*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;

**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.

## 1. Introduction

Biological systems are massively complex and heterologous in nature. To resolve the mysteries behind complex biological systems,  large-scale studies are conducted which yielded massive volumes of biological data including the genetic variations associated with specific phenotypes. Currently, we encounter an -omics revolution in which genome, epigenome, transcriptome, and other -omics can be readily characterized. With advancements in various -omics approaches, it is
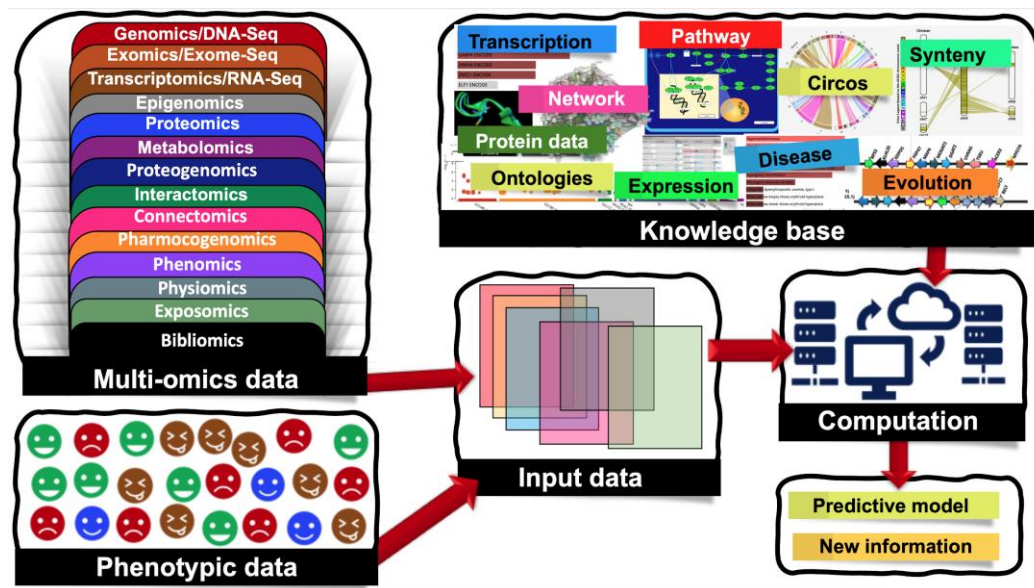
*Figure 1. Machine learning (ML) applications that combine multi-omics and phenotypic data. Multi-omics data are classified into the following groups: genomics/DNA-Seq - the study of the genetic material for an organism, it assesses DNA sequence and structural variations including single-nucleotide polymorphisms (SNPs), insertions and deletions, copy number variations (CNVs), and inversions; epigenomics - the measurement of DNA methylation, histone modifications (methylation, acetylation, phosphorylation, DP-ribosylation, and ubiquitination), and non-coding RNAs (microRNAs, long noncoding RNAs, small interfering RNAs); transcriptomics/RNA-Seq - the study of the transcriptome of an organism, exomics/exome-seq - the study of the exome of an organism (coding regions); proteomics - the study of the total proteins within an organism; metabolomics - the study of the total metabolites; proteogenomics - combined study of genomics and proteomics; interactomics - interactions between nucleotides, proteins and metabolites; connectomics - study of the connections, neural pathways in the brain; pharmocogenomics - the application of genomics to pharmacology; phenomics - observable phenotypes; physiomics - functional behaviour of an organism; exposomics - study of organism's environment and bibliomics( the literature concerning a topic).*

now possible to generate multi-omics data to  answer various biological problems. Nowadays, several types of -omics data are considered as depicted in Figure 1, and the numbers of different -omics data types are increasing day-by-day[1]. Also, there are complex cascades and interactions among different –omics data types. For example, genomic and epigenomic variations have the capacity to control or modulate the transcriptome and in turn affect the proteome. Here, epigenomics refers to the measurement of DNA methylation, histone modifications (methylation, acetylation, phosphorylation, DP-ribosylation, and ubiquitination), and non-coding RNAs (microRNAs, long noncoding RNAs, small interfering RNAs). Similarly, the epigenome of an organism refers to the entire collection of the molecules that modify the genome and control the genes to turn on and off. Since the epigenome shows how environmental factors influence the activity of genes, the study of the epigenome integrated with the study of the genome is crucial to fully account for phenomics. Accounting for such molecular deviations is crucial for making tangible improvements in biomarker analysis.

Traditional analyses attempted to untangle the molecular mechanisms of complex diseases using a single -omic dataset which contributed towards the identification of disease-specific mutations and epigenetic alterations. However, in the postgenomic era, it has been noticed that a single -omics dataset is not sufficient to explain disease hallmarks. It requires the combined analysis of various -omics datasets. Hence, recent studies are shifting towards multi-omics data analysis, where each of these different -omics data types are critical for deciphering molecular signatures of human diseases. Therefore, the integrated analysis of different data types has become a recent trend. For a holistic understanding of complex biological problems, it is becoming clear that integrations of different -omics data types are essential steps. However, it is a notorious task as handling heterogeneous  and noisy biological data is a challenging issue [2].

In addition to the 'omics' realm, another major reason for phenotypic differentiation is post-translational modifications (PTM). They can be both in the physiologically reasonable and the pathologically anomalous forms. Methods for bioinformatically incorporating PTM effects are emerging from the gradual improvement of sequence motifs or less directly from compensatory expression patterns that emerge when an organism seeks to correct for aberrant biochemistry arising from anomalous structural modifications.

