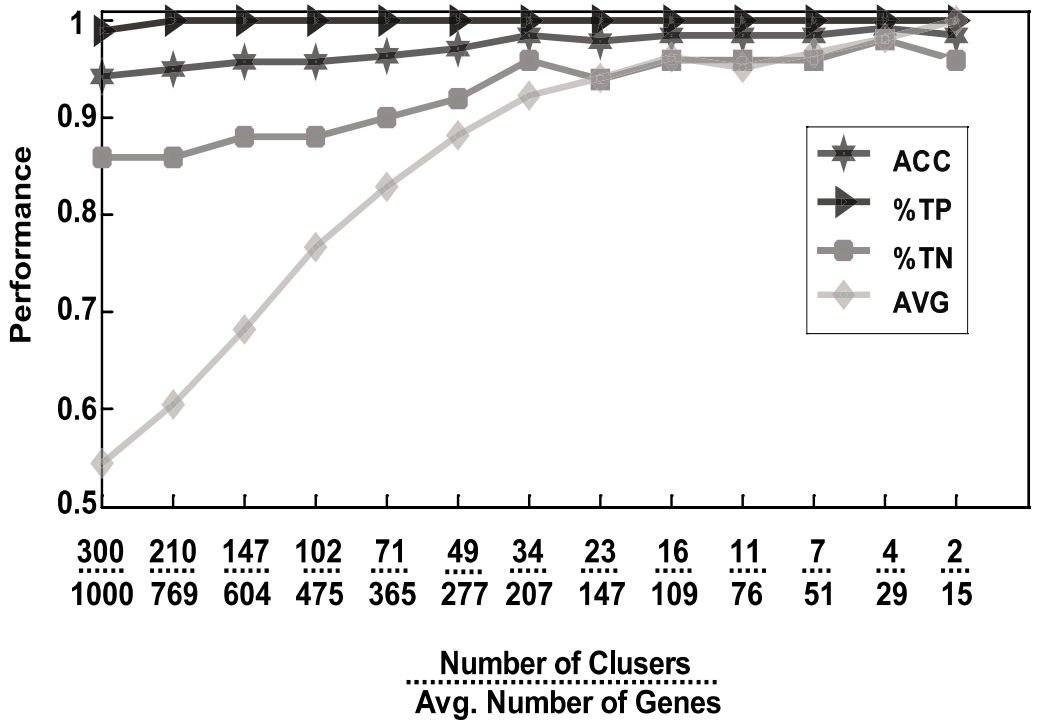
Comparison Results: SVM-RCE

Table 1: Summary results for the SVM-RCE, SVM-RFE and PDA-RFE method. Summary results for the SVM-RCE, SVM-RFE and PDA-RFE method applied on 6 public datasets. #c field is the number of clusters for the SVM-RCE method. The #g field is the number of genes in the associated #c clusters for SVM-RCE, while for the SVM-RFE and PDA-RFE indicates the number of genes used.

	Leukemia(I)			CTCL(I)			CTCL(II)			Head & Neck vs. Lung tumors (I)			Head & Neck vs. Lung tumors (II)			Prostate		
	#c	#g	ACC	#c	#g	ACC	#c	#g	ACC	#c	#g	ACC	#c	#g	ACC	#c	#g	ACC
SVM-RCE	2	12	99%	2	8	100%	2	8	91%	2	8	100%	2	9	100%	2	8	87%
	3	32	98%	9	32	100%	9	34	96%	8	32	100%	6	32	100%	11	36	95%
	28	100	97%	32	101	100%	28	104	96%	28	103	100%	25	103	100%	32	100	93%
SVM-RFE		11	96%		9	89%		8	84%		8	92%		8	98%		8	93%
		32	96%		32	94%		32	85%		32	90%		32	98%		36	95%
		102	97%		102	100%		102	87%		102	90%		102	98%		102	94%
PDA-RFE		8	96%		8	92%		8	83%		8	89%		8	70%		8	94%
		32	96%		32	92%		33	81%		31	96%		32	98%		32	94%
		104	96%		104	95%		108	79%		109	96%		102	98%		104	90%

Malik Yousef, Louise C. Showe and Michael K. Showe, Recursive Cluster Elimination (RCE) for Classification and Feature Selection from Gene Expression Data. **BMC Bioinformatics** 8:144 (2007) [Highly Accessed]

Comparison Results: SVM-RCE



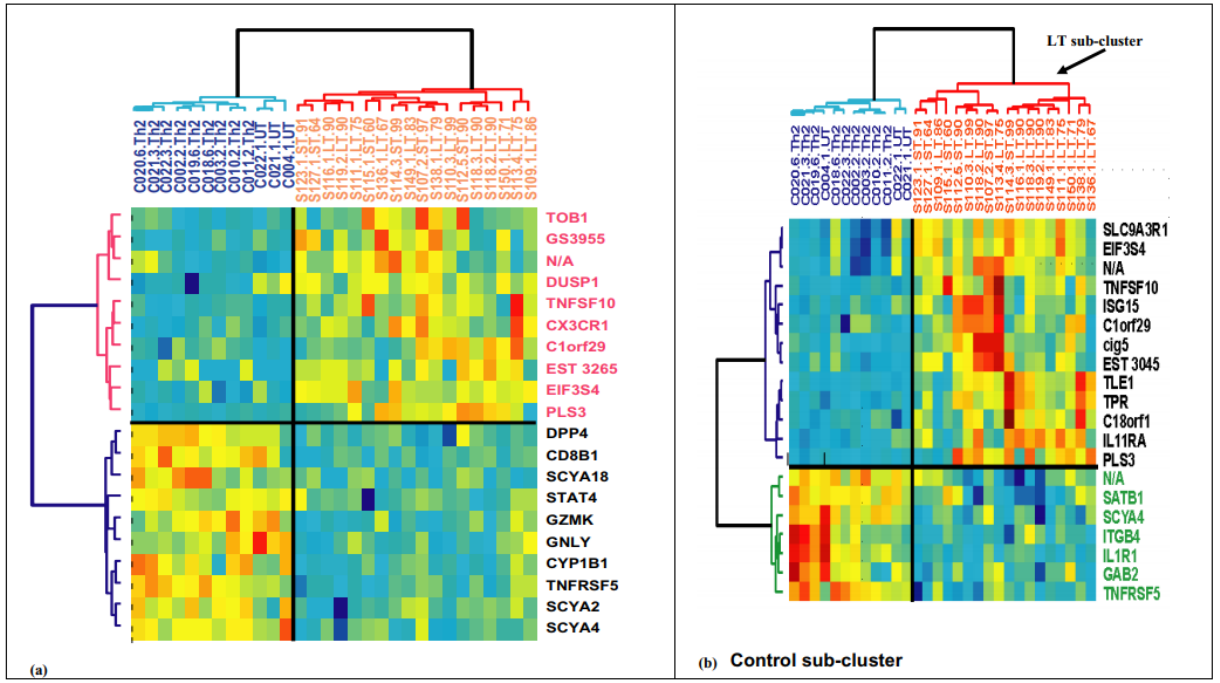


Figure 2
Hierarchal cluster of CTCL(I) on the top 20 genes from SVM-RFE and SVM-RCE. (a) Hierarchal cluster on the top 20 genes from SVM-RFE (b) Hierarchal cluster on the top 20 (~4 clusters) genes from SVM-RCE. Sample names that start with S are CTCL patients, while those that start with C are for controls. LT = long term, ST = short term.

Malik Yousef, Louise C. Showe and Michael K. Showe, Recursive Cluster Elimination (RCE) for Classification and Feature Selection from Gene Expression Data. **BMC Bioinformatics** 8:144 (2007) [Highly Accessed]

SVM-RCE-R - New version of SVM-RCE (2020)

Recursive Cluster Elimination based Rank Function (SVM-RCE-R) Implemented in Knime

F1000Research

Malik Yousef[1], Burcue Bakir, Amhar Jabeer, Rehman Qureshi[2] and Louise C. Showe[2]

$$R(W_1,W_2,W_3,W_4,W_5,W_6) = W_1acc + W_2sen+W_3spe+W_4xfm+W_5auc+W_6pres$$

Where the *acc* is the accuracy, *sen* is the sensitivity, *spe* is the specificity, *fm* is the f-measurement, *auc* is the area under curve and *pres* is precision.

Ranking Algorithm - R(Xs,M,f,r)

X_s: any subset of the input gene expression data X, the features are gene expression values

M {m1,m2,…,mp} is a list of groups produced by k-means.

f is a scalar ( ): split into train and test data

r: repeated times (iteration)

res={} for aggregation the scores for each m_i

Generate Rank for each m_i, Rank(m_i):

For each m_i in M

sm_i=0;

Perform r times (here r=5) steps 1-5:

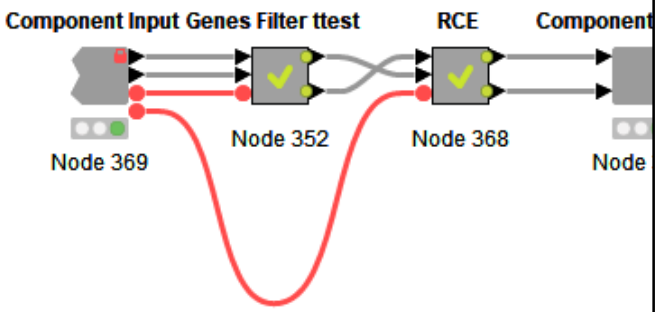
1. Perform stratified random sampling to split X_s into train X_t and test X_v data sets according to f (here 80:20)
2. Remove all genes (features) from X_t and X_v which are not in the group m_i
3. Train classifier on X_t using SVM
4. t = Test classifier on X_v –calculate performance
5. $sm_i = sm_i + t$;

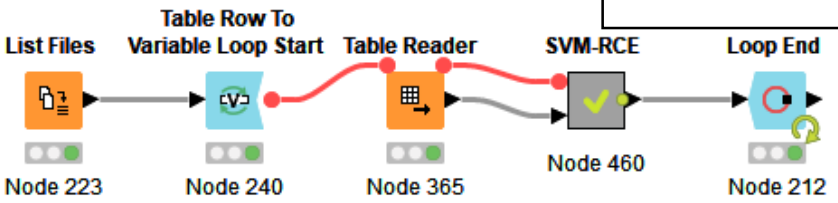
Score(m_i)= sm_i /r ; Aggregate performance
res= i=1pScore(mi)

Output

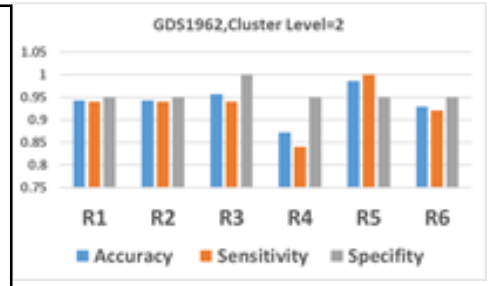
Return res (res = {Rank(m_1),Rank(m_2),…,Rank(m_p)})

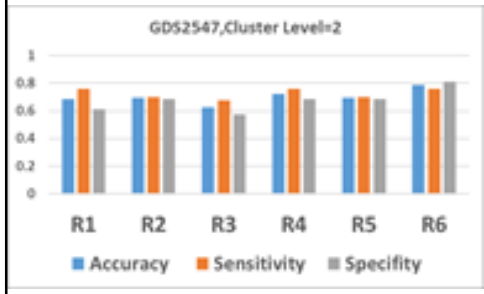
R1	**Acc=0.2 Spe=0.3 Sen= 0.4 Auc=0.1**
R2	Acc=1.0 the rest are zero
R3	Auc= 1.0 the rest are zero
R4	F1=1.0, the rest are zero
R5	Spe=0.2 , Sen=0.8
R6	Spe=0.8 Sen=0.2

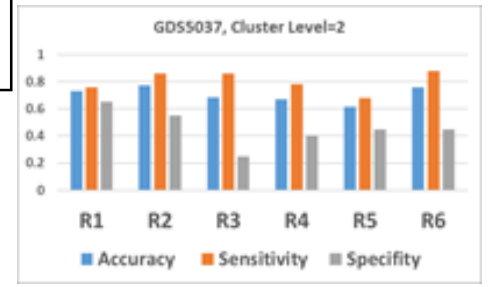




For example, the accuracy obtained with R5 is significantly greater than R4 by about 12% while reaching 4%-6% more than the other ranks. Interestingly we getting a 4% improvement over the standard rank we have been using on the old version of SVM-RCE which was R2.
GDS2547 data reach accuracy of ~79% applying R6 while getting 63% with R3 big difference of about 16%. Same as before a difference of about 9% over the standard rank used on previous version SVM-RCE. However, GDS5037 the max performance obtained with the standard rank R2 reaching a difference of 16% over the minimum values reached by R5.
Is this statistical significance? If so, how is this calculated?
Confidence intervals







SVM-RCE-R - New version of SVM-RCE (2020)

Recursive Cluster Elimination based Rank Function (SVM-RCE-R) Implemented in Knime

Malik Yousef[1], Burcue Bakir, Amhar Jabeer, Qureshi[2] and Louise C. Showe[2]

Ranking Algorithm - R(Xs,M,f,r)

X_s: any subset of the input gene expression data X expression values

M {m1,m2,…,mp} is a list of groups produced by k

f is a scalar ( ): split into train and test data

r: repeated times (iteration)

res={} for aggregation the scores for each m_i

Generate Rank for each m_i, Rank(m_i):

For each m_i in M

sm_i=0;

Perform r times (here r=5) steps 1-5:

1. Perform stratified random sampling t X_v data sets according to f (here 80:20
2. Remove all genes (features) from X_t group m_i
3. Train classifier on X_t using SVM
4. t = Test classifier on X_v –calculate perf
5. $sm_i = sm_i + t$;

Score(m_i)= sm_i /r ; Aggregate performance
res= i=1pScore(mi)

Output

Return res (res = {Rank(m_1),Rank(m_2),…,Rank(m_p)})

Next: SVM-RCE Optimal Rank Parameters

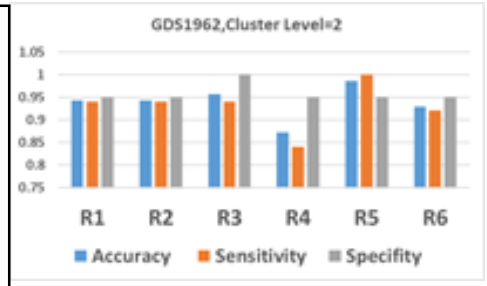
Developing some hyper-parameter optimization methods (such as Bayesian) to select the weights of the rank function.

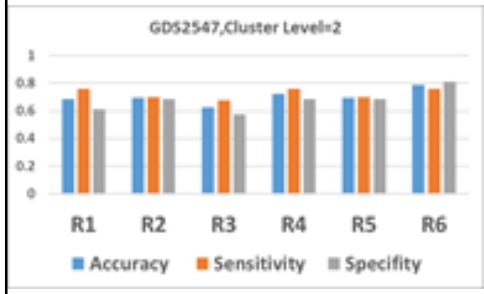
Yousef M., Jabeer A., Bakir-Gungor B. (2021**) SVM-RCE-R-OPT: Optimization of Scoring Function for SVM-RCE-R.** In: Kotsis G. et al. (eds) Database and Expert Systems Applications - DEXA 2021 Workshops. DEXA 2021. Communications in Computer and Information Science, vol 1479. Springer, Cham. https://doi.org/10.1007/978-3-030-87101-7_21

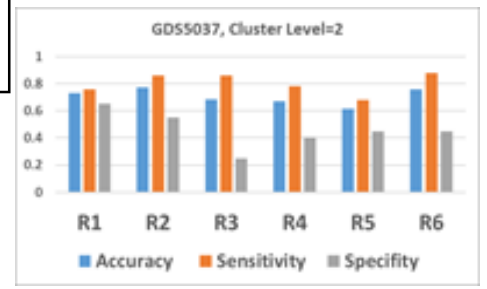
$$R(W_1,W_2,W_3,W_4,W_5,W_6) = W_1acc + W_2sen+W_3spe+W_4xfm+W_5auc+W_6pres$$

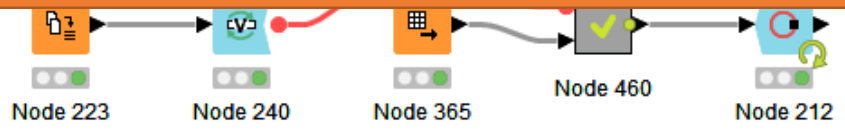
spe is the specificity, *fm* is the
s is precision.

racy obtained
greater than
reaching 4%-
other ranks.
tting a 4%
standard rank
on the old
ich was R2.
accuracy of
le getting 63%
of about 16%.
rence of about
rank used on
RCE. However,
performance
ndard rank R2
f 16% over the
ed by R5.
tance? If so,









SVM-RCE -> Future Work….?

Conclusion: SVM-RCE provides improved classification accuracy with complex microarray data sets when it is compared to the classification accuracy of the same datasets using either SVM-RFE or PDA-RFE. SVM-RCE identifies clusters of correlated genes that when considered together provide greater insight into the structure of the microarray data. Clustering genes for classification appears to result in some concomitant clustering of samples into subgroups.

Our present implementation of SVM-RCE groups genes using the correlation metric. The success of the SVM-RCE method in classification suggests that gene interaction networks or other biologically relevant metrics that group genes based on functional parameters might also be useful.