Due to the recent advancements in next-generation sequencing and microarray technologies, the cost of obtaining the gene expression profile of a sample is rapidly decreasing and hence, expression profiling has become a routine protocol in biological laboratories. The high turnaround of expression data is also coupled by the massive increase in the use of the revolutionary RNA-Seq method [3]. It is best exemplified by large oncogenomic expression profiles hosted at The Cancer Genome Atlas (TCGA) [4]. Mutations are the core causative agents of diseases like different cancers [5] when coupled with  gene expression profiles. These datasets provide sufficient information to scientists and physicians for deciphering the disease mechanisms. It is becoming clear that proper design of the RNA-seq can be used for mutational profiling as well as expression profiling [6]. This information also enables the design of platforms to assist diagnosis, to assess patients' prognosis, and to create patient treatment plans. For instance, van't Veer et al.  had  collected gene expression profiling datasets of primary breast tumors derived from a cohort of 117 young patients [7]. Machine Learning (ML) with feature selection was used to unravel a gene expression signature, which served as a signal for distant metastases even divergent conditions like lymph node negative [7].

Data analysis approaches of gene expression profilings have evolved rapidly as there are massive shifts from DNA microarray to RNA-seq based profiling. The earlier methods involved clustering approaches and traditional ML approaches. Since a large volume of biological knowledge becomes available, in literature there are obvious shifts from the pure data-oriented approaches to biological domain knowledge based integrative approaches. This fact has triggered bioinformatics researchers to suggest and develop advanced tools that consider the emerging biological knowledge and hence, they exploit this knowledge for deep analysis of the data.  There are many resources of biological knowledge, such as textual knowledge as more and more literature emerges, different databases and repositories such as miRTarBase [8] for microRNA, DNA Sequence Databases, Immunological Databases, Gene Expression Omnibus (GEO), Proteomics Resources, Protein Sequence Databases, TCGA, Gene Ontology (GO) and others.

Most feature selection algorithms that are applied on gene expression data are based on statistics and ML. However, most of them neglect the biological knowledge of the data that could contribute to perform better feature selection.  R. Bellazzi and B. Zupan [9] discussed recent developments in gene expression-based analysis methods focusing on studies (like associations; classification), and implications (like reverse engineering of gene-gene networks and resulting phenotypes). Authors surveyed the clustering approaches that group the genes using different distance measures like Euclidean distance and/or Pearson's correlation. Moreover, incorporating biological knowledge in the clustering algorithm is a very challenging task. The GOstats package [10] allows to define semantic similarity between the genes via incorporating the GO [11]. An additional study by Kustra and Zagdanski [12] used incorporation of GO annotation to expression data by inducing a correlation-based dissimilarity matrix to derive GO-based dissimilarity matrix.

The flood of -omics data and the need for more informative results urge the need for integrative approaches. The book of [13] is the first book on integrative data analysis and visualization in this area. It outlines essential techniques for integration of data derived from multiple sources. It is one of the first systematic books that overview the issue of biological data integration using analytical approaches. The book provides a framework for the creation and implementation of integrative

analytical methods for the study of biological data on a systematic scale. Additionally, a recent review [2] describes the principles of biological data integration along with different approaches and methods indicating the importance of utilizing ML for biomedical datasets. However, to the best of our knowledge, in the literature, there is no comprehensive survey on biological domain knowledge-based feature selection methods, except from the study of Perscheid et al [14] that compares the performances of traditional gene selection methods against integrative ones. Moreover, authors also proposed a straightforward method to integrate external biological knowledge with traditional gene selection approaches. They introduced a framework for the automatic integration of external knowledge for selected genes and their evaluation. Herein, we aimed to provide a selective review on such gene selection approaches. In this regard, for the analysis of gene expression datasets, we present the traditional and integrative gene selection approaches in Sections 2.1. and 2.2., respectively.

On the other hand, multi-dimensional biological data is another challenge as these data are often derived from limited numbers of samples because of the associated costs of biological data generation. This is an example of the 'curse of dimensionality' problem, as initially reported by Bellman [15] in 1961. It means that the dimensions of gene features and/or functional parameters are critical input variables and there are requirements of a minimal number of samples for estimation of an arbitrary function, where an increase in sample size improves the chance of function prediction. In Section 3 of this manuscript, we present a prospective solution to this problem by defining a novel grouping of features and estimating their contribution to the machine learning model for two-class classification problems. Once we evaluate the methods that select the features using a classifier in a traditional way in Section 3.1; we present integrative approaches that incorporate biological domain knowledge into ML to group and rank the genes in Section 3.2. In Section 4, we conclude our review with discussions and future prospects.

## 2. Gene Selection Approaches for Gene Expression Datasets

Gene selection approaches for gene expression datasets can be mainly categorized into two classes as traditional gene selection and integrative gene selection. While traditional gene selection approaches are solely based on statistical and computational analyses of the expression levels, integrative gene selection approaches incorporate domain knowledge from external biological resources during gene selection.

### 2.1. Traditional Gene Selection

Traditional gene selection approaches are heavily based on statistical and computational analyses of the actual expression levels. Recent reviews have summarized various methods for describing the selection process of disease specific features from large gene expression datasets [16], [17]. Primarily, these approaches are classified into three major classes as (i) filtering-based, (ii) wrapping-based, and (iii) embedding-based approaches. Briefly, the filtering approaches are based on F-statistic (ANOVA, t-test, etc.), not based on ML. Wrapping-based approaches are primarily learning techniques and these are used for exploration of usefulness of features. Whereas, embedding-based approaches are combining the feature selection and the classifier construction. Wei Pan carried out a comparative study on different filtering methods in [17] and he summarized similar and dissimilar points among three main methods (namely t-test method, regression modeling approach and mixture model approach).

Additional comparisons of filtering techniques are available in [16]. I. Inza [17] also carried out a comparison between filter metrics and wrapper sequential search procedure, which are applied on gene expression datasets. Also, hybrid, and ensemble approaches, which combine multiple

approaches are two additional categories of gene selection. Cindy et al [14] presents an overview of the recent gene selection methods, where each method is classified according to these five categories.

The traditional gene selection approach has several drawbacks. For example, the filtering approach evaluates the significance of each gene individually without considering the relationships and the interactions between the genes. Although the wrapping-based approaches can find the optimal set, yet it might be specific to the model used, such as SVM, decision trees or other models. In other words, it might be overfitting the data [18]. The main disadvantage of such methods are their difficulties for biological interpretation and they are unlikely to generate new biological knowledge.

## 2.2. Integrative Gene Selection

Although the traditional gene selection approaches became popular for a long time, they have several drawbacks when one needs to precisely identify the underlying biological processes. Alternatively, integrative gene selection approaches incorporate domain knowledge from external biological resources during gene selection [9], [18], which improves interpretability and predictive performance. One of the widely used external ontology resources is the Gene Ontology (GO) [19] which provides i) cellular component (CC), ii) molecular function (MF), and iii) biological process (BP) terms for the products of each gene. GO captures biological knowledge in a computable form that consists of a set of concepts and their relationships to each other. The first attempt to integrate biological background into statistical analysis/ML analyzes was to incorporate Gene Ontology (GO) [19] in clustering gene expression data [10]. Another widely used external ontology resource is Kyoto Encyclopedia of Genes and Genomes (KEGG), which is a pathway knowledge base providing manually curated pathways [20]. Yet another widely used external biological resource is DisGeNET, which is a meta knowledge base on gene-disease-variant associations [20].

One example of integrative gene selection approach is proposed by Qi and Tang, where they utilize the power of biological information contained in GO annotations to rank the genes [21]. The algorithm is designed in an iterative manner that starts by applying Information Gain (IG) to compute discriminative scores for each gene. The genes that have a score of zero are removed from the analysis. The second step is to integrate the biological knowledge, which is achieved by annotating those surviving genes with GO term. The third step is to score the GO terms as the mean of their associated genes' discriminative scores, which were computed before using IG. The final gene set is created as follows: Starting from the highest ranked GO terms, the genes with the highest discriminative scores are chosen. These genes are removed from the annotated genes and this procedure is repeated until the final gene set is complete. Using multiple cancer datasets, Qi and Tang showed that their proposed method can achieve better results, as compared to using IG only.

Another example of integrative gene selection approach is SoFoCles [22], which uses GO terms to find semantically similar genes. In order to assign a discriminative score to each gene, SoFoCles utilizes a classic filter approach like χ2, ReliefF, or IG. The initial set of candidate genes are composed of the top n ranked genes. Genes receive a similarity score based on their associated GO terms. Then, the genes which have a high similarity score, i.e. the genes that are semantically very similar to the candidate genes, are added to the set of candidates. The experiments conducted on SoFoCles showed that the incorporation of biological knowledge to the gene selection process improves the results.

Yet another study by Fang et al [18] combines KEGG and GO terms with IG . The authors initially apply IG on the dataset as the filtering step and then check the GO and KEGG annotations of the remaining genes. Then, the authors use association mining and calculate the interestingness of the frequent itemsets by averaging the original discriminative scores (from IG) of the included genes. The final gene set is generated via selecting the highest ranked genes from the top n frequent itemsets. They evaluated this method using GO, using KEGG, and using both terms against IG only and

against Qi and Tang's approach. Although their proposed approach slightly increased the overall accuracy, the main advantage of this approach was that it used a much lower number of genes.

The integrative gene selection approach that is proposed by Raghu et al. [23] makes use of KEGG, DisGeNET, and further genetic meta information [20]. In their approach, for each gene, i) the importance score and ii) the gene distance metrics are computed. The importance score is calculated via combining a gene-disease-association score from DisGeNET with the gene expression levels in the data. The gene distance is defined as the physical distance between two genes (in terms of their chromosomal locations) and their associations to the same diseases. Both of the scores (importance score and gene distance) are later used to find maximally relevant and diverse gene sets. As compared to variance-based gene selection techniques, the use of the top n genes according to the importance score resulted in slightly better performance in predictive modeling task.

The integrative approach of Quanz et al. aims to map genes into KEGG pathways and then uses these pathways as features for further pattern mining [24]. In their approach, they make use of a global test to extract KEGG pathways which are related to the phenotypes of a dataset. In their feature extraction step, the genes in each pathway are then transformed into one single feature by applying mean normalization or logistic regression. In this way, the data is represented as the number of pathways, which can be considered as a feature reduction step and it provides dramatic reduction. For instance, for the diabetes data, 17 pathways, out of approximately 300 pathways are selected and hence, for the classification task, the dimensionality is reduced from 22,283 to 17. Even though this approach was not tested on multiclass problems like cancer (sub-) type classification, the experiments on binary classification problems showed an improved performance over different traditional approaches .

Mitra et al. adopted Clustering Large Applications based upon RANdomized Search (CLARANS) method to feature (gene) selection problem via utilizing biological knowledge [25] . Their reduced feature set is composed of gene clusters, which are the medoids of biologically enriched sets. Later on, the authors attempted to use a fuzzy clustering technique instead of CLARANS, and developed a technique called FCLARANS for feature selection [26].

In [27], authors proposed an integrative gene (feature) selection approach based on sample clustering technique, which utilizes gene annotation information from GO. On the generated gene-GO term matrix, they applied Partitioning Around Medoids clustering. In their method, the optimal number of clusters (k) is chosen by comparing their silhouette index values. For the selected k number of clusters, the medoids are used as the selected gene subset. They reported that the integration of biological knowledge during the gene selection process not only reduces the dimensionality of the feature space, but also increases the accuracy of sample classification.

The related studies that are presented until this point are highly specific to a single knowledge base, e.g. KEGG pathway or GO terms. On the other hand, Perscheid et al [14] proposed an approach that can flexibly combine traditional gene selection approaches with several knowledge bases. They comparatively evaluated the performance of traditional gene selection approaches with integrative gene selection approaches. Their study concluded that the integration of external data especially improves on simple traditional filter approaches, e.g. Information Gain. Once external biological data is integrated, such traditional filter approaches become compatible with more complex machine learning approaches at very similar classification accuracies, but far less computational running times and a more transparent and thus interpretable computation processes.

The above-mentioned studies proposed predictive models, but most of the time, instead of obtaining high predictive accuracies in these models, the scientists are curious about the biological meaning of the predictive model. The 'black box' nature of the predictive model can hamper its interpretation.