2009

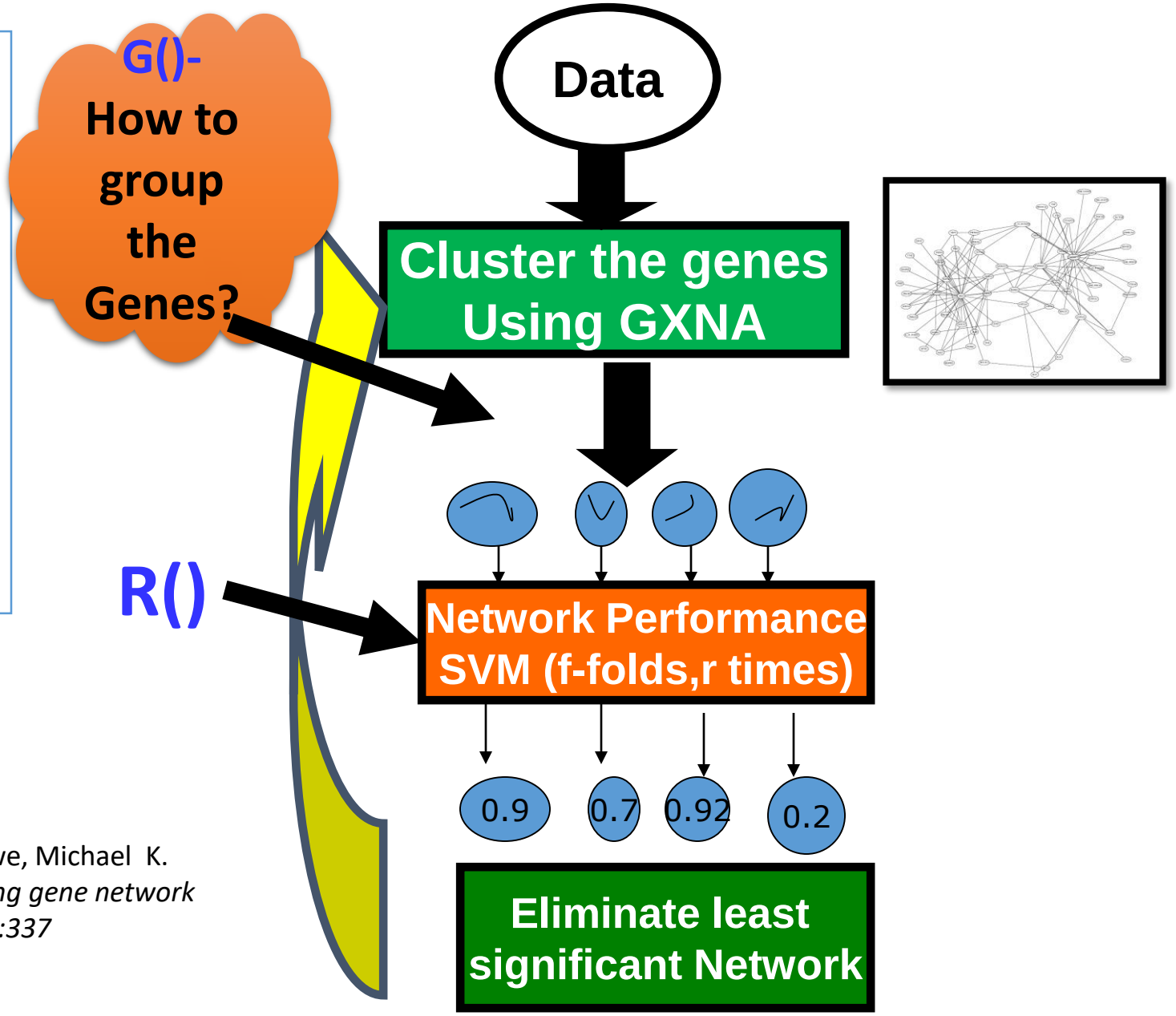
2019 - We have developed maTE and still way to go….

Malik Yousef, Louise C. Showe and Michael K. Showe, Recursive Cluster Elimination (RCE) for Classification and Feature Selection from Gene Expression Data. **BMC Bioinformatics** 8:144 (2007)

SVM-RNE based GXNA - Gene Expression Network Analysis

- GXNA is uses gene expression data set and prior biological interaction network, it suggests differentially expressed pathways or gene sub networks.
- GXNA uses statistical method for scoring sub-networks while a search algorithm is used to determine the sub-networks with high score.

*Nacu, S., et al., Gene expression network analysis and applications to immunology. Bioinformatics, 2007. 23(7): p. 850-858.



BMC Part of Springer Nature
BMC Bioinformatics

Malik Yousef, Mohamed Ketany, Larry Manevitz , Louise C Showe, Michael K. Showe , (2009). *Classification and biomarker identification using gene network modules and support vector machines*. **BMC Bioinformatics**, *10:337* doi:10.1186/1471-2105-10-337

Results : SVM-RNE

Table 1: Summary results for the SVM-RNE, SVM-RCE and SVM-RFE algorithms.

	CTCL(I)			CTCL(II)			Lymphocyte		
	c	g	ACC	c	g	ACC	c	g	ACC
SVM-RNE	2	4	100%	2	5	91%	4	13	80%
	4	8	97%	4	8	90%	6	18	79%
	14	31	96%	14	33	90%	11	30	74%
	24	55	92%	30	69	89%			
SVM-RCE	2	8	96%	2	8	76%	2	13	96%
	3	12	96%	9	34	89%			
	9	32	97%	19	71	91%			
	15	51	97%	28	104	91%			
	32	101	96%						
							6	39	92%
							10	64	92%
SVM-RFE		9	89%		8	84%		12	81%
		32	94%		32	85%		18	79%
		102	100%		102	87%		30	77%

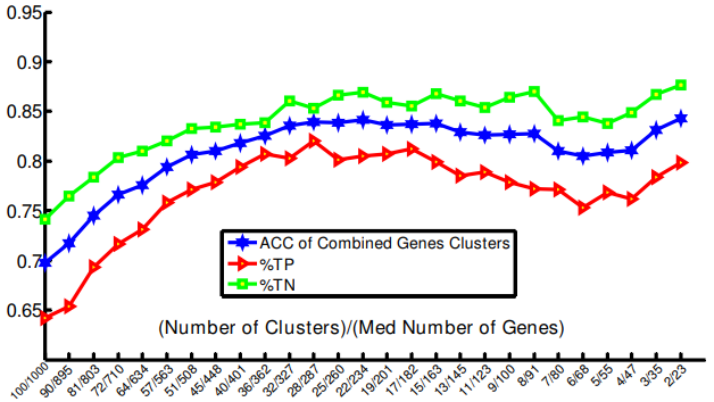
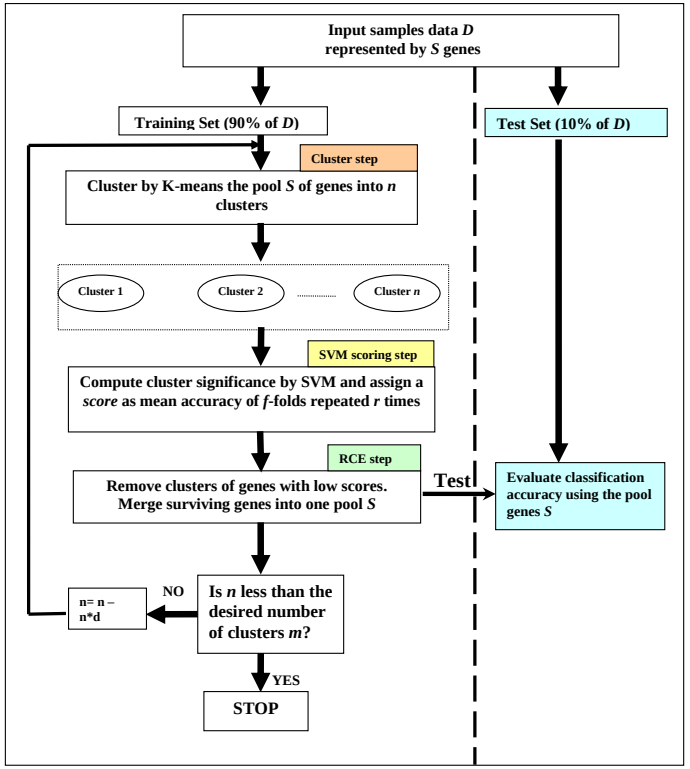
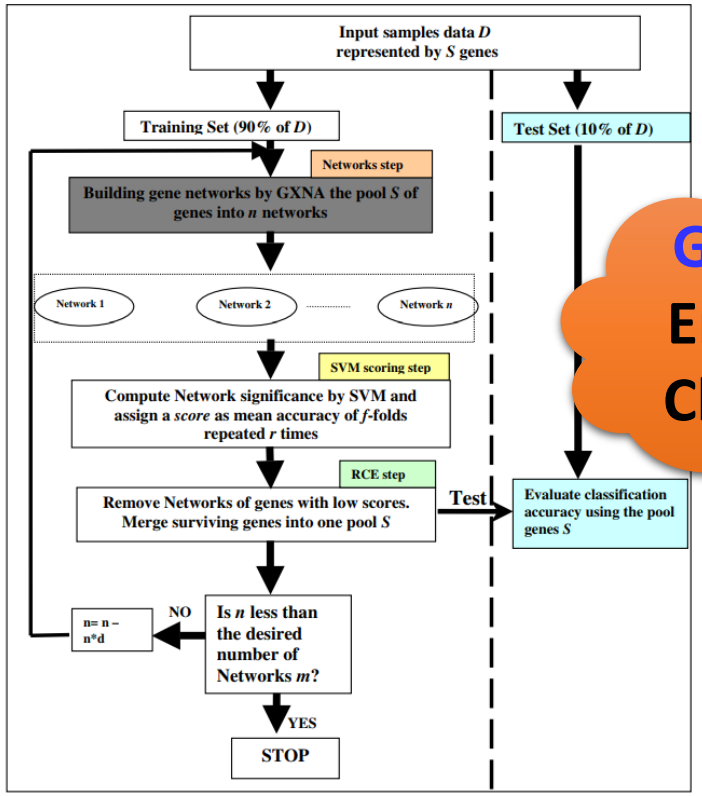


Figure 2
Classification performance of SVM-RCE on the Lymphocyte data set. All of the values are an average of 100 iterations of SVM-RCE. ACC is the accuracy, TP is the sensitivity, and TN is the specificity of the remaining genes determined on the test set. The x-axis shows the median number of clusters and number of genes in the clusters at each step.

Selection of Significant Clusters of Genes based on Ensemble Clustering and Recursive Cluster Elimination (RCE)

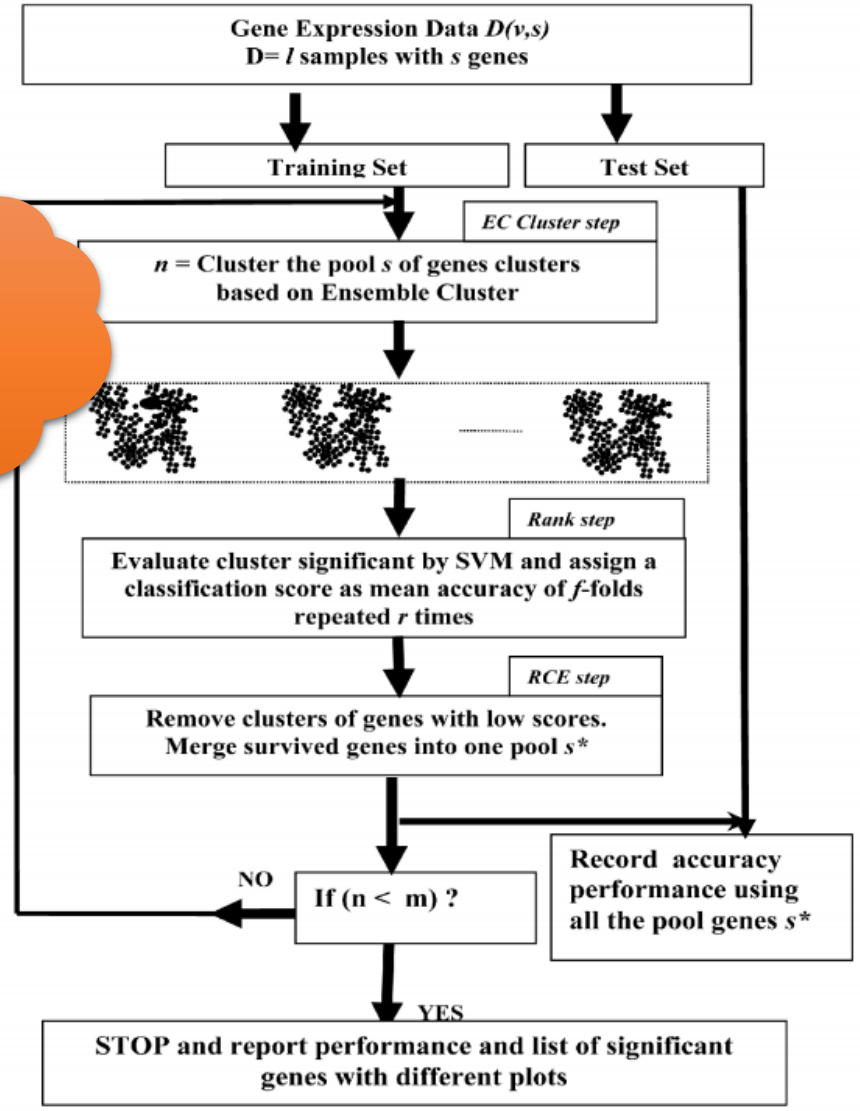




SVM-RCE

SVM-RNE

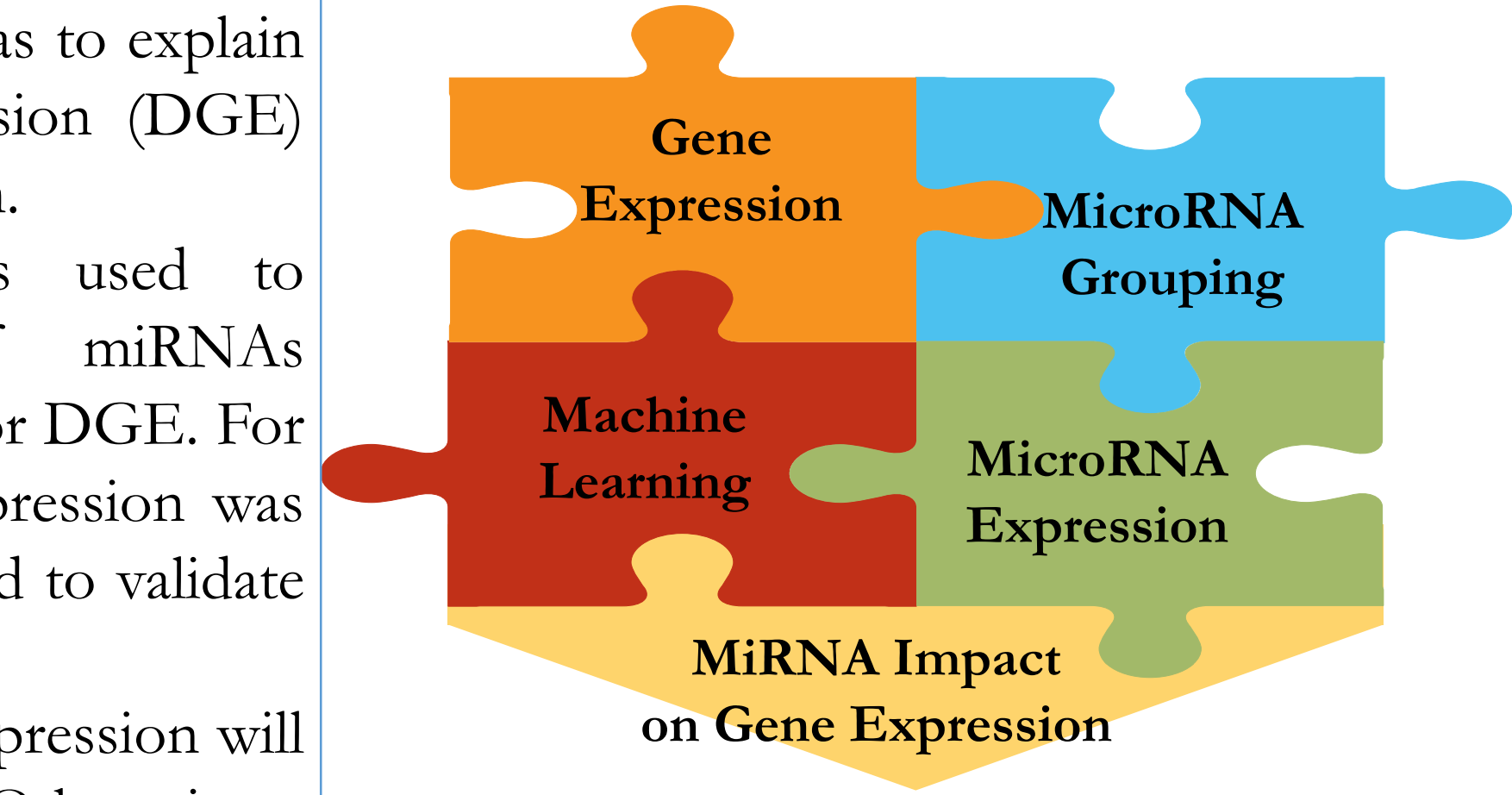
G()- uses Ensemble Clustering



Abdalla L, Khalifa W, Showe LC, **Yousef M** (2017) Selection of Significant Clusters of Genes Based on Ensemble Clustering and Recursive Cluster Elimination (RCE**). J Proteomics Bioinform** 10:186-192. doi: 10.4172/jpb.1000439

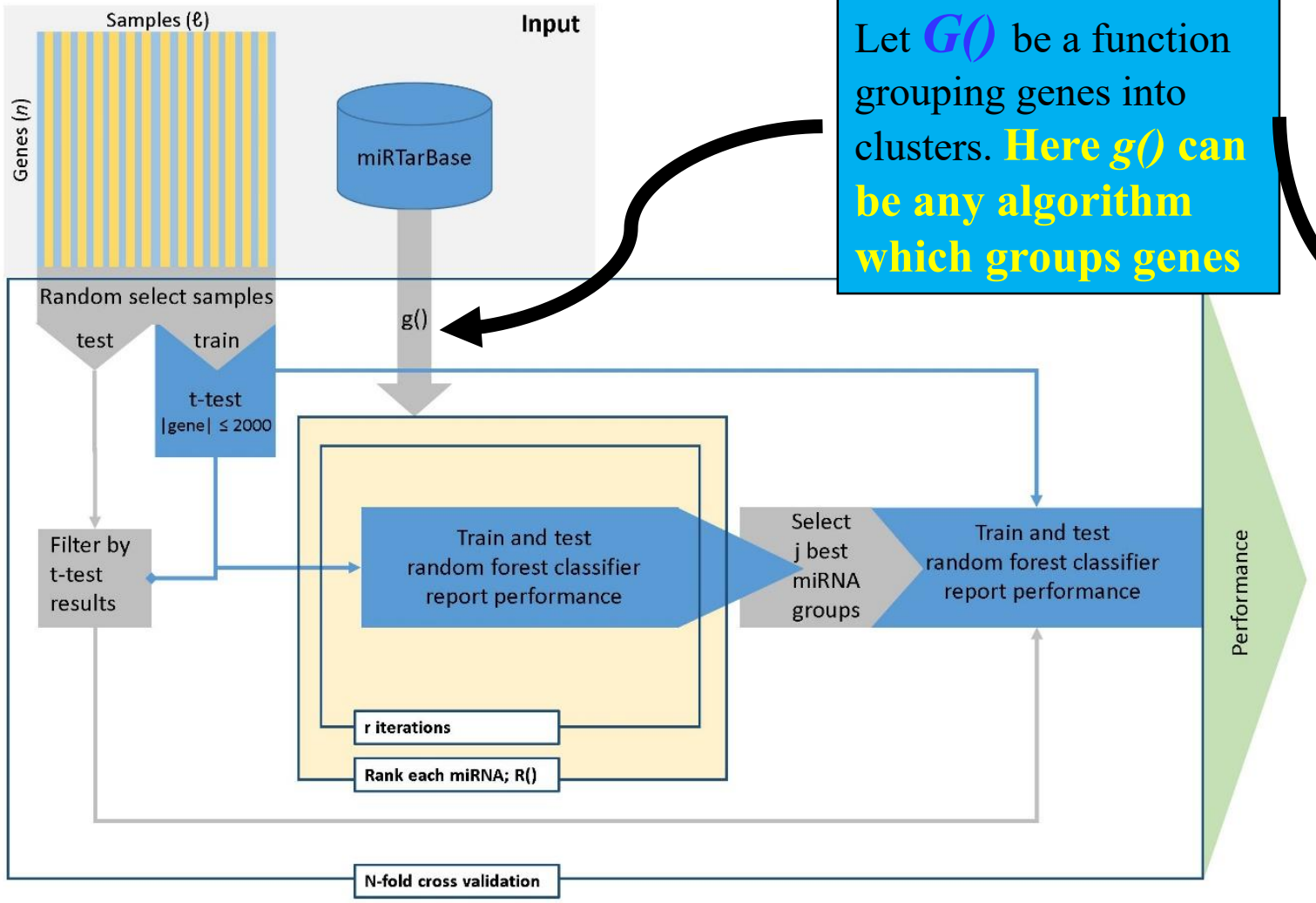
maTE: Discovering Expressed MicroRNA - Target Interactions

- The aim of the study was to explain differential gene expression (DGE) via microRNA regulation.
- Machine learning was used to determine a set of miRNAs potentially responsible for DGE. For one dataset miRNA expression was measured and it was used to validate the results.
- In the future miRNA expression will be added to the puzzle. Other pieces such as differenial protein expression will be evaluated in the future.



Malik Yousef, Loai Abddallah, Jens Allmer, maTE: Discovering Expressed MicroRNA - Target Interactions, **Bioinformatics**, , btz204, https://doi.org/10.1093/bioinformatics/btz204

maTE workflow



Let $G()$ be a function grouping genes into clusters. **Here $g()$ can be any algorithm which groups genes**

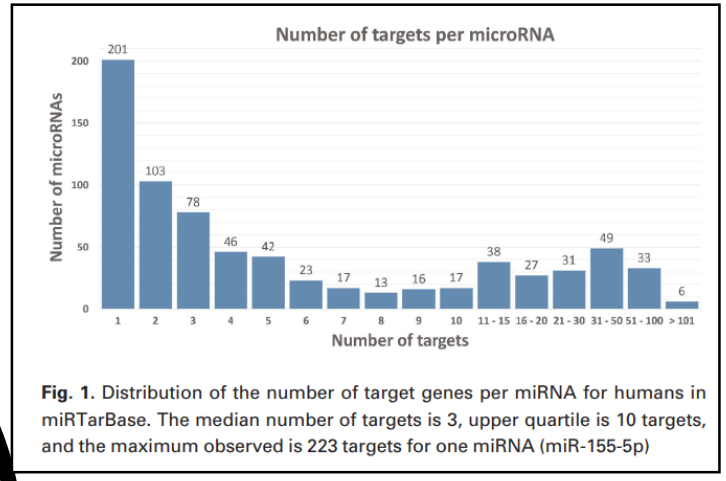
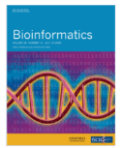


Fig. 1. Distribution of the number of target genes per miRNA for humans in miRTarBase. The median number of targets is 3, upper quartile is 10 targets, and the maximum observed is 223 targets for one miRNA (miR-155-5p)

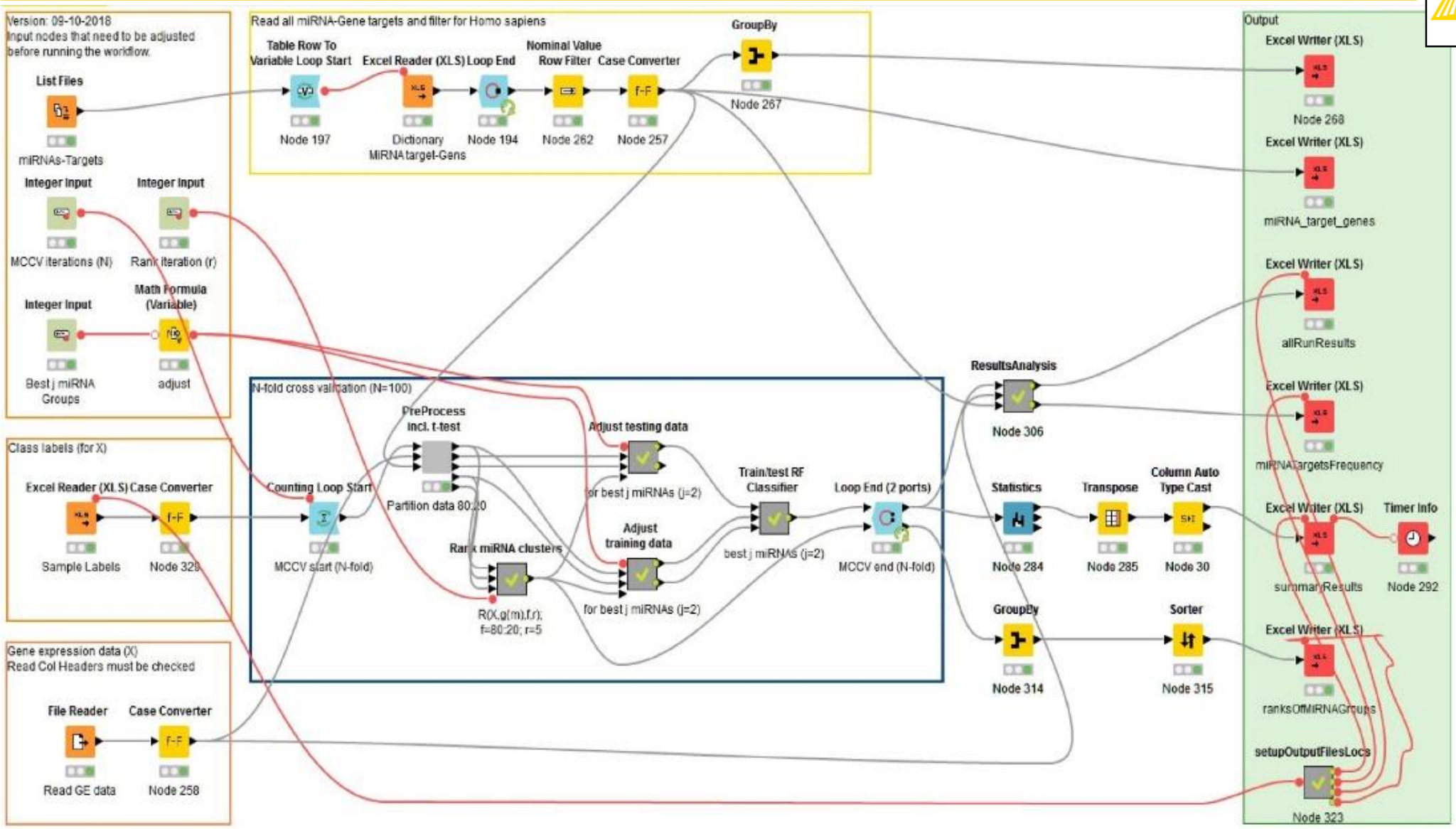
MicroRNA	Target Genes List
HSA-LET-7A-3P	CCND1, CCND2, E2F2
HSA-LET-7D-5P	HMGA2, APP, DICER1, SLC11A2, IL13, MPL, AGO1, TNFRSF10B, COL3A1
HSA-MIR-103A-2-5P	PDCD10
HSA-MIR-129-2-3P	SOX4, UBE2F, CCP110, BCL2L2, MYC, CDK6
HSA-MIR-140-5P	HDAC4, VEGFA, PDGFRA, DNMT1, DNPEP, SOX2, OSTM1, FGF9, TGFBR1, ALDH1A1, SOX9, IGF1R, FZD6, RALA, PAX6, HDAC7, LAMC1, ADA, MMD, PIN1, STAT1, GALC, HMGN5, SOX4, FGFRL1, SMURF1
HSA-MIR-638	OSCP1, SP2, SOX2, CDK2, STARD10, PLD1, PTEN
HSA-MIR-944	S100PBP, HECW2

Malik Yousef, Loai Abddallah, Jens Allmer, *maTE: Discovering Expressed MicroRNA - Targe Interactions,* **Bioinformatics**, , btz204, https://doi.org/10.1093/bioinformatics/btz204

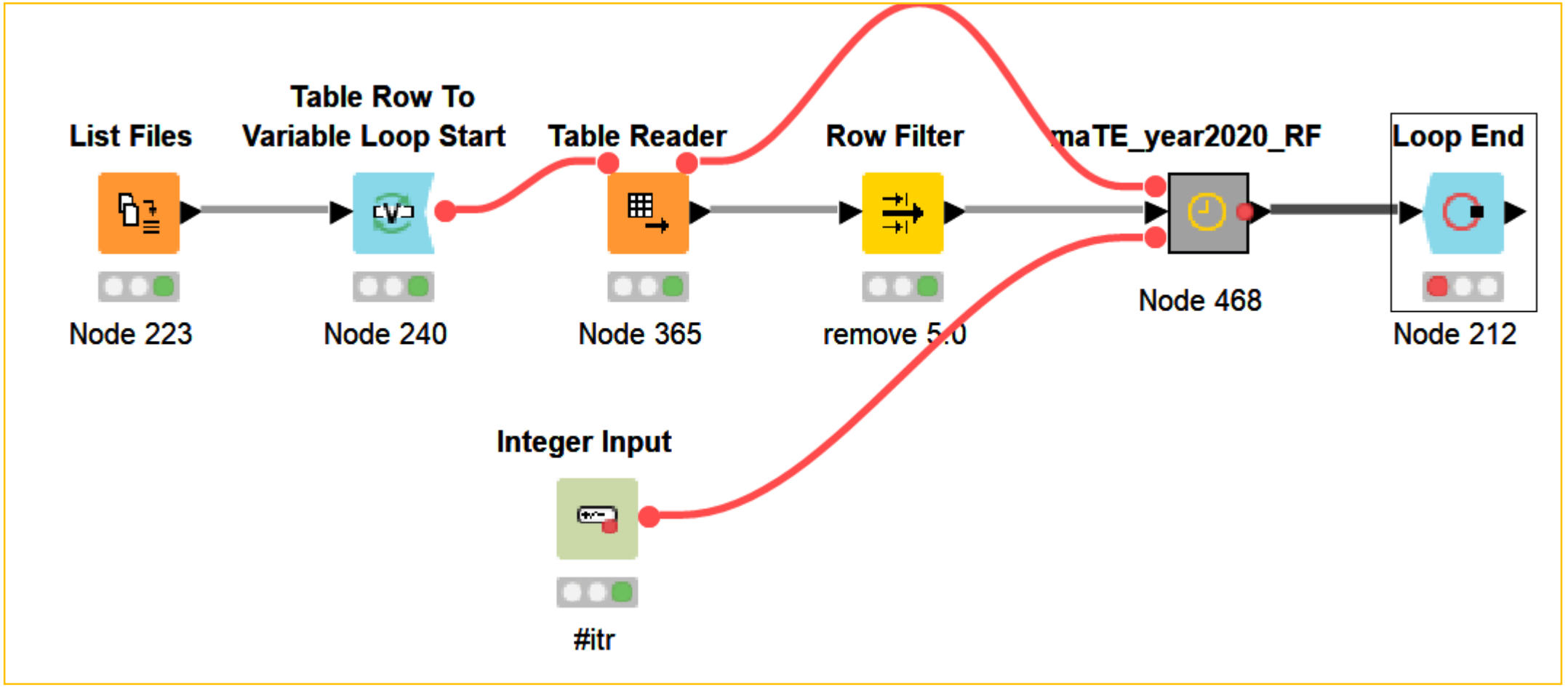


maTE workflow

Implementation of maTE using the data analytics platform KNIME

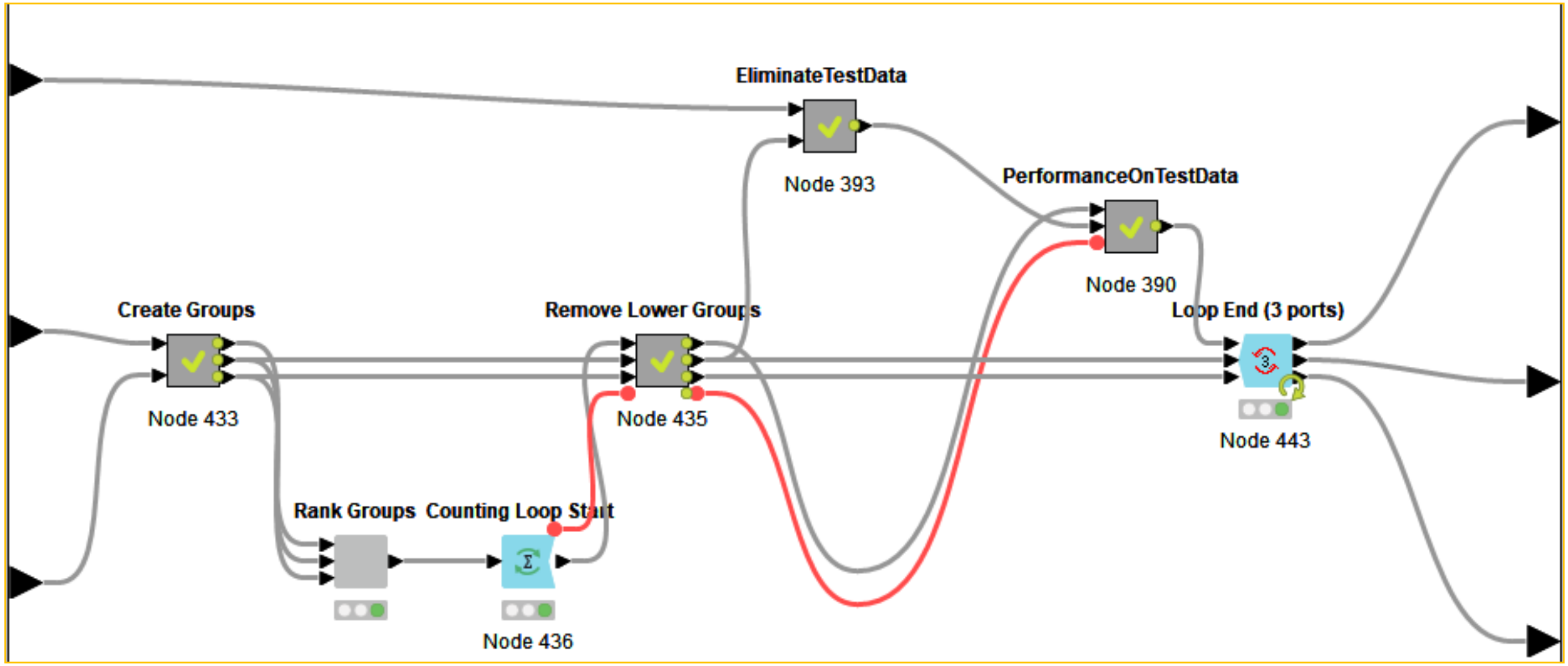


maTE workflow(new version)



maTE workflow(new version)

Creat Groups->Rank->Test



maTE Algorithm

Algorithm 1. The Ranking method $R()$, a main component of the maTE algorithm.

Ranking Algorithm - $R(X_s, g(M), f, r)$

X_s: any subset of the input gene expression data X, the features are gene expression values

M $\{m_1, m_2, \ldots, m_p\}$ is a list of miRNAs

Grouping function g(M)- for each m_i, associate the names of genes (Genes ID) that are targeted by miRNA m_i (See Table 2).

f is a scalar ($0 \leq f \leq 1$): split into train and test data

r: repeated times (iteration)

res={} for aggregation the scores for each m_i

Generate Rank for each m_i-*Rank*(m_i):

For each m_i in M

sm_i=0;

Perform r time (here $r = 5$) Steps 1–5:

1. Perform stratified random sampling to split X_s into train X_t and test X_v datasets according to f (here 80:20)
2. Remove all genes (features) from X_t and X_v which are not targets of m_i
3. Train classifier on X_t (here Random Forest)
4. t = Test classifier on X_v—calculate performance
5. $sm_i = sm_i + t$;

$Score(m_i) = sm_i / r$; Aggregate performance

$res = \cup_{i=1}^{p} Score(m_i)$

Output

Return res ($res = \{Rank(m_1), Rank(m_2), \ldots, Rank(m_p)\}$)

Algorithm 2. The overall algorithm of maTE, which depends on the $R()$ method (see Algorithm 1).

maTE Algorithm

Objective

maTE aims to select j miRNAs with target genes that can best classify samples by expressions.

Input

X: gene expression data with two-class labels, the features are genes expression.

M $\{m_1, m_2, \ldots, m_p\}$: list of miRNAs (here from miRTarBase) where p is the number of miRNAs.

Grouping function g(M)- for each m_i associate the names of genes (Genes ID) that are targeted by miRNA m_i (See Table 2).

Algorithm

M*={} empty list

Perform N-fold cross-validation (here $N = 100$):

Randomly split data by samples into train (X_t) and test (X_v) parts,

performs Steps 1–6:

1. X_{tf} = filter genes (features) from training data by t-test (here P-value ≤ 0.05 and maximum number of filtered genes ≤ 2000)
2. X_{vf} = remove all genes from X_v that are not in X_{tf}
3. miR_p = **R(X_{tf}, g(M), *f*, *r*)** (here $f = 80{:}20$ with stratified random sampling; $r = 5$). $R()$ is the procedure in Algorithm 1, the output will be miR_p = $\{Rank(m_1), Rank(m_2), \ldots, Rank(m_p)\}$
4. M*= Sort(miR_p) according to performance; best first
5. $M^* = \{m^*_1, m^*_2, .., m^*_j\}$, Select best j miRNAs (here $j = 2$)
6. Filter X_{tf} and X_{vf} by $g(M^*)$, now X_{tf} and X_{vf} represented by genes that are targeted by miRNA from M^*.