The information excerpted from the model may require further processing and careful interpretation with corresponding biological knowledge may be needed. The interpretation of the complicated cases may be quite challenging and such an interpretation may currently be out of reach. Although the joint analysis of multiple biological data types has the potential to enlighten our understanding of complex biological phenomena; the data integration is challenging due to the heterogeneity of different data types. For example, an expression profile, as obtained from a transcriptomic study, is a vector of real values and the length of a vector is equal to the number of genes in the genome. But, the genetic variants as obtained from a genomic study, are categorical and they have different vector lengths. While different studies [1, 4] proposed several strategies for data integration; the best practices about which -omics data types can be integrated and how to integrate these biological data are still needed.

Feature selection and discovering the molecular explanation of diseases describe the same process, where the first one is a computer science term and the second one is used in the biomedical sciences. In 2007, Yousef et al. proposed a new feature selection method, Support Vector Machines-Recursive Cluster Elimination (SVM-RCE), to group/cluster genes for gene expression data analysis. This study invented the "Recursive Cluster Elimination '' phrase for the first time in the machine-learning domain and introduced it to the computational community. Hence, this study became a pioneer study in this field. Interests in this approach have increased over time and several studies have successfully applied the SVM-RCE approach to identify the features/genes that are directly associated with a disease/condition [28]. This growing interest is based on the reconsideration of how feature selection in biological datasets can benefit from incorporating the biomedical relationships of the features in the selection process. The usefulness of SVM-RCE then led to the development of maTE [29], which uses the same approach based on the interactions of microRNAs (miRNA) and their gene targets. Also, in literature, the biological information buried in genetic interaction networks are utilized for classification studies. For example, SVM-RNE (SVM with Recursive Network Elimination) integrates network information with recursive feature elimination based on SVM [30]. It is shown that SVM-RNE has good performance and also improves the biological interpretability of the results. Studies similar to SVM-RCE and SVM-RNE were later carried out by different groups [31], [32] , which indicates the importance and the merit of the SVM-RCE approach. The study of [33] has a slightly modified SVM-RCE algorithm on the disease state prediction step. Additionally, they used the already invented term of "recursive cluster elimination".

The study of Zhao, X. et al. [34] has used the SVM-RCE tool for comparison and used expression profiles for identifying microRNAs related to venous metastasis in hepatocellular carcinoma. Another similar study to SVM-RNE is carried out by Johannes M. et al. [35] for integration of pathway knowledge into a reweighted recursive feature elimination approach for risk stratification of cancer patients. A recent tool SVM-RCE-R [36] is an updated version of SVM-RCE, which is implemented in Knime [37], uses random forest classifier with additional important features such as suggesting a new approach of ranking the clusters.

The term of "knowledge driven variable selection (KDVS)" is a similar term to "integration of biological knowledge", and both of them are used in the process of feature selection. An additional similar study that applied KDVS to SVM-RNE is presented by [38], in which the authors proposed a framework that uses a priori biological knowledge in high–throughput data analysis.

The RCE algorithm [28] considers similar features/genes and applies a rank function to the feature group. Since it uses k-means as the clustering algorithm, we refer to these groups as clusters, but it could include other biological or more general functions combined with the features, as was suggested in several studies[29], [30] . In the original paper of SVM-RCE, the contribution to the accuracy is achieved in distinguishing specific classes for ranking the clusters. The data for that ranking is divided into training and testing with the data represented by each gene/feature being

assigned to a specific cluster of features. The rank function is then applied as the mean of m times repeats of the training-testing performance while recording different measurements of accuracy (sensitivity, specificity etc).

In Table 2, we summarize the specifications, advantages and disadvantages of the presented integrative gene selection approaches.

## 3 Grouping and Ranking of the Genes for Classification Problem

The genes that are involved in the same biological process are likely to be co-expressed [39]. Therefore, one potential way of discovering gene function is to group genes with a similar expression profile. Thus, different clustering algorithms [40] were considered to perform the grouping step. This was the first approach and more advanced approaches that use biological information in order to group the genes are later proposed. In this section, we will introduce a generic approach for grouping that is accompanied by ranking and classification. The presented model is used by different studies and other similar studies are still ongoing.

The main aim of the generic approach is to search and determine significant groups/clusters of features based on one or more biological grouping functions (will be referred as bioF() throughout the rest of this paper), that are integrated with ML algorithms. The generic approach is presented in Figure 2. The advantage of those systems is that the grouping of the genes/features is in the hand of the researcher that is actually based on available biological knowledge. The researcher will provide how genes or features should be grouped and then the algorithm will proceed to score and rank those groups in terms of the classification problem. The final model will be built from the top n groups according to the researcher's settings. The outcome of the algorithm is different from the traditional current approaches (like SVM-RFE [41]), where the algorithm takes as input the data of
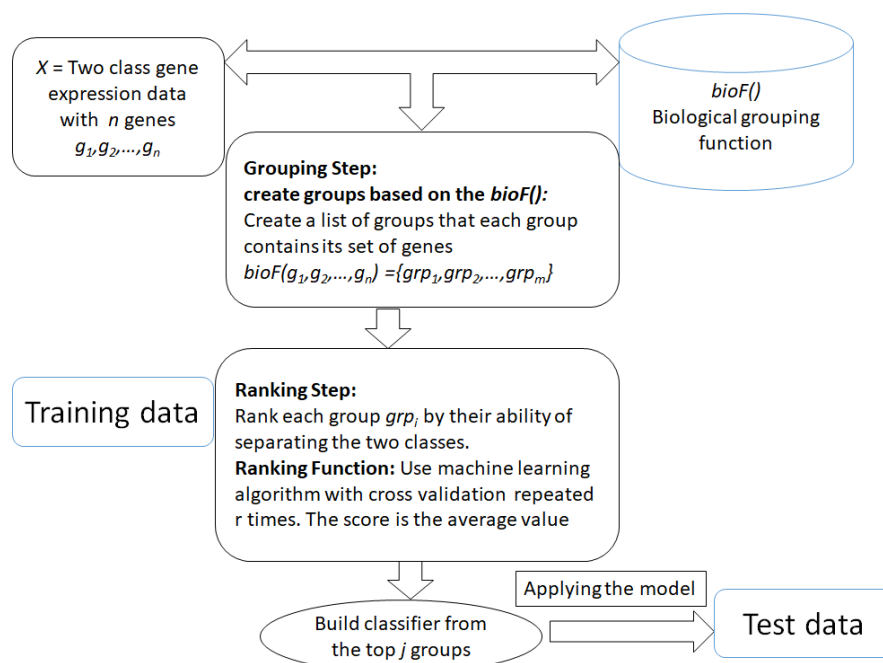


*Figure 2. The generic framework of the algorithm that is based on biological integration for grouping, ranking and classification.*

gene expression with class labels. Then the outcome is just a list of significant genes that are able to distinguish the two-classes. With the integration framework, the researcher will get a more informative list of significant groups/clusters with its genes list that is able to distinguish the two-classes. Additionally, the researcher can use the computational approach of grouping that is based

on clustering approaches such as k-means or others and specify different measurements for ranking the groups/clusters based on his interest and his research aims. The outcome of the algorithm will be more specific to the researcher's interest.

The generic approach mainly consists of two main components. The first component is the grouping step relying on the bioF() function that is based on biological knowledge to group the genes into groups. For example, bioF() might be disease related genes, then the function will group the genes into groups where each group is associated with one disease. Another possibility is grouping the genes that are targeted by specific miRNAs, such as in the maTE [29] tool. One interesting use of bioF() is that it allows to create different biological groupings, such as creating groups related to miRNA, groups related to disease, groups related to KEGG pathways, and others. However, the grouping can also be based on clustering algorithms such as k-means, as suggested in SVM-RCE [28] for grouping correlated genes. Similarly, SVM-RNE [30] incorporates another tool GXNA [42] to create the groups. GXNA utilizes gene expression profiles and prior biological information to suggest differentially expressed pathways or gene networks.

## 3.1. Traditional Approach of Feature/Gene Selection Using a Classifier

There are many classifiers that were used to fit the data in order to rank the features and perform the process of features selection. The simple one is the linear model, where the coefficients of the variable/features can serve as a measurement for the features rank. We will be describing the first approach that suggests the RFE procedure. RFE refers to recursive feature elimination. The approach uses SVM as the linear model.

### SVM-RFE (Support Vector Machines with Recursive Feature Elimination)

Let's assume that we are given a set of points called S. S consists of m points $x_i \in R^d$ in dimension d. Let's assume that we have two-class labels denoted by $y_i \in \{-1, +1\}$. We call S a linear separable set if there is a hyperplane of equation $w \cdot x + w_0 = 0$ (we refer to it as hyperplane $(w, w_0)$) that separates the points with label +1 from the points with label -1. The signed distance $d_i$ of a point $x_i$ to the separating hyperplane $(w, w_0)$ is given by $d_i = (w \cdot x_i + w_0)/||w||$.

For simplicity, let's define the optimal separating hyperplane to be $f(x) = w_0 + w_1 x_1 + w_2 x_2 + ... + w_d x_d$, where the $(x_1, x_2, ... x_d)$ is the features and $w = (w_1, w_2, .., w_d)$ is the corresponding weights. SVM is actually the solution of finding the optimal linear function, as developed by Vapnik [43]. It is obvious that the contribution of features with lower weights is non-significant to the sign of the $f(x)$. So one can consider removing those features in order to perform dimension reduction of the feature space.

SVM-RFE, which stands for support vector machines with recursive feature elimination, is firstly introduced by Isabelle Guyon et al [41] and applied to gene expression data. The primary goal of SVM-RFE is to use SVM (linear svm) to compute the weights of the features. The weights are actually the ranks of the features. SVM-RFE performs an iterative step to remove features with low ranks. The RFE procedure can be described as:

1. Train the classifier on the given data.
2. Assign rank for each feature as its weight.
3. Remove one feature or percentage (10%) with the smallest weight.
4. Repeat steps 1-3 until reaching a predefined number of genes.

Different studies [44] have emerged as extended variations of the original SVM-RFE algorithm. However, SVM-RFE has some limitations that other studies have reported to suggest an

improvement approach. One such limitation is that SVM-RFE is designed as a greedy method and tries to find out superlative possible combinations leading to binary classification, where these combinations may not be biologically significant. To overcome this limitation, a novel feature selection algorithm, sigFeature [45] based on SVM and t statistic was developed.

## 3.2. Biological Domain Knowledge Based ML Approaches

ML is becoming a very powerful computational approach in the field of bioinformatics. In this respect, the first step in utilizing ML is to apply unsupervised and supervised algorithms on biological data. However, in the era of big biological data, we come up with emerging approaches that integrate biological knowledge with ML. A recent review [46] on ML and complex biological data discuss the challenges and hurdles of the analysis and discovery from complex biological data. They predict that in the very near future, more researchers will be interested in applying ML to complex biological data. In this section, we will review different approaches that consider the biological structure or knowledge for the process of feature selection. The following approaches follow the generic approach presented in Figure 2.

### SVM-RCE (Support Vector Machines with Recursive Cluster Elimination)

SVM-RCE [28], [36] is based on the concept of grouping and ranking where k-means clustering algorithm is used for performing the related grouping of bioF(). Correlated genes are hypothesized to have similar biological functions. Then the rank component is applied to assign a score for each cluster, indicating its significance in terms of the classification of the two given classes. In order to perform the rank component, each cluster is considered by representing the data based on the genes that belong to it while keeping the class labels of the original data. Now the data is transferred to cluster genes representation. The rank component performs internal cross-validation and aggregates the performance outcome as the score of the cluster. The RCE procedure is applied to remove the least ranked groups. SVM-RCE results show that the classification accuracy is superior to other approaches suggesting that the classification results are more interpretable and it creates new hypotheses for future investigation.

### SVM-RNE (Support Vector Machines with Recursive Network Elimination)

SVM-RNE [30] is an extended version of the SVM-RCE approach that uses the tool GXNA [42] as the bioF() for grouping the genes into sub-networks of genes. Then a similar procedure of ranking is applied as described in SVM-RCE. SVM-RNE also performs the recursive elimination procedure by ranking firstly all the groups and then removing the least significant groups. The algorithm proceeds by applying again the GXNA to suggest groups. This process is repeated until satisfying some predefined constraints on the number of groups.