Train classifier using X_{tf} and X_{vf} (here random forest)

Test classifier using X_{vf}

Output

Report performance (e.g. average accuracy)

maTE : Knime WorkFlow

Row ID	S miRNA Target Name	D Accuracy	I ...	S Unique concatenate(Target Gene)
Row23	HSA-MIR-106A-5P	0.861	23	E2F1, CDKN1A, HIPK3, MYLIP, RB1, APP, ARID4B, VEGFA, IL10, FAS, TGFBR2, CYP19A1, PTEN, SIRPA, SLC2A3, BMP2, STAT3, ATM, CASP7, BCL10, RUNX3, TIMP2, MAPK9, LIMK1, FASTK, ULK1, HIF1A, RBL2, APC, MFN2, CXCL8, MYB, ATG7, RARB, HMGA2, CCND1, CDX2, MGST2, ERCC1, RND3
Row692	HSA-MIR-891B	0.833	332	CBLB
Row554	HSA-MIR-545-3P	0.806	194	LRP1, SNAI2, CDK4
Row104	HSA-MIR-135A-5P	0.792	104	JAK2, NR3C2, APC, HOXA10, MYC, SMAD5, BMPR2, STAT6, MTSS1, ROCK1, VLDLR, TXNIP, IRS2, HTR1A, SLC6A4, SIAH1, BCL2, KLF8, ESRRA, RUNX2, ROCK2, FOXO1, PHLPP2, E2F1, DAPK2, KLF4, PTPRD, PTK2, CEBPD, PPM1E, MMP11, EGFR
Row113	HSA-MIR-139-5P	0.792	113	IGF1R, MCL1, CXCR4, JUN, NR5A2, NOTCH1, RAP1B, ROCK2, FAM162A, RHOT1, ZHX2, PIK3CA, NFKB1, HRAS, OIP5, ACTC1, ADGRL4, TPD52, MET, BCL2, PDE4D, WNT1, MMP11, FOS, SMARCA4
Row152	HSA-MIR-15A-5P	0.792	152	MYB, CDC25A, CCND2, BCL2, CCND1, CCNE1, BACE1, DMTF1, BRCA1, AKT3, CADM1, TMEM184B, APP, UCP2, VEGFA, TSPYL2, CHUK, TP53, FGF7, CLCN3, CRKL, MN1, HMGA1, HMGA2, IFNG, PURA, RECK, FOXO1, REPIN1, YAP1, CARM1, SOX5, HSPA1B, WEE1, CHEK1, PHLPP1, RET, CXCL10, ...
Row181	HSA-MIR-18A-5P	0.792	181	ESR1, PTEN, CTGF, TNFSF11, NR3C1, HIF1A, TGFBR2, SMAD4, HSF2, ATM, DICER1, PHLPP1, PIAS3, BCL2, TBPL1, SMAD2, SDC4, STK4, BCL2L10, DNMT1, IRF2, MEF2D, TNFAIP3, NR1I2, NCOA3, NEDD9, CDK19, SMAD3, FCGR2B, NEO1
Row387	HSA-MIR-372-3P	0.792	27	WEE1, TRPS1, MBNL2, KLF13, CDKN1A, ERBB4, NR4A2, VEGFA, TGFBR2, RHOC, NFIB, CDK2, CCNA1, TNFAIP1, DKK1, BTG1, LEFTY1, ATAD2, PHLPP2
Row711	HSA-MIR-95-5P	0.792	351	ESR2
Row208	HSA-MIR-19A-3P	0.771	208	HOXA5, MECP2, PTEN, ESR1, CCND1, ERBB4, NR4A2, ATXN1, KAT2B, SOCS1, BCL2L11, TGFBR2, BMPR2, SMAD4, KIT, TLR2, TNF, SUZ12, RAB13, MSMO1, ABCA1, PSAP, DPYSL2, VPS4B, MYCN, RAB14, PHLPP1, MXD1, TF, IMPDH1, NPEPL1, SIVA1, TNFRSF12A, SOCS3, DNMT1, RHOB, TNFAIP...
Row8	HSA-LET-7E-5P	0.75	8	HMGA2, EIF3J, SMC1A, WNT1, CCND1, MPL, MYCN, AGO1, IGF1R, MMP9, IGF1, LIN28A, AURKB, ARID3A, FASLG, TNFAIP3, EZH2, TNFRSF10B, PLK1
Row48	HSA-MIR-124-3P	0.75	48	BDNF, EFNB1, NR3C2, BACE1, ADIPOR2, CEBPA, CTGF, MAPK14, RELA, CDK4, CDK6, PTBP1, AHR, SLC16A1, B4GALT1, IQGAP1, SNAI2, CAV1, CTDSP1, ITGB1, MYO10, NFATC1, RHOG, ELK3, PPP1R13L, CDK2, CCL2, NR3C1, VIM, SMYD3, E2F6, AR, ROCK2, EZH2, IL6R, HMGA1, ROCK1, PIK3...
Row116	HSA-MIR-141-3P	0.75	116	ZEB2, ZEB1, DLX5, BAP1, KLF5, STK3, TGFB2, SFPQ, CLOCK, BRD3, UBAP1, PTEN, ZFPM2, TRAPPC2B, EIF4E, CTBP2, CDYL, ACVR2B, MAPK14, PPARA, NR0B2, E2F3, SHC1, VAC14, TCF7L1, ELMO2, RASSF2, KLHL20, RIN2, HOXB5, ERBIN, KLF11, PTPRD, WDR37, STAT4, YAP1, CDC25C, HDGF,...
Row127	HSA-MIR-146A-5P	0.75	127	CXCR4, TLR2, FADD, TRAF6, IRAK1, ROCK1, BRCA2, BRCA1, FAF1, NFKB1, CDKN1A, EGFR, CD40LG, FAS, ERBB4, SMAD4, TLR4, WASF2, STAT1, CD80, L1CAM, CARD10, COPS8, ELAVL1, PTGS2, CCL5, PTGES2, SIKE1, CXCL12, PRKCE, RAC1, LAMC2, RNF11, SOS1, CASP7, BCLAF1, CPM, NO...
Row150	HSA-MIR-155-5P	0.75	150	MEIS1, TAB2, MECP2, SOCS1, MSH6, MSH2, MLH1, INPP5D, DET1, SMAD5, HIVEP2, ZNF652, ZIC3, BACH1, JARID2, APC, TRIP13, TRIM32, TBCA, SMAD1, SDCBP, RHEB, POLE3, PKN2, PICALM, PHC2, NARS, MYO10, DHX40, CLDN1, CEBPB, ARFIP1, TM6SF1, LDOC1, JADE1, RHOA, AGTR1, FG...
Row154	HSA-MIR-15B-5P	0.75	154	CCNE1, RECK, BCL2, CCND1, VEGFA, IFNG, PURA, CHEK1, SMAD7, FOXO1, SMURF1, PPM1D, WEE1, FUT2, KDR, HNF1A, AKT3, TRIM29, INSR, CCND3, RAB1A, MTSS1, OIP5, SOCS3, TGFB1, TBR1, SMAD2, BAX, MMP9, TRIM14
Row159	HSA-MIR-17-5P	0.75	159	CCL1, GPR137B, NABP1, NPAT, YES1, JAK1, PTEN, CDKN1A, PTPRO, PKD2, BCL2L11, E2F1, MAP3K12, BCL2, MEF2D, APP, VEGFA, MAPK9, FBXO31, HIF1A, TGFBR2, BMPR2, CCND1, NCOA3, SMAD4, ICAM1, SELE, CCND2, E2F3, RB1, RBL1, RBL2, WEE1, RND3, SMURF1, TCF3, TCEAL1, HBP1, ...
Row214	HSA-MIR-200B-3P	0.75	214	ZEB2, BAP1, ZEB1, RERE, ETS1, FN1, WASF3, ZFPM2, RNF2, E2F3, VEGFA, FLT1, KDR, RND3, CCNE2, BCL2, XIAP, SMAD2, CREB1, KLHL20, ELMO2, PTPRD, ERBIN, WDR37, TCF7L1, VAC14, HOXB5, RIN2, RASSF2, KLF11, SHC1, MYB, QRSL1, SUZ12, WNT1, DNMT3A, DNMT3B, SP1, MSN, FER...
Row230	HSA-MIR-20A-5P	0.75	230	HIF1A, TCEAL1, CCND1, E2F1, BMPR2, CDKN1A, TGFBR2, MAP3K12, BCL2, MEF2D, PTEN, APP, VEGFA, BCL2L11, SMAD4, CCND2, E2F3, MAPK9, RB1, RBL1, RBL2, WEE1, IRF2, KIT, EGLN3, PPARG, BAMBI, PURA, ARHGAP12, TSG101, SIRPA, UBE2C, STAT3, LIMK1, PHLPP2, GJA1, DUSP2, ITG...
Row258	HSA-MIR-222-3P	0.75	258	STAT5A, CDKN1B, MMP1, FOXO3, CDKN1C, KIT, PPP2R2A, TIMP3, FOS, ICAM1, ESR1, PTEN, SELE, SSSCA1, DIRAS3, ETS1, DICER1, RECK, TRPS1, CERS2, GRB10, GJA1, SSX2IP, DKK2, ARID1A, MGMT, VGLL4, GNAI3, PRDM1, GNAI2, ABCG2, PLXNC1, TNFSF10, TP53, CORO1A, TCEAL1, SMA...
Row278	HSA-MIR-26B-5P	0.75	278	PTGS2, EPHA2, CDK6, CCNE1, ABCA1, ARL4C, GATA4, CHORDC1, PLOD2, NR2C2, TAB1, EZH2, USP9X, KPNA2, RB1, NAMPT, IGF1R, PTEN, ULK2, TRAF5, SMAD1, HAS2, COL1A2, CTGF, TLR4, IGF1, ST8SIA4, PDE4A, FH, HGF, JAG1, LARP1
Row331	HSA-MIR-320A	0.75	331	POLR3D, TAC1, TFRC, MCL1, HSPB6, AQP1, AQP4, NPR1, MAPK1, IGF1R, HOXA10, ARPP19, VDAC1, MYC, ITGB3, GNAI1, RAC1, PBX3, NRP1, NFATC3, TRPC5, PTEN, BANP, RAB14, SUZ12, FOXM1, KITLG, RUNX2, MTDH, YWHAZ, VIM, USP14, ABCG2, CTNNB1, CRKL, ESRRG, RAB11A, NOD2, ...
Row365	HSA-MIR-34A-5P	0.75	5	JAG1, MYC, MYB, MET, CDK4, CCNE2, MYCN, CDK6, CCND1, WNT1, E2F3, VAMP2, SIRT1, BCL2, NOTCH1, HNF4A, YY1, MAGEA12, MAGEA6, MDM4, MAP2K1, VEGFA, IFNB1, E2F1, NOTCH2, CD44, MAP3K9, CEBPB, SPI1, ZAP70, PDGFRA, NANOG, SOX2, IMPA1, IMPDH2, ULBP2, SYT1, STX1A, ...
Row477	HSA-MIR-489-3P	0.75	117	PTPN11, GSE1, SNAI2, PAX3, SPIN1, PROX1, MMP7
Row485	HSA-MIR-494-3P	0.75	125	PTEN, CDK6, ARNTL, BCL2L11, SDC1, PROS1, ZEB1, BAG1, MYC, TFAM, FOXJ3, HNRNPA3, PDIA3, RAD23B, SYNCRIP, ARHGAP5, MCC, CNR1, SLC26A3, CXCR4, CFTR, HOXA10, IGF1R, UGT2B17, TFPI, ATF3, MAP2K1, ITPR1, BCL2, MYH2, AKT1, BIRC5
Row533	HSA-MIR-520B	0.75	173	CDKN1A, MICA, CXCL8, MAP3K2, CCND1, CD46, PFKP, EGFR, LAMTOR5
Row657	HSA-MIR-675-5P	0.75	297	RB1, TGFBI, MITF, CDC6, REPS2, ATP8A2, TGFB1, HDAC4, HDAC5, HDAC6, DDB2
Row672	HSA-MIR-765	0.75	312	NTRK3, HNF4A, INPP4B, EMP3, HES1
Row118	HSA-MIR-142-3P	0.75	118	RAC1, ARNTL, TGFBR1, PROM1, ROCK2, CCNT2, TAB2, HMGA1, IL1A, HSPA1B, APC, LPP, ABCG2, LGR5, HOXA10, HOXA7, LRRC32, HMGB1, DOCK6, THBS4, EGR2, PRKCA
Row244	HSA-MIR-215-5P	0.75	244	WNK1, ACVR2B, CTNNBIP1, ALCAM, RB1, XIAP, PTPRT, NID1, SIGLEC8, ZEB2
Row15	HSA-MIR-100-3P	0.733	15	CXCL8
Row18	HSA-MIR-101-5P	0.729	18	PRKDC, ATM, ACKR3, VEGFA, CCDC88A, RAC1, SOX9, TLR2, STMN1, RAB1A, FOS, EZH2
Row60	HSA-MIR-125B-1-3P	0.729	60	S1PR1, TACSTD2, SGPL1, BIK, TP53, MAP2K7, FRZB, ITGA9
Row156	HSA-MIR-16-2-3P	0.729	156	RARB
Row17	HSA-MIR-101-3P	0.722	17	MYCN, ATXN1, EZH2, APP, FOS, MCL1, PTGS2, ATM, ATP5B, DUSP1, STMN1, RAB5A, SOX9, DNMT3A, FMR1, VEGFA, RAC1, CDH5, ZEB1, PTGER4, CPEB1, MTOR, CFTR, KLF6, ZEB2, RAP1B, PIM1, JAK2, MET, NLK, MITF, PRDM1, VEGFC, NOTCH1, PRKAB1, RHOA, SRF, CDK8, EYA1, VHL, GRSF...
Row201	HSA-MIR-197-3P	0.722	201	TUSC2, NSUN5, CD82, PMAIP1, MTHFD1, FOXJ2, MAPK1, RAN, FOXO3
Row64	HSA-MIR-126-5P	0.722	64	PTPN7, ADAM9, MMP7, CXCL12, MYC, VEGFA, CRK, CYLD
Row661	HSA-MIR-708-5P	0.722	301	ZEB2, BIRC5, AKT2, CD44, EYA3, NNAT, CASP2, CNTFR, SMAD3, IKBKG, KDM1A, AKT1, CCND1, MMP2, EZH2, PARP1, BCL2
Row123	HSA-MIR-144-5P	0.717	123	TGIF1, ROCK1, ROCK2, MET, CCNE1, CCNE2, SMAD4
Row606	HSA-MIR-608	0.717	246	CDC42, CD44, NAA10, BCL2L1
Row3	HSA-LET-7B-5P	0.708	3	CDC34, HMGA1, RPIA, ACTG1, HMGA2, CDC25A, CDK6, CCND1, CCND2, NRAS, CCNA2, LIN28A, IFNB1, NR2E1, PDGFRA, IGF2BP2, PRDM1, CYP2J2, CCNA1, AGO1, IRS2, IGF1R, COL3A1, CPEB1, CPEB3, ACVR1, LRIG1, TGFBR1, AKT2, TLR4, LGR4, E2F2, HRAS, TNFRSF10B, EZH2
Row25	HSA-MIR-106B-5P	0.708	25	ITCH, APP, CDKN1A, E2F1, KAT2B, VEGFA, BCL2L11, RB1, TCEAL1, CCND1, CCND2, E2F3, MAPK9, PTEN, RBL1, RBL2, WEE1, PURA, APC, CASP7, JAK1, SETD2, ATG16L1, SMAD7, STAT3, TWIST1, HIF1A, TRIM8, CASP8, RUNX3, MMP2, MFN2, FAM129A, RHOC, PRRX1, FYN, SLC2A4, PTENP1, ...
Row108	HSA-MIR-136-5P	0.708	108	MTDH, BCL2, RASAL2, IL6
Row129	HSA-MIR-146B-5P	0.708	129	NFKB1, CDKN1A, MMP16, TRAF6, IRAK1, TLR4, PDGFRA, HNRNPD, IL6, PAX8, RARB, EGFR, ERBB4, SLC5A5, KIT
Row183	HSA-MIR-1908-5P	0.708	183	APOE, DNAJA4, NKIRAS2, SKI
Row227	HSA-MIR-208A-3P	0.708	227	CDKN1A, MED13, ETS1, CACNA1C, CACNB2, QKI
Row260	HSA-MIR-223-3P	0.708	260	MEF2C, STMN1, LMO2, E2F1, RHOB, NFIX, CHUK, FBXW7, IGF1R, LIF, SP3, EPB41L3, SLC2A4, ARTN, FOXO1, HSP90B1, SCARB1, PARP1, CDK2, ECT2, PTBP2, ATM, CYB5A, TOX, PRDM1, STAT5A, CARM1, POLR3G, FOXO3, CDC27, SP1, IL6, CXCL2, ABCB1, CAPRIN1, PAX6, CFTR, MYL9, STAT...
Row295	HSA-MIR-29B-3P	0.708	295	TGFB3, HDAC4, CTNNBIP1, COL5A3, COL4A2, COL1A1, SP1, CDK6, BACE1, SFPQ, DNMT3B, DNMT3A, MCL1, BCL2, DNMT1, S100B, VEGFA, TET1, TCL1A, MMP15, MMP24, GRN, FGG, FGA, FGB, COL3A1, COL4A1, MMP2, ADAM12, NID1, HMGA2, TGFB2, TGFB1, BMP1, PTEN, NASP, PPP1R13B, ...

Output of maTE

#Groups	#Genes (Mean)	Accuracy (Mean)	Sensitivity (Mean)	Specifity (Mean)	F-measure (Mean)	Area Under Curve (Mean)	Recall (Mean)	Precision (Mean)	Cohen's kappa (Mean)
10	24.7	0.97	0.96	1.00	0.98	0.98	0.96	1.00	0.95
9	22.9	0.97	0.98	0.95	0.98	0.97	0.98	0.98	0.93
8	20.8	0.99	0.98	1.00	0.99	0.98	0.98	1.00	0.97
7	18.2	0.96	0.96	0.95	0.97	0.95	0.96	0.98	0.90
6	16.5	0.94	0.94	0.95	0.96	0.96	0.94	0.98	0.88
5	14.5	0.94	0.94	0.95	0.96	0.95	0.94	0.98	0.88
4	12	0.96	0.94	1.00	0.96	0.98	0.94	1.00	0.92
3	9.7	0.96	0.96	0.95	0.97	0.96	0.96	0.98	0.90
2	7.5	0.94	0.94	0.95	0.95	0.96	0.94	0.98	0.87
1	3.8	0.90	0.90	0.90	0.92	0.94	0.90	0.96	0.77

Group	#Frequency	Score
HSA-MIR-145-5P	35	3.5
HSA-MIR-195-5P	21	2.1
HSA-MIR-199A-5P	21	2.1
HSA-MIR-372-3P	21	2.1
HSA-MIR-27B-3P	19	1.9
HSA-MIR-101-3P	17	1.7
HSA-MIR-199A-3P	17	1.7
HSA-MIR-144-3P	16	1.6
HSA-LET-7B-5P	15	1.5
HSA-LET-7A-5P	14	1.4
HSA-MIR-126-3P	12	1.2
HSA-MIR-23A-3P	12	1.2
HSA-MIR-34A-5P	11	1.1
HSA-MIR-105-5P	10	1
HSA-MIR-124-3P	10	1
HSA-MIR-139-3P	10	1
HSA-MIR-146B-5P	10	1
HSA-MIR-210-3P	10	1
HSA-MIR-495-3P	10	1
HSA-MIR-630	10	1

Gene	#Freq	Score
VEGFA	84	8.4
EZH2	63	6.3
SOX9	45	4.5
FZD7	39	3.9
UHRF1	36	3.6
WEE1	34	3.4
TUG1	33	3.3
APOE	32	3.2
LDHA	32	3.2
CD44	26	2.6
COL4A2	26	2.6
CD40	25	2.5
RHOC	24	2.4
ANGPT2	22	2.2
ILK	22	2.2
ZEB1	22	2.2
ARHGAP12	21	2.1
EIF4EBP1	21	2.1
ELN	21	2.1
STAT1	20	2
IGFBP2	19	1.9
PRRX1	19	1.9
SWAP70	19	1.9
ELAVL1	18	1.8
AKT1	17	1.7
VIM	17	1.7
ALDH5A1	16	1.6
COL4A1	16	1.6
CXCR4	16	1.6
FAS	16	1.6

Results

Table 3. Accuracy results for both methods, SVM-RCE and maTE

Dataset	SVM-RCE					maTE				
	SE	SP	ACC	stdev	#G	SE	SP	ACC	stdev	#G
GDS1962	0.97	1.00	0.98	0.06	44	0.96	1.00	0.98	0.05	66
GDS2519	0.87	0.90	0.88	0.14	24	0.64	0.57	0.61	0.10	62
GDS3268	0.89	0.88	0.88	0.08	42	0.78	0.71	0.74	0.07	84
GDS3900	1.00	1.00	1.00	0.00	64	1.00	0.95	0.98	0.01	86
GDS3929	0.98	0.96	0.97	0.05	81	0.50	0.57	0.54	0.10	26
GDS2547	0.89	0.81	0.85	0.08	54	0.87	1.00	0.83	0.07	34
GDS5499	0.96	0.95	0.95	0.07	59	0.79	0.97	0.88	0.09	90
GDS3646	0.96	0.93	0.95	0.10	29	0.42	0.63	0.53	0.16	29
GDS3874	0.97	0.97	0.97	0.00	17	0.77	0.90	0.84	0.15	52
GDS3837	0.97	0.96	0.96	0.05	63	0.76	0.99	0.88	0.04	79

Malik Yousef, Loai Abddallah, Jens Allmer, maTE: Discovering Expressed MicroRNA - Target Interactions, **Bioinformatics**, , btz204, https://doi.org/10.1093/bioinformatics/btz204

CogNet

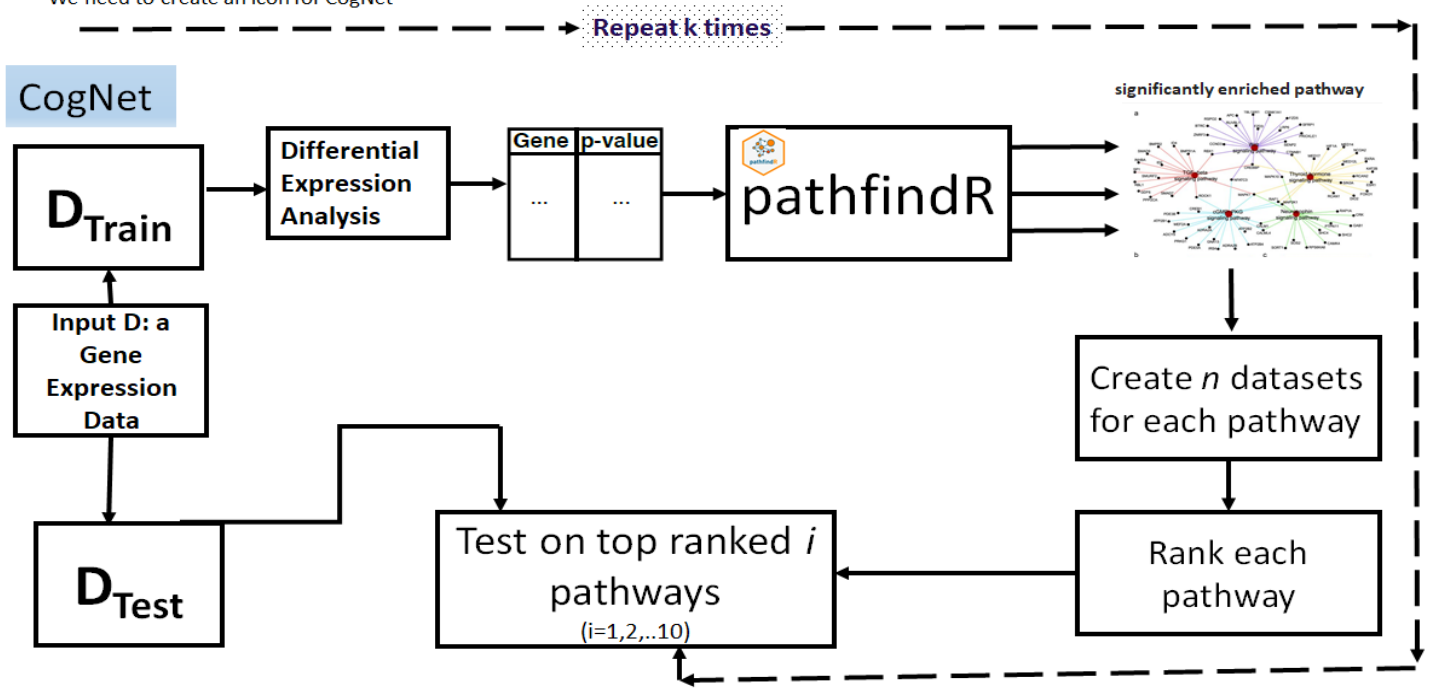
Classification of Gene Expression Data based on ranked Active-Subnetwork-Oriented KEGG Pathway Enrichment Analysis

Malik Yousef[1,2], Ege Ülgen[3] and Osman Ugur Sezerman[3]

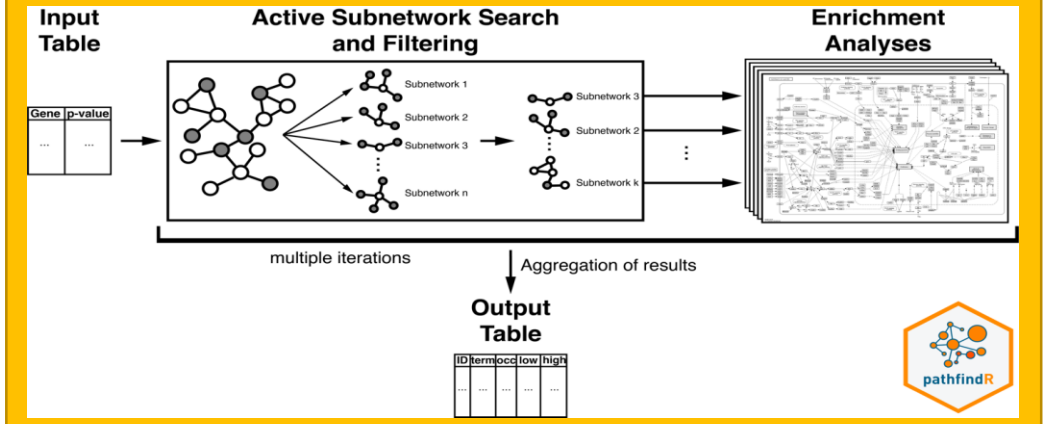
[1]Department of Information Systems, Zefat Academic College, Zefat, 13206, Israel.
[2]Galilee Digital Health Research Center (GDH), Zefat Academic College, Israel
[3]Department of Biostatistics and Medical Informatics,
School of Medicine, Acibadem Mehmet Ali Aydinlar University, Istanbul, Turkey

PeerJ Computer Science

* We need to create an icon for CogNet

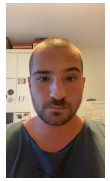


pathfindR



Active-subnetwork-oriented KEGG pathway enrichment analysis of the proteins

E. Ulgen, O. Ozisik, and O. U. Sezerman, "PathfindR: An R package for comprehensive identification of enriched pathways in omics data through active subnetworks," Front. Genet., 2019, doi: 10.3389/fgene.2019.00858.



Department of Biostatistics anc Medical Informatics
Faculty of Medicine, Acibadem University

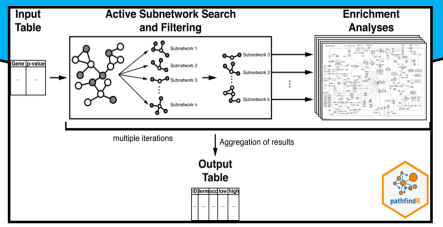
The Generic Approach

Data X = Two class gene expression data with *n* genes $g_1, g_2, \ldots, g_n$

G()- Grouping Component:
Create a list of groups that each group contains its set of genes
$G(g_1, g_2, \ldots, g_n) = \{grp_1, grp_2, \ldots, grp_m\}$

pathfindR

G()-Grouping Function



Training data

R()- Ranking Component :
Ranking each group grp_i by their ability of separating the two classes.
Ranking Function: Use machine learning algorithm with cross validation repeated *r* times. The rank is the average value

With pathfindR, our aim was likewise to exploit interaction information to extract the most relevant pathways. We aimed to combine together active subnetwork search and pathway enrichment analysis

Build classifier from the top *j* groups

Applying the model

Testing data

For example, Mootha et al. (2003) showed that the oxidative phosphorylation genes are significantly associated with **diabetes as a group**, while none of the genes in the pathway showed significant **change individually**

Integrating Biological Knowledge with Gene Expression Profiles for Survival Prediction of Cancer
Xi Chen and Lily Wang

Output of CogNet

#Groups	#Genes (Mean)	Accuracy (Mean)	Sensitivity (Mean)	Specifity (Mean)	F-measure (Mean)	Area Under Curve (Mean)	Recall (Mean)	Precision (Mean)	Cohen's kappa (Mean)
10	123	0.97	0.98	0.95	0.98	1.00	0.98	0.98	0.93
9	110.3	0.97	0.98	0.95	0.98	0.99	0.98	0.98	0.93
8	100.5	0.97	0.98	0.95	0.98	0.99	0.98	0.98	0.93
7	88.5	0.99	0.98	1.00	0.99	0.99	0.98	1.00	0.97
6	75.5	0.97	0.98	0.95	0.98	0.98	0.98	0.98	0.93
5	68.1	0.96	0.98	0.90	0.97	0.99	0.98	0.97	0.89
4	51.7	0.97	0.98	0.95	0.98	0.98	0.98	0.98	0.93
3	39.1	0.97	0.98	0.95	0.98	0.98	0.98	0.98	0.93
2	25.5	0.97	0.98	0.95	0.98	0.98	0.98	0.98	0.93
1	12.4	0.96	0.98	0.90	0.97	0.99	0.98	0.97	0.89

KEGG	#Freq	Score
hsa05163	35	3.5
hsa04072	33	3.3
hsa05224	29	2.9
hsa05215	22	2.2
hsa05213	20	2
hsa04510	19	1.9
hsa04934	19	1.9
hsa05206	19	1.9
hsa05210	19	1.9
hsa05231	19	1.9
hsa04066	17	1.7
hsa04144	17	1.7
hsa04520	17	1.7
hsa04310	16	1.6
hsa04550	14	1.4
hsa05165	14	1.4
hsa04014	13	1.3
hsa04015	13	1.3
hsa05226	13	1.3
hsa04540	12	1.2

Gene	#Freq	Score
TP53	33	3.3
PLD1	31	3.1
TCF7L2	31	3.1
EGFR	30	3
EGFR	30	3
EGFR	29	2.9
EGFR	27	2.7
TP53	27	2.7
VEGFA	27	2.7
VEGFA	27	2.7
EGFR	26	2.6
FYN	26	2.6
AGPAT5	25	2.5
ARAP3	25	2.5
FYN	25	2.5
NRAS	24	2.4
TCF7L2	24	2.4
TCF7L2	24	2.4
TCF7L2	24	2.4
ZEB1	23	2.3

CogNet

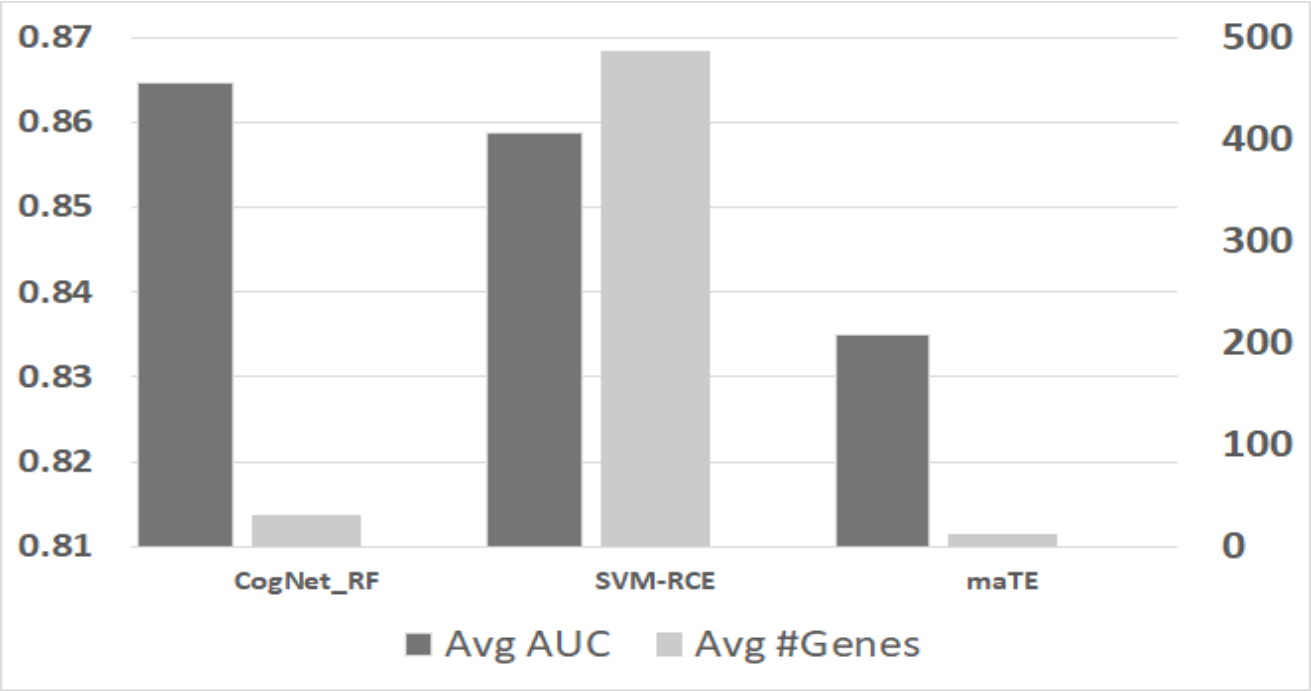


Figure presents the average of all AUC results over the 13 datasets on the 10 clusters/groups, for each tool (Avg AUC bar), while the Avg#Genes bar represent the average of the number of genes over the 13 datasets on the 10 clusters/groups.

- We have considered 13 gene expression data sets to test **CogNet** and for comparison with SVM-RCE and maTE
- For each tool other than **SVM-RCE**, we have obtained the performance over the top 1-10 groups that were ranked by the Ranking **R()** stage
- The purpose of the comparison is not to prove a higher performance, even though it outperform **maTE** and gets similar performance to **SVM-RCE** while with advantage of using a very less number of genes than **SVM-RCE**.

CogNet- Validation of the Results

- We have conducted further analysis using the datasets (GSE15573, GSE4107 and GSE55945).
- We have run the **CogNet** tool on those three datasets obtaining the performance **on top significant pathways.**
- For each dataset, the top 10 groups (pathways) identified by **CogNet** were manually examined in the literature for a possible association with the disease under study.
- Literature support for the top 10 pathways per each dataset are presented in next slide.

CogNet- Validation of the Results

Dataset	Investigated disease	ID	Pathway	Literature support	PMID
GSE15573	Rheumatoid Arthritis	hsa04932	Non-alcoholic fatty liver disease (NAFLD)	None	None
GSE15574	Rheumatoid Arthritis	hsa04120	Ubiquitin mediated proteo sis	Aberration of this system leads to the dysregulation of cellular homeostasis and the development of multiple inflammatory and autoimmune diseases, including rheumatoid arthritis.	16978533
GSE15575	Rheumatoid Arthritis	hsa0501	on	None	None
GSE15576	Rheumatoid Arthritis	hsa	mogenesis	None	None
GSE15577	Rheumatoid Arthritis		Retrograde endocannabinoid signaling	Endocannabinoids, a group of endogenous bioactive lipids, have immunomodulatory effects able to influence both inflammation and pain in rheumatic disease, including rheumatoid arthritis.	29164003, 28857069
GSE15578	Rheumatoid Arthritis	hsa05010	Alzheimer disease	None	None
GSE15579	Rheumatoid Arthritis	hsa04140	Autophagy	Deregulation of autophagic pathway has recently been implicated in the pathogenesis of several autoimmune diseases, including rheumatoid arthritis.	30072986

Partial Results

For GSE15573, the rheumatoid arthritis dataset, **4 out the top 10** pathways were found to be supported by literature to be associated with rheumatoid arthritis biology.