### maTE

Disease development mechanisms mainly involve changes in the transcript levels and protein abundance. MicroRNAs (miRNAs) are instrumental in regulating the gene expression and hence, they affect transcript levels and protein abundance. The fact that microRNAs target more than one mRNA serve us to group the genes into groups where each group consists of the list of genes targeted by a specific microRNA. In other words, the *bioF()* biological grouping function here is the biological association between microRNA and its set of targets. The *bioF()* grouping function is based on the database mirTarBase [8]. mirTarBase has accumulated more than three hundred and sixty thousand miRNA-target interactions. Thus a novel approach called maTE [29] is developed.

Table 1 presents partially the result of applying *bioF()* on mirTarBase. Additionally, it performs a computation procedure in order to score/rank the importance for each group for the classification tasks.

| MicroRNA group name | Target Genes List |
|---|---|
| HSA-MIR-147A | VEGFA, ACVR1C, MCM3, NDUFA4, PSMA3, HIF3A, SLC22A3, MCM3, NDUFA4, PSMA3, HIF3A, VEGFA, ACVR1C, MCM3, NDUFA4, PSMA3, HIF3A, SLC22A3 |
| HSA-MIR-18B-5P | ESR1, MDM2, CTGF, TNRC6B, HIF1A, SMAD2, FOXN1, IGF1, IGF1, CTGF, HIF1A, SMAD2, FOXN1, ESR1, MDM2, CTGF, TNRC6B, HIF1A, SMAD2, FOXN1, IGF1, IGF1 |
| HSA-MIR-19B-3P | BACE1, PTEN, PTEN, PTEN, ATXN1, HIPK3, ARID4B, MYLIP, ESR1, KAT2B, SOCS1, BCL2L11, BCL2L11, TGFBR2, TGFBR2, BMPR2, BMPR2, TLR2, PPP2R5E, PPP2R5E, CYP19A1, GCM1, HIPK1, SMAD4, MYCN, MXD1, BCL3, DNMT1, TNFAIP3, PKNOX1, MTUS1, PITX1, PTEN, PTEN, PTEN, ATXN1, ESR1, NCOA3, KAT2B, SOCS1, TGFBR2, BMPR2, CUL5, TLR2, HIPK1, MXD1, BCL3, TNFAIP3, MTUS1, PITX1, BACE1, PTEN, PTEN, PTEN, PTEN, ATXN1, HIPK3, ARID4B, MYLIP, ESR1, NCOA3, KAT2B, SOCS1, BCL2L11, BCL2L11, TGFBR2, TGFBR2, BMPR2, BMPR2, CUL5, TLR2, PPP2R5E, PPP2R5E, CYP19A1, GCM1, HIPK1, SMAD4, MYCN, MXD1, BCL3, DNMT1, TNFAIP3, PKNOX1, MTUS1, PITX1 |
| HSA-MIR-210-5P | CFB |

*Table 1. Example of microRNA and their targets list*

The input to the maTE tool is the gene expression data, list of microRNAs and its target genes. The main component of the tool is to produce a group of genes based on the miRNA target information and then rank each group by applying random forest with cross validation which is repeated r times. The average of the accuracy for each iteration is actually the rank for a specific group. Then the groups are ranked according to the rank values. The model will be built considering the genes on the top j groups. The default value of *j* is 2. We apply the training part of the data a t-test statistics in order to remove noisy genes. The test part of the data is used in order to estimate the performance of the tool.

## CogNet

CogNet tool is a classification of gene expression data based on ranked active-subnetwork-oriented KEGG pathway enrichment analysis [47]. CogNet is based on biological knowledge as a function for grouping the genes for the task of ranks and classification. The pathfindR tool serves to be the biological grouping function allowing the main algorithm to rank active-subnetwork-oriented KEGG pathway enrichment analysis [48]. Even Though, the main aim of the current tools is not improving results of any existing tools, the performance of the CogNet outperforms a similar to maTE while getting similar performance to SVM-RCE. CogNet was tested on 13 gene expression datasets that include a variety of diseases.

CogNet provides a list of significant KEGG pathways including its genes that are able to separate the classes of the data. The list would serve the biology researcher for deep analysis and better interpretability of the role of KEGG pathways in the data or the case that is being studied. As a future work, we would develop CogNet to explore the effectiveness of different combinations of the KEGG pathways in the data. In the current version, we treat each KEGG pathway individually.

## miRcorrNet

Due to the advances in technology, both mRNA and microRNA expression profiles can be generated allowing integrative analysis aiming to uncover the functional effects of RNA expression in complex diseases, such as cancer. Most of the approaches that integrate miRNA and mRNA are based on

statistical methods, such as Pearson correlation, combined with enrichment analysis approaches. In this study [49], a novel tool called miRcorrNet, which performs machine learning-based integration to analyze miRNA and mRNA gene expression profiles. miRcorrNet groups mRNA genes based on their correlation to miRNA expression. Then, these groups are subject to a rank function for classification. We have tested our tool on TCGA data miRNA-seq and mRNA-seq expression compared to other tools. The performance results show that the tool is working as good as other tools in terms of accuracy measurements reaching AUC above 95%. Moreover, we conducted a deep biological analysis to explore the list of significant miRNAs. Accumulated results suggest that miRcorrNet is able to accurately prioritize pan-cancer regulating high-confidence miRNAs.