For GSE4107, comparing colorectal cancer patients to healthy controls, **5 out of the top 10 pathways** were supported by literature to be associated with colorectal cancer.

Finally, for GSE55945, the dataset comparing human prostate benign and malignant tissue**, 6 out of the top 10 pathways** were found to be associated with prostate cancer.

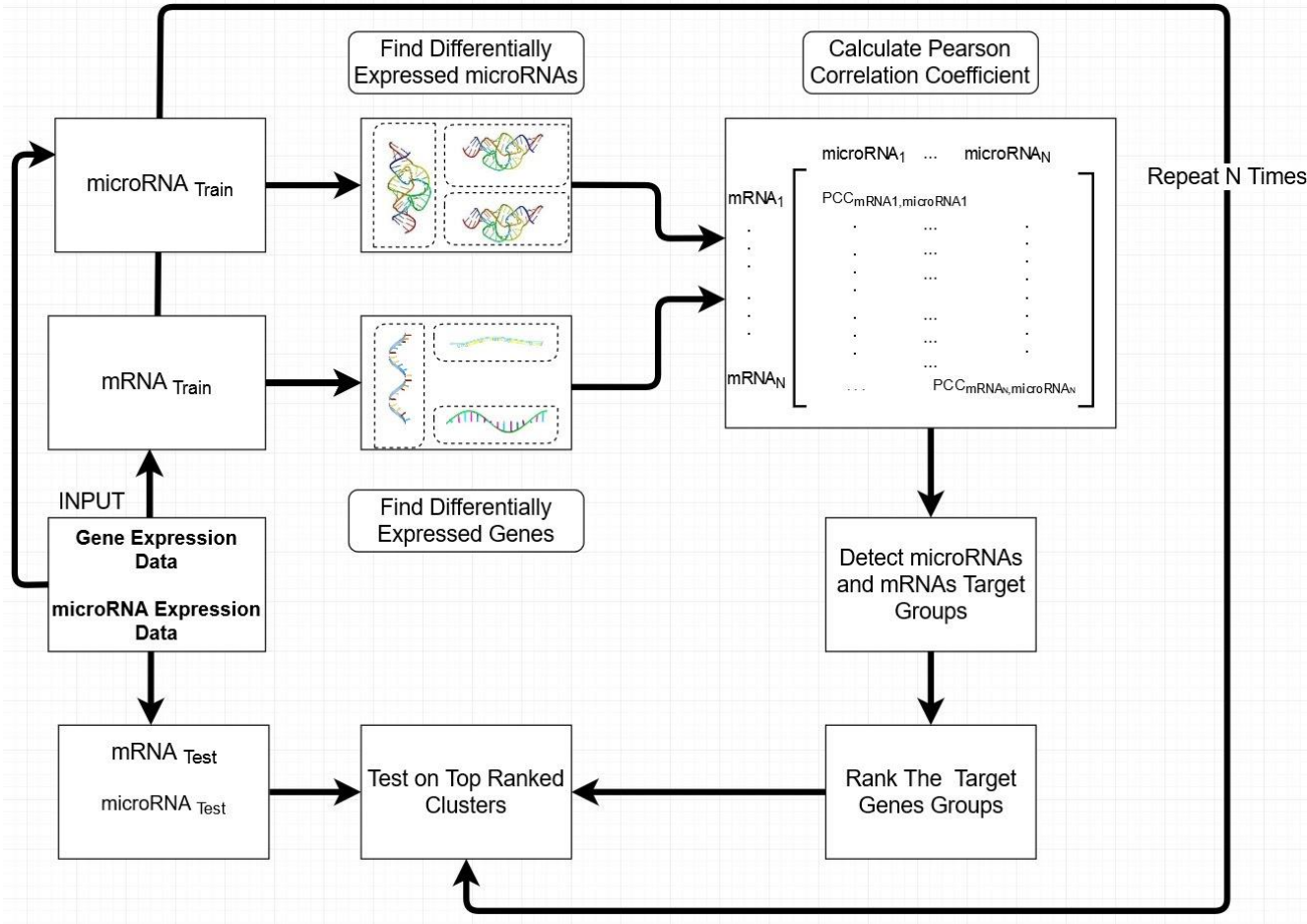
miRcorrNet

Integrated microRNA Gene Expression and mRNA Expression Based Machine Learning combined with Features Grouping and Ranking

Malik Yousef[1,2], Gokhan Goy[3], Ramkrishna Mitra[4], Christine M. Eischen[4], and Burcu Bakir-Gungor[3]

1 Department of Information Systems, Zefat Academic College, Zefat, 13206, Israel.
2 Galilee Digital Health Research Center (GDH), Zefat Academic College, Israel
3 Department of Computer Engineering, Abdullah Gül University, Kayseri, 38090, Turkey
[4]Thomas Jefferson University

Gokhan Goy
PhD Student

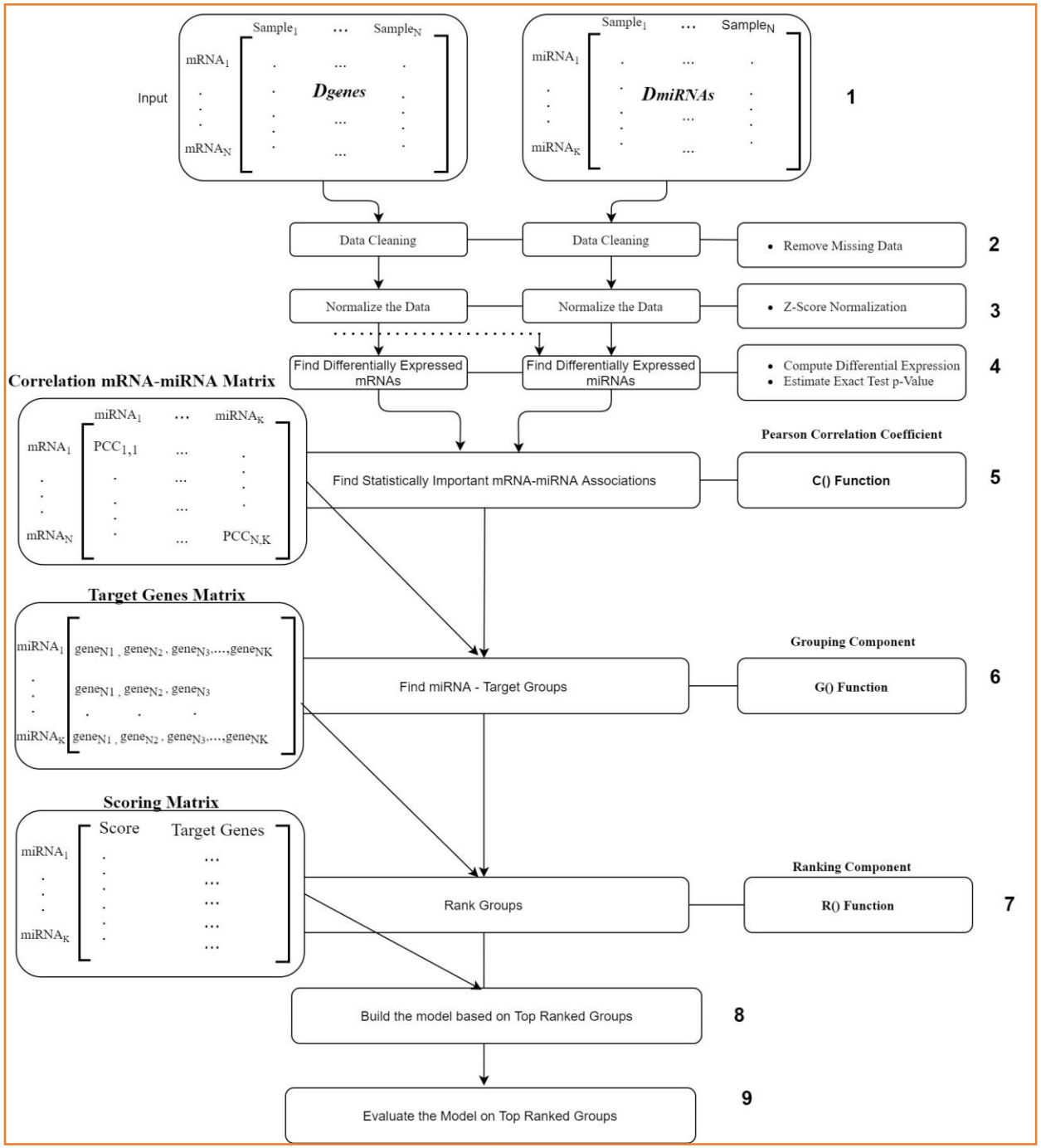


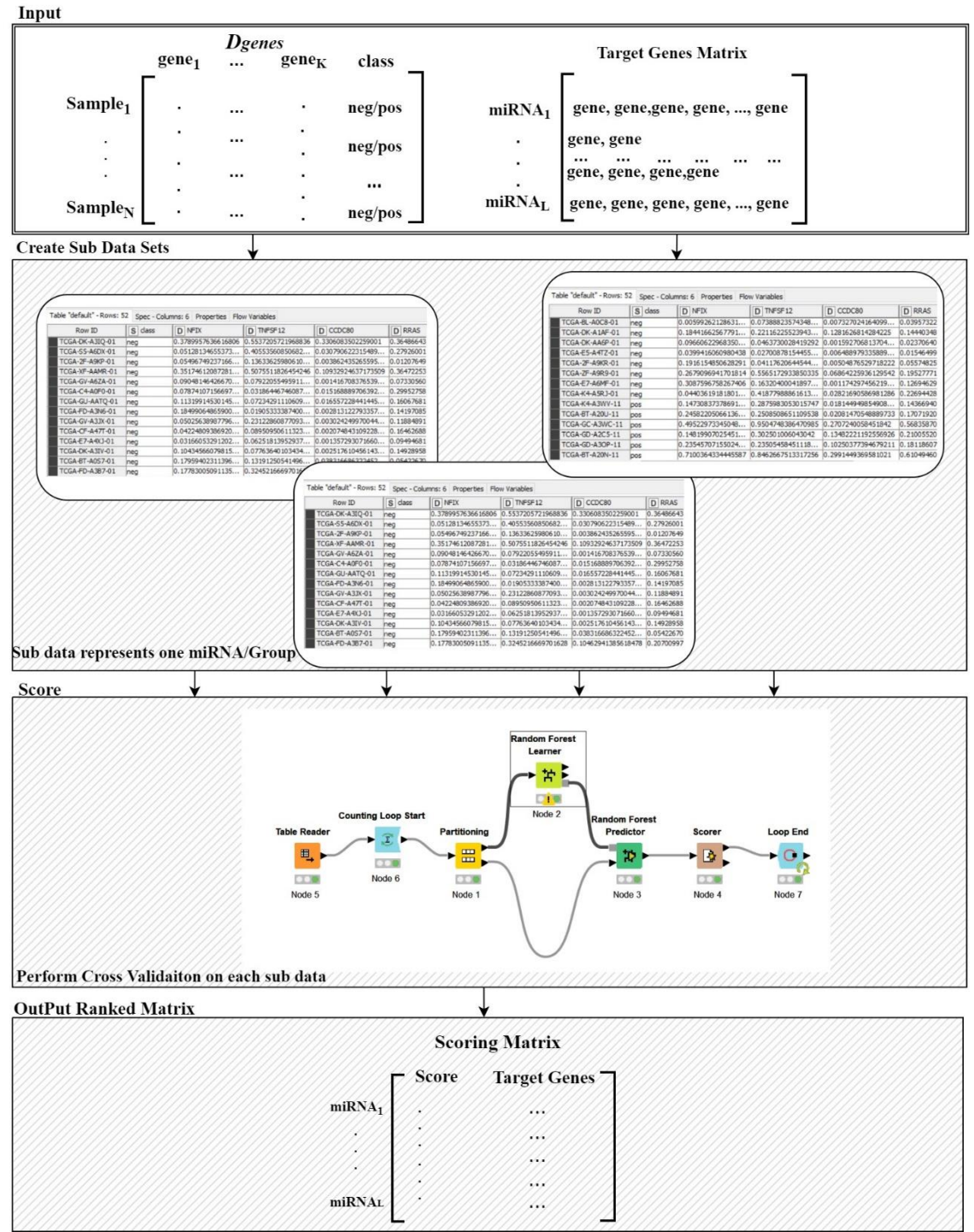
Ramkrishna Mitra, Christine M. Eischen from Thomas Jefferosn Univeristy has contoributed to the Biological Interpretation of the Results

Jefferson
Philadelphia University + Thomas Jefferson University
HOME OF SIDNEY KIMMEL MEDICAL COLLEGE

PeerJ Life & Environment

Yousef M, Goy G, Mitra R, Eischen CM, Jabeer A, Bakir-Gungor B. miRcorrNet: machine learning-based integration of miRNA and mRNA expression profiles, combined with feature grouping and ranking. PeerJ. 2021 May 19;9:e11458. doi: 10.7717/peerj.11458. PMID: 34055490; PMCID: PMC8140596.


Input
Sample1 … SampleN
mRNA1
mRNAN
Dgenes
DmiRNAs
miRNAK
1
Data Cleaning
Remove Missing Data
2
Normalize the Data
Z-Score Normalization
3
Find Differentially Expressed mRNAs
Find Differentially Expressed miRNAs
Compute Differential Expression
Estimate Exact Test p-Value
4
Correlation mRNA-miRNA Matrix
PCC1,1
PCCN,K
Pearson Correlation Coefficient
Find Statistically Important mRNA-miRNA Associations
C() Function
5
Target Genes Matrix
Grouping Component
Find miRNA - Target Groups
G() Function
6
Scoring Matrix
Score
Target Genes
Ranking Component
Rank Groups
R() Function
7
Build the model based on Top Ranked Groups
8
Evaluate the Model on Top Ranked Groups
9


Input
Dgenes
gene1 … geneK class
Sample1
SampleN
neg/pos
Target Genes Matrix
miRNA1
miRNAL
gene, gene,gene, gene, ..., gene
gene, gene
gene, gene, gene,gene
Create Sub Data Sets
Sub data represents one miRNA/Group
Score
Random Forest Learner
Node 2
Counting Loop Start
Node 6
Table Reader
Node 5
Partitioning
Node 1
Random Forest Predictor
Node 3
Scorer
Node 4
Loop End
Node 7
Perform Cross Validaiton on each sub data
OutPut Ranked Matrix
Scoring Matrix
Score
Target Genes

Output of miRcorrNet

#Group	#Genes (Mean)	Accuracy (Mean)	Sensitivity (Mean)	Specifity (Mean)	F-measure (Mean)	Area Under Curve (Mean)	Recall (Mean)	Precision (Mean)	Cohen's kappa (Mean)
10	66.0	0.96	1.00	0.94	0.95	1.00	1.00	0.90	0.92
9	60.8	0.98	0.98	0.98	0.97	1.00	0.98	0.97	0.96
8	56.8	0.98	0.97	0.99	0.97	1.00	0.97	0.97	0.96
7	56.1	0.98	0.95	0.99	0.96	1.00	0.95	0.99	0.95
6	54.6	0.97	0.94	0.99	0.96	1.00	0.94	0.98	0.94
5	44.0	0.97	0.96	0.98	0.96	1.00	0.96	0.97	0.94
4	43.7	0.98	0.95	0.99	0.97	0.99	0.95	0.99	0.95
3	37.7	0.98	0.97	0.98	0.97	1.00	0.97	0.97	0.95
2	36.0	0.98	0.96	0.99	0.97	1.00	0.96	0.98	0.95
1	26.0	0.98	0.96	0.99	0.97	1.00	0.96	0.98	0.95

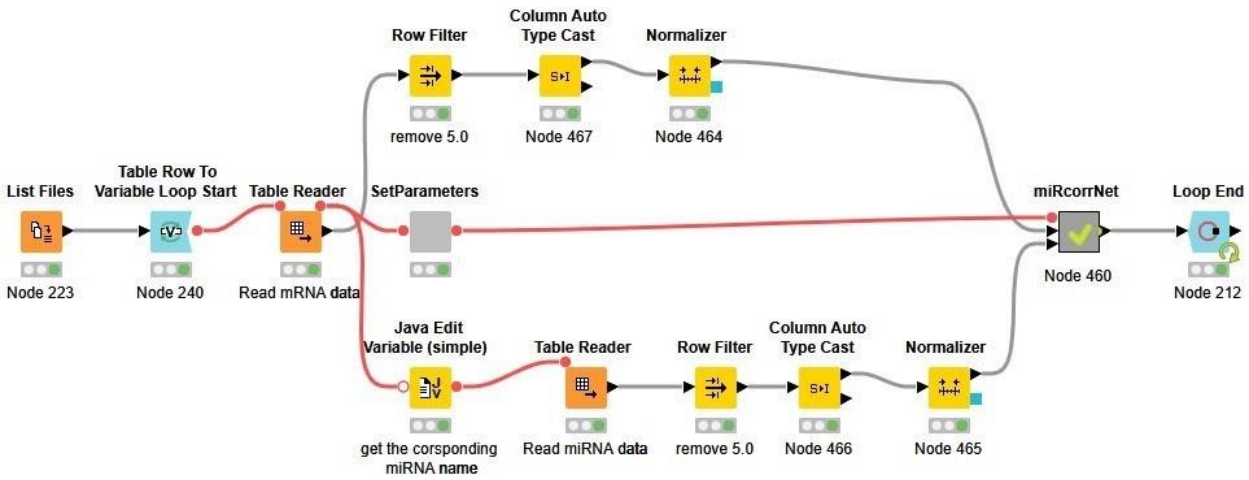
miRNA	p-value(Score)	Targets	#genes
hsa-miR-21-5p	0.000000	MME,FIGF,F8,IL33,MAB21L1,LOC572558,TGFBR2,RBMS3,MYCT1,CCDC46,TEF,ANXA1,HLF,DST,ADAMTS5,LHFP,CDC14C,LDB2,TXNIP,ABCA10,FMO2,PLSCR4,PDE2A	23
hsa-miR-10b-5p	0.000053	H2AFY,PYGO2,CCT3	3
hsa-miR-200c-3p	1.000000	CCDC46,LHFP,ERG	3
hsa-let-7c	1.000000	TIMM17A,NUSAP1,H2AFY	3