| Tool Name | Incorporated Biological Knowledge | Methodology | Advantage/Disadvantage | Ref |
|---|---|---|---|---|
| N/A | GO | Rank the genes uses information gain (IF)incorporated with GO terms | **A:**The novelty of this work is to evaluate genes based on not only their individual discriminative powers but also the powers of GO terms that annotate them**.** | [21] |
| N/A | GO | $\chi 2$, ReliefF, or IG | **A:** including biological knowledge in the gene selection process improves results. | [22] |
| N/A | Combines KEGG and GO terms | Utilizes graphical causal modeling IG as an initial filter search for GO and KEGG annotations Frequent Items | **A**: Method is capable of intelligently selecting genes for learning effective causal networks. **D:** no significant improvement in accuracy | [18] |
| N/A | KEGG, DisGeNET, and further genetic meta information | Gene-disease-association score from DisGeNET Gene distance metrics | | [23] |
| N/A | KEGG pathways | Uses these pathways as features for further pattern mining | **A:** Reduce the dimension of the data by transforming to KEGG feature space **A**: Improved performance over different traditional approaches | [24] |
| N/A | Gene ontology (GO) | RAN-domized search (CLARANS) | **A**: Reducing the dimension dramatically | [25] |
| SVM-RCE | Genes related are correlated | SVM and K-means | **A:** Discover significant of clusters **D:** Might loos important | [28], [36] |

| | | | genes because they were in lower ranked clusters | |
|---|---|---|---|---|
| SVM-RNE | GXNA for creating sub-networks from gene expression | SVM, GXNA | **A:** Reducing the dimension of the data by considering sub-networks <br> **D:** The sub-networks are created as a prediction of the gene expressions data | [30] |
| maTE | microRNA genes targets | Random Forest Group the genes that associated with microRNA | **A:** A novel approach of integrating microRNA into gene expression <br> **D:** The size of the groups might be large and might rank these groups highly as a result of that. | [29] |
| CogNet | | Random Forest, based on pathFindR tool | **A:** Improve the results of the pathFindR tool by ranking its groups. | [47] |
| miRcorrNet | | Random Forest based on the correlation with miRN expressions | **A:** Novel approach for integrating miRNE and mRNA expressions using machine learning. | [49] |

*Table 2. Summary Table of the presented methodologies that integrate biological knowledge. While "A" refers to the advantages, "D" refers to the disadvantages of the methods.*

## 4 Conclusions

As we have more advanced high-throughput technologies, big transcriptomic datasets became available, and extracting insights from long lists of differentially expressed genes became a challenge. Since the gene expression data typically includes small sample size but high dimensions and noise, the major challenge is the detection of disease-related information from vast amounts of redundant data and noise. Hence, the gene (feature) selection and the removal of redundant/irrelevant genes has been a key step to address this problem. For gene expression data analysis, most of the existing feature selection methods rely on expression values alone to select the genes; and biological knowledge is integrated at the end of the analysis in order to gain biological insights or to support the initial findings. However lately the gene selection process shifted from being purely data-centric to more incorporative analysis with additional biological knowledge. Integrative gene selection approaches incorporate domain knowledge from external biological resources during gene selection [9], [18], which improves interpretability and predictive performance. One of the widely used external ontology resources is the GO [19] which captures biological knowledge in a computable form that consists of a set of concepts and their relationships to each other. As another alternative, pathway-based analysis approaches aim to investigate the aggregation of the genes that are part of a functional unit, where these functional units are predefined by prior biological knowledge. These pathway-based methods rely on statistical tests that aim to detect damaged functionalities, which may result in disease phenotype. Several studies reported that the genetic variations occurring at multiple loci often disturb signal transduction, regulatory and metabolic pathways, which causes severe changes in phenotype [18]. In this regard, a widely used external ontology resource is KEGG,

which is a knowledge-base of manually curated pathways [20]. Yet another widely used external biological resource is DisGeNET, which is a meta knowledge-base for gene-disease-variant associations [20].

High-throughput profiling technologies currently enable us to concurrently measure gene expression levels for tens of thousands of genes in a single experiment, but they have some drawbacks. The high dimensionality of the gene expression data and relatively small sample sizes make the interpretation of the data a complicated, and often overwhelming task. Although sample sizes have continued to grow in recent years, new and efficient feature selection algorithms are still needed to overcome the challenges in the existing methods [4]. Hence, it is an active research topic in the field of bioinformatics.

At present, ML  is applied on specific data in order to explain and answer a specific biological query in the biological knowledge domain. One of the challenges of future integrative model-based ML is the ability to combine different biological resources to enhance our understanding of multiple biological questions. Once the full potential of the available data is achieved, it can be used in the development of gene-based diagnostic tests, drug discovery studies and in the development of therapeutic strategies for improving the public health.

To sum up, since biological systems are quite complex and they have interconnected nature, a single model that is trained on a single dataset can only benefit from a small portion of the entire biomedical knowledge. For this reason, in order to get the complete picture of molecular biology and medicine, the integration of diverse biological resources and multi-omics data is crucial. In the field of gene expression data analysis, there are still many challenges that the community needs to solve, such as the integration of gene expression datasets that are generated by different research groups for the same phenotype (which will help to overcome the batch effect) and an additional obstacle is the integration of non-similar data that each one tackles a specific disease.

**References**

[1]   Y. Hasin, M. Seldin, and A. Lusis, "Multi-omics approaches to disease," *Genome Biol.*, vol. 18, no. 1, p. 83, Dec. 2017, doi: 10.1186/s13059-017-1215-1.

[2]   M. Zitnik, F. Nguyen, B. Wang, J. Leskovec, A. Goldenberg, and M. M. Hoffman, "Machine learning for integrating data in biology and medicine: Principles, practice, and opportunities," *Inf. Fusion*, vol. 50, pp. 71–91, Oct. 2019, doi: 10.1016/j.inffus.2018.09.012.

[3]   Z. Wang, M. Gerstein, and M. Snyder, "RNA-Seq: a revolutionary tool for transcriptomics," *Nat. Rev. Genet.*, vol. 10, no. 1, pp. 57–63, Jan. 2009, doi: 10.1038/nrg2484.

[4]   K. Tomczak, P. Czerwińska, and M. Wiznerowicz, "The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge," *Contemp. Oncol. Poznan Pol.*, vol. 19, no. 1A, pp. A68-77, 2015, doi: 10.5114/wo.2014.47136.