Gene	p-value(Score)
FIGF	9.71496E-50
NR3C1	4.73611E-46
F8	1.13558E-43
MME	1.8344E-37
NKAPL	2.53738E-28
LOC572558	1.26865E-22
CXORF36	3.03815E-22
MAB21L1	9.4826E-21
HLF	6.19181E-18
RBMS3	1.16879E-17
ADAMTS5	8.69323E-16
C18ORF34	3.96603E-15
TEF	3.54006E-14
KANK1	1.10046E-12
CDC14B	1.44556E-12
CHL1	3.4941E-12
CD34	1.0216E-11
SHE	1.85052E-11
IGSF10	3.63798E-11
DST	1.36597E-10

RobustRankAggreg R package developed by (Kolde et al., 2012)

miRcorrNet

Method	#Genes	ACC	Sen	Spe	AUC
maTE	7.48	0.96	0.94	0.96	0.98
miRcorrNet	141.1	0.96	0.94	0.97	0.98
svm-rfe	8	0.84	0.85	0.85	0.91
svm-rfe	125	0.96	0.97	0.95	0.98



	miRcorrNet Performance										
	BLCA	**BRCA**	**KICH**	**KIRC**	**KIRP**	**LUAD**	**LUSC**	**PRAD**	**STAD**	**THCA**	**UCEC**
#Grp	AUC	AUC	AUC	AUC	AUC	AUC	AUC	AUC	AUC	AUC	AUC
10	0.98	1.00	1.00	0.99	1.00	1.00	1.00	0.95	0.96	1.00	0.99
7	0.98	1.00	1.00	0.99	1.00	1.00	1.00	0.95	0.98	1.00	0.99
5	0.99	1.00	1.00	0.99	1.00	1.00	1.00	0.96	0.97	1.00	0.99
2	0.97	1.00	1.00	0.99	1.00	1.00	1.00	0.96	0.98	1.00	0.99
1	0.97	0.99	1.00	0.99	1.00	1.00	1.00	0.95	0.93	0.99	0.99

miRcorrNet prioritizes pan-cancer regulating miRNA

Prof. Christine M. Eischen, Ph.D.
Herbert A. Rosenthal Professor of Cancer Biology
Vice Chair, Department of Cancer Biology
Co-Leader, Molecular Biology and Genetics Program (SKCC)
Thomas Jefferson University,PA

Ramkrishna Mitra,PHD
Research Instructor
Department of Cancer Biology
Thomas Jefferson University, PA

Ramkrishna Mitra, Christine M. Eischen from Thomas Jefferosn Univeristy has contoributed to the Biological Interpretation of the Results

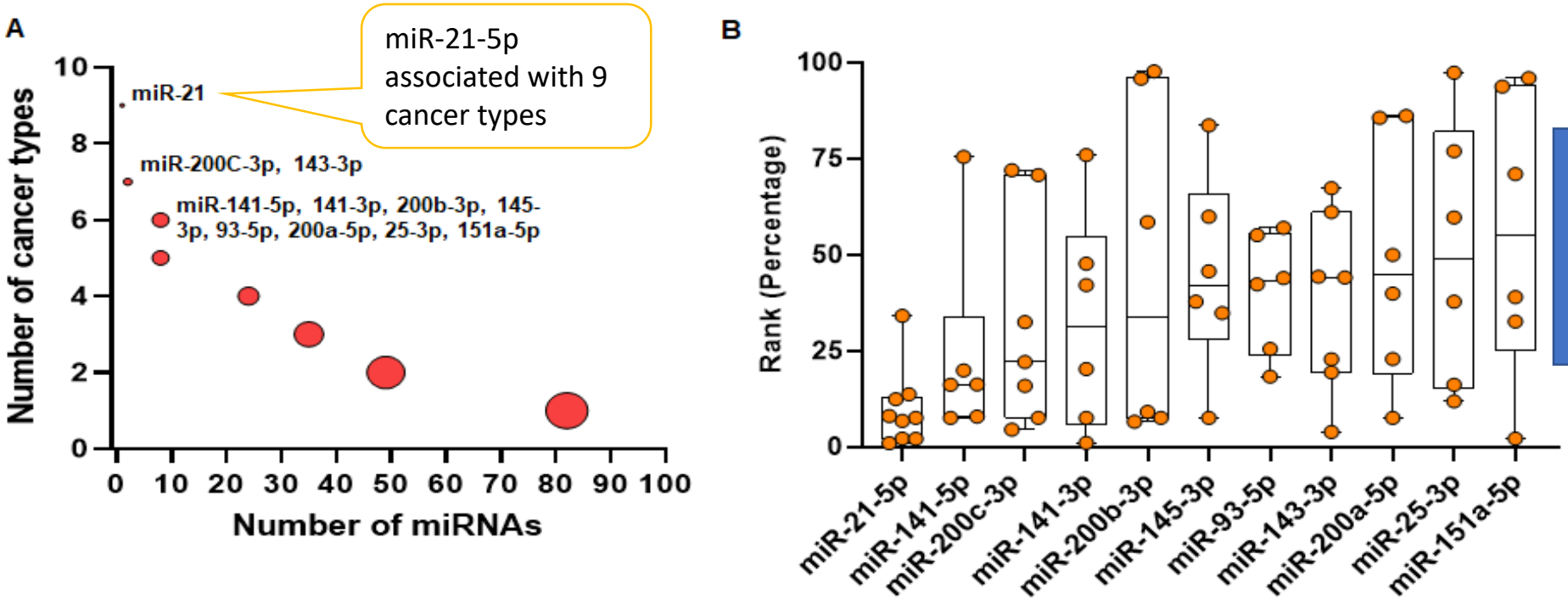
Jefferson
Philadelphia University + Thomas Jefferson University
HOME OF SIDNEY KIMMEL MEDICAL COLLEGE

Biological Interpretation of the Results

miRcorrNet Prioritizes Pan-Cancer Regulating miRNAs

As a test of miRcorrNet, we employed it to identify critical miRNAs in 11 TCGA cancer types. Among the >2500 mature miRNAs, detected in miRBase database (Release 22.1) (…Birgaoanu & Griffiths-Jones, 2019). miRcorrNet prioritized only a few (13 ~ 92, …f them as critical in a specific cancer. We investigated how recurrently these …e prioritized across the cancer types. We determined that 11 miRNAs had …e regulation across 6 or more cancer types (Fig. 4). Among these miRNAs, miR-…sociated with 9 cancer types, miR-200C and miR-143-3p were associated with 7 … and 8 other miRNAs had recurrent association with 6 cancer types (Fig. 4A). …as not only associated with the highest number of cancer types, but is also …ne of the top ranked miRNAs consistently across the cancer types (Fig. 4B). The …e based on the frequency score derived from the miRcorrNet algorithm. miR-21-…known onco-miRNA whose elevated expression is linked with suppression of …ssor genes associated with proliferation and apoptosis across numerous cancer …padhyay et al., 2010; Feng & Tsao, 2016). Moreover, diagnostic and prognostic …21-5p and its implication in drug resistance had also been observed in many …Zhang et al., 2012; Faragalla et al., 2012; Wang et al., 2014; Yang et al., 2015; … 2016; Gaudelot et al., 2017; Emami et al., 2018) .The literature-based evidence validated that miRcorrNet accurately predicted miR-21-5p as a critical pan-cancer regulator. The miRNAs miR-141-5p, miR-200C-3p, miR-141-3p, and miR-200b-3p were ranked as 2nd to 5th, respectively (Fig. 4B). Both mature strands of miR-141 were prioritized as top critical miRNAs in our study, and concordant dysregulation of miR-141-5p and miR-141-3p across cancer types has recently been observed (45). However, due to the historical belief that one mature strand is degraded during miRNA biogenesis, little is known about the coordinated regulatory roles of this 5p/3p pair. Here our study indicates that the miR-141 5p/3p pair mediates recurrent regulations across the cancer types, suggesting that they may be critical

miRcorrNet prioritizes pan-cancer regulating miRNA



miR-21-5p was not only associated with the highest number of cancer types, but is also regarded as one of the top ranked miRNAs consistently across the cancer types

More than 2500 miRNAs are available. Among there miRNAs, **miRcorrNet** prioritized only a few (13~92, median=3) as a critical for each specific cancer type.

A) Eleven miRNAs, potentially regulate 6 or more cancer types, are highlighted.

B) Ranks of these 11 miRNAs in individual cancer types are denoted by dots.

These miRNAs are sorted based on their median rank.

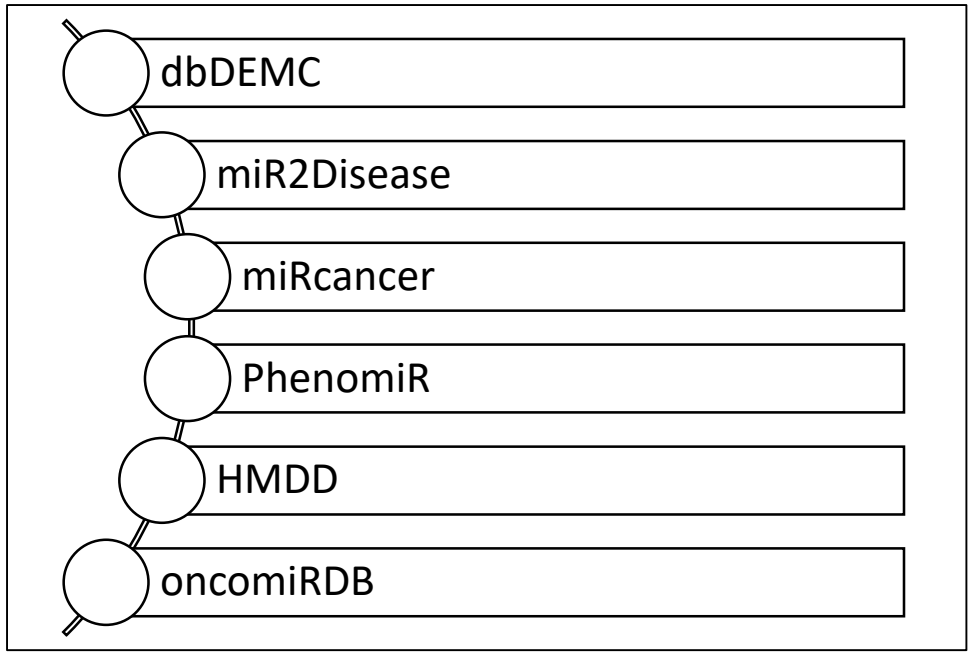
miR-21-5p is a well-known onco-miRNA whose elevated expression is linked with suppression of tumor suppressor genes associated with proliferation and apoptosis across numerous cancer types (Bandyopadhyay et al., 2010; Feng & Tsao, 2016).

Validation of miRNA-Disease Relations

miRcorrNet Findings

miRNA	p-value(Score)	Targets
hsa-miR-21-5p	8,42423E-33	RASGEF1C,SOX10,NOVA1,PCSK2,(
hsa-miR-22-3p	1,20936E-12	MEIS1
hsa-miR-16-5p	0,006982937	EVC,ZNF154,PPP3CB
hsa-miR-1976	0,011501451	FAM168B,AHNAK,ACOX2,PJA2,DI
hsa-miR-182-5p	0,125595903	CSGALNACT1,ACOX2,CD99L2,ARH
hsa-miR-576-5p	0,126381607	MID2,ZBTB4
hsa-miR-92a-3p	0,301933719	SOX10,NOVA1,AQP1,EVC,LRRK2,I
hsa-miR-26b-3p	0,92385325	SOX10,RRAGD,ARHGAP24
hsa-miR-15b-5p	1	EVC,SETD7,ARHGAP1,AXL,SERINC
hsa-miR-361-3p	1	DYNC1LI2
hsa-miR-15b-3p	1	TNFSF12,LRRK2,C10orf72,CRTAP,A
hsa-miR-361-5p	1	PDZRN3,TPD52L1,PJA2,SLC25A4,E
hsa-miR-484	1	EVC,LRRK2,MID2,MEF2A,CAPZA2,
hsa-let-7a-3p	1	NPR2,TSPAN18,CELF2,TNFSF12,GI
hsa-miR-23a-3p	1	RRAGD,STAT5B
hsa-miR-34a-5p	1	NFIX,ARHGAP24,SETBP1,ZHX3,NF

miRNA – Disease Relationship Databases



- To be able to validate our miRNA-Disease relationships we looked for the databases we found in the literature. As a result we saw that our findings are robust.

Validation of miRNA-Disease Relations

BLCA			BRCA		
miRNA Name	*Score*	*Evidence*	*miRNA Name*	*Score*	*Evidence*
hsa-miR-21-5p	7.32	dbDEMC, miR2Disease, miRCancer	hsa-miR-21-5p	9.66	dbDEMC, miR2Disease, miRCancer
hsa-miR-22-3p	4.67	miRCancer	hsa-miR-10b-5p	7.98	dbDEMC, miR2Disease, miRCancer
hsa-miR-148b-3p	4.06	dbDEMC, miR2Disease	hsa-miR-200c-3p	5.26	dbDEMC, miR2Disease, miRCancer

GeNetOntology: Gene Expression based Ontology Grouping and Ranking

GeNetOntology

Group and Rank (D)

Input: D Gene Expression Data with two-class labels

CreateGroups (D)

G= run Ontology grouping function on D

[$G = \{groups_i = [gene_{i1}, gene_{i2}, ..., gene_{ik}]\}$, $i = 1,...,n_t$

return G

RankGroups(G)

For each group t in G

R = {rank (t)}

[R is the collection of the groups and its ranks]

Gene Ontology Data

- Gene Ontology Data
  - Molecular Signatures Database

Subset of Gene Ontology	#Gene Sets
Biological Process	7573
Molecular Functions	1001
Cellular Component	1697

Gene Set	Genes
Mitochondrial Genome Maintenance	AKT3, PPARGC1A, POLG2, PARP1, DNA2, TYMP, FLCN, PRIMPOL, STOX1, SLC25A4, LIG3, MEF2A, MPV17, OPA1, MSTO1, SLC25A36, TOP3A, TP53, PIF1, SESN2, SLC25A33, MGME1, LONP1
Single Strand Break Repair	AP002495.1, ERCC8, PARP1, APLF, ERCC6, SIRT1, LIG4, APTX, TDP1, TERF2, TNP1, XRCC1
Small Nucleolar Ribonucleoprotein Complex Assembly	RUVBL2, PIH1D2, NUFIP1, SNU13, ZNHIT6, PIH1D1, SHQ1, TAF9, RUVBL1, NAF1, ZNHIT3

Results

#Groups	#Genes (Mean)	Accuracy (Mean)	Sensitivity (Mean)	Specificity (Mean)	Area Under Curve (Mean)
10	107.1	0.96	0.98	0.90	0.99
9	100	0.96	0.96	0.95	0.99
8	85.9	0.93	0.96	0.85	0.99
7	68.3	0.94	0.96	0.90	0.99
6	55.8	0.94	0.96	0.90	0.99
5	44.7	0.96	0.96	0.95	0.99
4	36	0.93	0.94	0.90	0.98
3	29.8	0.91	0.94	0.85	0.98
2	21.8	0.90	0.92	0.85	0.97
1	12.3	0.87	0.90	0.80	0.95

Results for GDS1962 and Molecular Function

Robust Rank Aggregation

Gene	p-value(Score)
LPL	0.000116606
PON2	0.000361272
BCHE	0.000378668
DNAJB1	0.000382702
CNN3	0.000539879
CXCR4	0.000553584
RAB3B	0.000643896
GTSE1	0.000683143
MYD88	0.000778579
GBP1	0.001069778

Robust Rank Aggregation

Cluster	p-value (Score)	Genes	#genes
GO_VITAMIN_D_BINDING	0.08	IRX5,CYP2R1,GC,VDR,CALB1,S100G,KL	7
GO_HYDRO_LYASE_ACTIVITY	0.08	CBSL,HTD2,CA5B,L3HYPDH,PARK7,UROC1,FAHD2B,ECI1,ECHS1,EHHADH,HACD2,ENO1,ENO2,ENO3,ALAD,FASN,FH,TGDS,CA14,GMDS,CYP2S1,HADHA,HADHB,HSD17B4,CA5BP1,IREB2,CA13,ENO4,HACD4,ACO1,ACO2,PCBD1,FAHD2A,APIP,HACD3,AUH,ECHDC2,ENOSF1,NAXD,ECHDC1,CA10,UROS,CA1,CA2,CA3,CA4,CA5A,CA6,CA7,CA8,CA9,CA11,CA12,ECHDC3,DGLUCY,PCBD2,CBS,HACD1,CDYL	59
GO_CARBON_OXYGEN_LYASE_ACTIVITY	0.08	CBSL,POLQ,HTD2,CA5B,L3HYPDH,PARK7,UROC1,FAHD2B,ECI1,ECHS1,EHHADH,HACD2,ENO1,ENO2,ENO3,ALAD,FASN,FH,TGDS,CA14,NEIL2,XRCC6,POLL,GMDS,CYP2S1,HADHA,HADHB,HSD17B4,CA5BP1,IREB2,CA13,ENO4,HACD4,ACO1,NTHL1,OGG1,ACO2,PCBD1,FAHD2A,APIP,HACD3,AUH,NEIL3,ECHDC2,ENOSF1,NAXD,ECHDC1,CA10,PTS,RPS3,ETNPPL,TPI1,UROS,XRCC5,CA1,CA2,CA3,CA4,CA5A,CA6,CA7,CA8,CA9,CA11,CA12,NEIL1,ECHDC3,DGLUCY,HMGA2,PCBD2,CBS,ALKBH1,HACD1,CDYL	74

GeNetOntology Results

Dataset	maTE				Our Approach			
	SE	SP	ACC	#G	SE	SP	ACC	#G
GDS1962	0.96	1.00	**0.98**	66.00	0.97	0.96	**0.96**	44.00
GDS2519	0.64	0.57	**0.61**	62.00	0.67	0.57	**0.61**	168.00
GDS3268	0.78	0.71	**0.74**	84.00	0.77	0.63	**0.71**	118.00
GDS3929	0.50	0.57	**0.54**	26.00	0.79	0.05	**0.58**	28.00
GDS2547	0.87	1.00	**0.83**	34.00	0.72	0.76	**0.74**	59.00
GDS5499	0.79	0.97	**0.88**	90.00	0.98	0.81	**0.93**	62.00
GDS3646	0.42	0.63	**0.53**	29.00	0.93	0.26	**0.73**	108.00
GDS3874	0.77	0.90	**0.84**	52.00	0.99	0.41	**0.77**	167.00

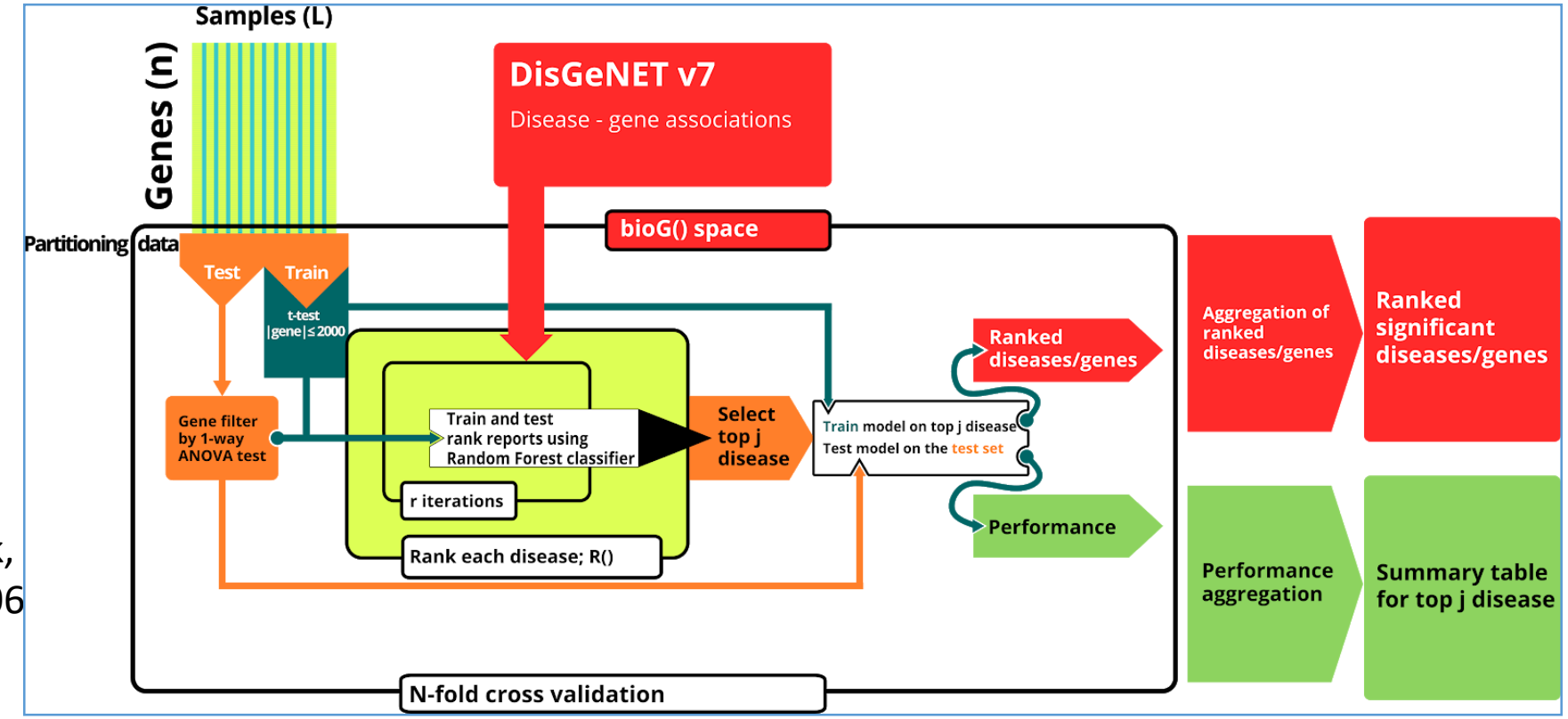
PredDisGeNetML - Discover Gene Biomarkers that Associated with Disease utilizing Knowledge-based Machine Learning

Computational genome Biology
Institute of Bioinformatics
Bangalore India

Abhishek Kumar

Institute of Bioinformatics, International Technology Park, Bangalore, 56006 Karnataka, India

Ayushman Kumar Banerje



PredDisGeNetML - Discover Gene Biomarkers that Associated with Disease utilizing Knowledge-based Machine Learning

Groups	BLCA	BRCA	KICH	KIRC	KIRP	LUAD	LUSC	PRAD	STAD	THCA	UCEC
	AUC (mean)										
10	0.989	0.999	1	1	1	1	0.998	0.961	0.996	0.997	0.993
9	0.985	0.999	1	1	1	1	0.999	0.964	0.997	0.998	0.992
8	0.988	0.999	1	1	1	1	0.999	0.961	0.996	0.997	0.991
7	0.988	0.999	1	1	1	1	0.998	0.961	0.996	0.997	0.993
6	0.988	0.999	1	1	1	1	0.998	0.961	0.996	0.997	0.993
5	0.983	0.999	1	1	1	0.999	0.998	0.958	0.996	0.997	0.993
4	0.986	0.999	1	1	1	0.998	0.998	0.96	0.997	0.997	0.992
3	0.984	0.999	1	1	1	0.999	0.998	0.961	0.994	0.997	0.993
2	0.981	0.999	1	1	1	0.998	0.997	0.956	0.993	0.997	0.993
1	0.97	0.998	0.99	1	0.999	0.993	0.995	0.954	0.988	0.997	0.98

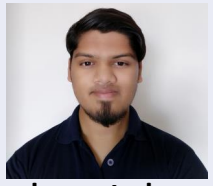
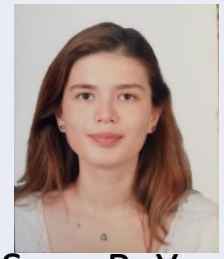
	#Genes										
10	192.86	311.27	144.41	170	158.6	196.21	147.42	311.73	376.37	313.21	137.91
9	182.16	296.84	124.99	142	146.99	182.8	137.45	295.43	355.46	299.19	128.45
8	165.45	277.03	107.86	138	134.37	163.86	127.99	281.34	338.54	274.81	118.35
7	143.8	260.08	93.4	137	118.26	143.65	115.88	260.56	313.58	250.87	107.98
6	125.99	228.28	82.61	123	107.65	122.58	99.66	242.22	286.55	236.2	92.45
5	108.43	192.01	66.06	116	89.86	109.7	85.11	217.93	255.51	208.56	79.57
4	92.76	151.31	47.39	116	76.38	90.94	65.92	194.99	215.33	182.48	68.97
3	70.59	118.21	26.92	113	59.11	63.91	47.39	168.24	168.61	147.5	57.39
2	45.95	88.75	11.83	113	36.73	43.95	27.47	132.19	113.39	105.82	38.34
1	17.45	45.65	3.78	22	14.75	17.87	12.36	83.18	58.48	57.3	19.47

PredDisGeNetML - Discover Gene Biomarkers that Associated with Disease utilizing Knowledge-based Machine Learning

KIRC			
Disease	**p-value**	**#genes**	**List of genes**
ANAPLASTIC THYROID CARCINOMA	0.000591	22	MME(1), ROS1(1), RAP1GAP(1), PLAU(1), CD70(1), CD82(1)...
CARCINOMA OF BLADDER	0.001181	103	SHBG(1), SLC14A1(1), KNG1(1), BAX(1), HOXB5(1), TYMS(1)...
COMMON ACUTE LYMPHOBLASTIC LEUKEMIA	0.001772	3	KNG1(1), MME(1), BCL2(1)
DUCTAL BREAST CARCINOMA	0.002363	13	TCF21(1), AFAP1L2(1), PLG(1), CD82(1), PADI2(1), BCL2(1)...
GASTRIC MUCOSA-ASSOCIATED LYMPHOID TISSUE LYMPHOMA	0.002953	2	BCL2(1), EPCAM(1)
INTRAHEPATIC CHOLANGIOCARCINOMA	0.003544	27	SHBG(1), BAX(1), TYMS(1), GPC3(1), OGG1(1), ROS1(1)...
LYMPHOMA, NON-HODGKIN	0.004135	44	BAX(1), SLC23A1(1), MME(1), TYMS(1), RIT1(1), OGG1(1)...
MALIGNANT NEOPLASM OF COLON STAGE IV	0.004725	7	TYMS(1), MYCN(1), KLK6(1), NDRG1(1), SELE(1), ALB(1)...
NEUROECTODERMAL TUMOR, PRIMITIVE	0.005316	14	SFRP1(1), PCSK2(1), MYCN(1), CAPS(1), ENO2(1), MTOR(1)...
PAPILLARY THYROID CARCINOMA	0.005907	75	BAX(1), PKHD1L1(1), MME(1), GPC3(1), PITX2(1), ROS1(1)...

Example output file showing top 10 disease from the RobustRankAggreg Node, for the KIRC dataset, where #genes column shows the count of genes in the List of genes column, the disease name with its corresponding p-value.

Ongoing Projects

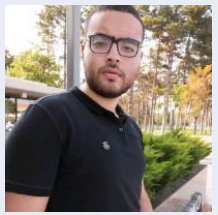
	Project Title	Status
 Gokhan Goy	miRcorrNet M3	
 Amhar Jabeer	Discover microNRA Biomarker that Associated with Disease utilizing Knowledge based Machine Learning	
	Optimization of Scoring function for SVM-RCE-R	
 Sena B. Yengec-Tasdemir	Normalization of the groups size effect	

ZEFAT ACADEMIC COLLEGE

ABDULLAH GÜL ÜNİVERSİTESİ

Burcu Bakir-Gungor

Ongoing Projects

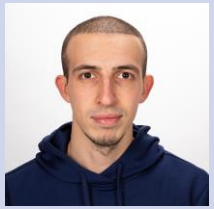
	Project Title	Status
 Ahmet Sayici	**GeNetOntology:** Gene Expression based Ontology Grouping and Ranking	First stage is submitted to ISMB 2021 conference
 Fatih Coşkun	**TextNetTopics**: Text Mining Based Topics Ranks	
 Fatma Ozdemir	**GeNetKEGG**: Gene Expression based KEGG PathWay Grouping and Ranking	
 Anas Nadifi	**MicroBiomeNet**: Machine Learning Analysis of Metagenomics Datasets: Colon Cancer Dataset	

ZEFAT ACADEMIC COLLEGE

ABDULLAH GÜL ÜNİVERSİTESİ

Burcu Bakir-Gungor

Ongoing Projects

	Project Title	Status
 Osman Bul	Discovering Potential Biomarkers of Type 2 Diabetes from Human Gut Microbiata via Different Feature Selection Methods	

Burcu Bakir-Gungor

Ongoing Projects

Khiam Research Group

MultiKOC : Multi one-class classifier based on K-Means clustering

Rank the groups based on distance measurements

Ensemble one-class machine leanring

Abhishek Kumar
Computational genome Biology
Institute of Bioinformatics
Bangalore India

Prediction of Candidate Disease Genes through Utilizing Knowledge based Machine Learning

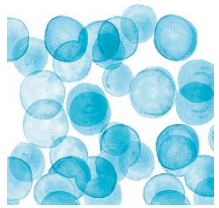
Ongoing Projects

European Cooperation in Science and Technology

Statistical and machine learning techniques in human microbiome studies

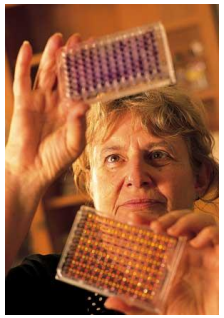
Review
Applications of machine learning in human microbiome studies: a review on feature selection, biomarker identification, disease prediction and treatment

Frontiers in Microbiology
Systems Microbiology



Jens Allmer
Professor for Medical Informatics and Bioinformatics at Hochschule Ruhr West University of Applied Sciences

Integration of Biological Knowledge (e.g. Pathways) to Classification Problem



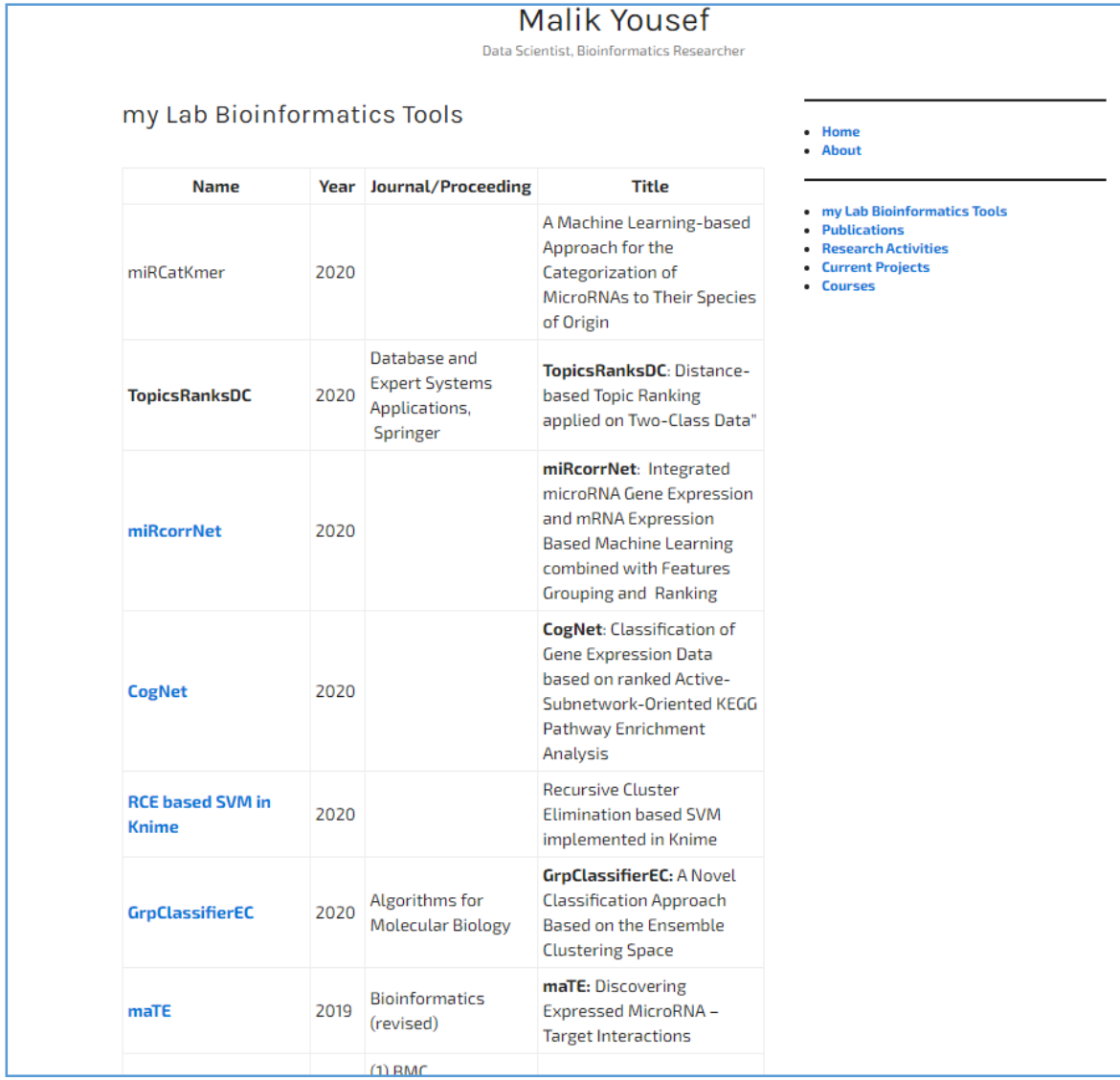
Louise C. Showe
The Wistar Institute Cancer Center

Developing SVM-RCE

Our Bioinformatics Tools

https://malikyousef.com/

https://github.com/malikyousef?tab=repositories



Name	Year	Journal/Proceeding	Title
miRCatKmer	2020		A Machine Learning-based Approach for the Categorization of MicroRNAs to Their Species of Origin
TopicsRanksDC	2020	Database and Expert Systems Applications, Springer	**TopicsRanksDC**: Distance-based Topic Ranking applied on Two-Class Data"
miRcorrNet	2020		**miRcorrNet**: Integrated microRNA Gene Expression and mRNA Expression Based Machine Learning combined with Features Grouping and Ranking
CogNet	2020		**CogNet**: Classification of Gene Expression Data based on ranked Active-Subnetwork-Oriented KEGG Pathway Enrichment Analysis
RCE based SVM in Knime	2020		Recursive Cluster Elimination based SVM implemented in Knime
GrpClassifierEC	2020	Algorithms for Molecular Biology	**GrpClassifierEC:** A Novel Classification Approach Based on the Ensemble Clustering Space
maTE	2019	Bioinformatics (revised)	**maTE:** Discovering Expressed MicroRNA – Target Interactions
		(1) BMC	

Prof. Jens Allmer

[1]Medical Informatics and Bioinformatics, Hochschule Ruhr-West University of Applied Sciences, 45407 Mülheim an der Ruhr, Germany

Prof. Larry M. Manevitz

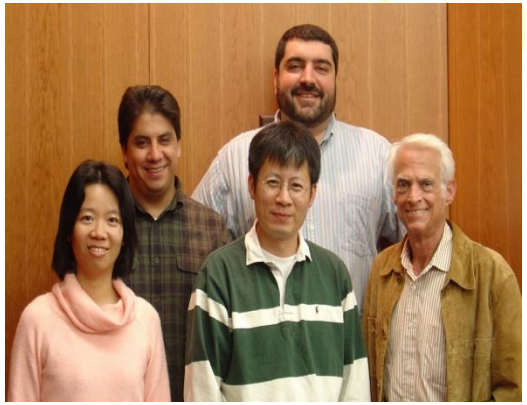
Department of Computer Science, Faculty of Social Sciences, University of Haifa, Israel

Mohamed Ketany

Prof. Louise C. Showe Lab

Molecular Biology

Computational



Rehman Qureshi,PHd

Associate Scientist - The Wistar Institute

Dr. Loai Abedallah

Management Information Systems, The Max Stern Yezreel Valley College Israel

Dr. Burcu Bakir-Gur

Department of Computer Engineering
Abdullah Gül University

Gokhan Goy
PhD Student

Amhar Jabeer
Undergraduate Student

Prof. Christine M. Eischen, Ph.D.
Herbert A. Rosenthal Professor of Cancer Biology
Vice Chair, Department of Cancer Biology
Co-Leader, Molecular Biology and Genetics Program (SKCC)
Thomas Jefferson University,PA

Ramkrishna Mitra,PHD
Research Instructor
Department of Cancer Biology
Thomas Jefferson University, PA

Prof. Ugur Sezerman Lab
Department of Biostatistics and Medical Informatics
Faculty of Medicine,
Acibadem University

Ege Ülgen, PhD Student