[5]   C. Fiala and E. P. Diamandis, "Mutations in normal tissues—some diagnostic and clinical implications," *BMC Med.*, vol. 18, no. 1, p. 283, Dec. 2020, doi: 10.1186/s12916-020-01763-y.

[6]     Q. Sheng, S. Zhao, C.-I. Li, Y. Shyr, and Y. Guo, "Practicability of detecting somatic point mutation from RNA high throughput sequencing data," *Genomics*, vol. 107, no. 5, pp. 163–169, 2016, doi: 10.1016/j.ygeno.2016.03.006.

[7]     L. J. Van't Veer *et al.*, "Gene expression profiling predicts clinical outcome of breast cancer," *Nature*, 2002, doi: 10.1038/415530a.

[8]     C. H. Chou *et al.*, "miRTarBase 2016: Updates to the experimentally validated miRNA-target interactions database," *Nucleic Acids Res.*, vol. 44, no. D1, 2016, doi: 10.1093/nar/gkv1258.

[9]     R. Bellazzi and B. Zupan, "Towards knowledge-based gene expression data mining," *J. Biomed. Inform.*, vol. 40, no. 6, pp. 787–802, Dec. 2007, doi: 10.1016/j.jbi.2007.06.005.

[10]    S. Falcon and R. Gentleman, "Using GOstats to test gene lists for GO term association," *Bioinformatics*, vol. 23, no. 2, pp. 257–258, Jan. 2007, doi: 10.1093/bioinformatics/btl567.

[11]    T. G. O. Consortium, *Gene ontology: tool for the unification of biology. The Gene Ontology Consortium*, vol. 25. 2000.

[12]    R. Kustra and A. Zagdanski, "Incorporating Gene Ontology in Clustering Gene Expression Data," in *19th IEEE Symposium on Computer-Based Medical Systems (CBMS'06)*, Salt Lake City, UT, 2006, pp. 555–563, doi: 10.1109/CBMS.2006.100.

[13]    F. Azuaje and J. Dopazo, Eds., *Data analysis and visualization in genomics and proteomics*. Hoboken, NJ: John Wiley, 2005.

[14]    C. Perscheid, B. Grasnick, and M. Uflacker, "Integrative Gene Selection on Gene Expression Data: Providing Biological Context to Traditional Approaches," *J. Integr. Bioinforma.*, vol. 16, no. 1, Feb. 2019, doi: 10.1515/jib-2018-0064.

[15]    R. Bellman, *Adaptive control processes: A guided tour. (A RAND Corporation Research Study)*. 1961.

[16]    C. Lazar *et al.*, "A survey on filter techniques for feature selection in gene expression microarray analysis.," *IEEEACM Trans. Comput. Biol. Bioinforma. IEEE ACM*, vol. 9, no. 4, pp. 1106–19, 2012, doi: 10.1109/TCBB.2012.33.

[17]    I. Inza, P. Larrañaga, R. Blanco, and A. J. Cerrolaza, "Filter versus wrapper gene selection approaches in DNA microarray domains," *Artif. Intell. Med.*, vol. 31, no. 2, pp. 91–103, 2004, doi: 10.1016/j.artmed.2004.01.007.

[18]    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.

[19]    M. Ashburner *et al.*, "Gene Ontology: tool for the unification of biology," *Nat. Genet.*, vol. 25, no. 1, pp. 25–29, May 2000, doi: 10.1038/75556.

[20]    M. Kanehisa and S. Goto, "DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes," *Nucleic Acids Res*, vol. 28, no. 1, pp. 27–30, 2000.

[21]    J. Qi and J. Tang, "Integrating gene ontology into discriminative powers of genes for feature selection in microarray data," in *Proceedings of the 2007 ACM symposium on Applied computing - SAC '07*, Seoul, Korea, 2007, p. 430, doi: 10.1145/1244002.1244101.

[22]    G. Papachristoudis, S. Diplaris, and P. A. Mitkas, "SoFoCles: Feature filtering for microarray classification based on Gene Ontology," *J. Biomed. Inform.*, vol. 43, no. 1, pp. 1–14, Feb. 2010, doi: 10.1016/j.jbi.2009.06.002.

[23]    V. K. Raghu, X. Ge, P. K. Chrysanthis, and P. V. Benos, "Integrated Theory-and Data-Driven Feature Selection in Gene Expression Data Analysis," in *2017 IEEE 33rd International Conference on Data Engineering (ICDE)*, San Diego, CA, USA, Apr. 2017, pp. 1525–1532, doi: 10.1109/ICDE.2017.223.

[24] B. Quanz, M. Park, and J. Huan, "Biological pathways as features for microarray data classification," in *Proceeding of the 2nd international workshop on Data and text mining in bioinformatics - DTMBIO '08*, Napa Valley, California, USA, 2008, p. 5, doi: 10.1145/1458449.1458455.

[25] S. Mitra and S. Ghosh, "Feature Selection and Clustering of Gene Expression Profiles Using Biological Knowledge," *IEEE Trans. Syst. Man Cybern. Part C Appl. Rev.*, vol. 42, no. 6, pp. 1590–1599, Nov. 2012, doi: 10.1109/TSMCC.2012.2209416.

[26] S. Ghosh and S. Mitra, "Gene selection using biological knowledge and fuzzy clustering," in *2012 IEEE International Conference on Fuzzy Systems*, Brisbane, Australia, Jun. 2012, pp. 1–9, doi: 10.1109/FUZZ-IEEE.2012.6250797.

[27] S. Acharya, S. Saha, and N. Nikhil, "Unsupervised gene selection using biological knowledge : application in sample clustering," *BMC Bioinformatics*, vol. 18, no. 1, p. 513, Dec. 2017, doi: 10.1186/s12859-017-1933-0.

[28] M. Yousef, S. Jung, L. C. Showe, and M. K. Showe, "Recursive Cluster Elimination (RCE) for classification and feature selection from gene expression data," *BMC Bioinformatics*, vol. 8, 2007, doi: 10.1186/1471-2105-8-144.

[29] M. Yousef, L. Abdallah, and J. Allmer, "maTE: discovering expressed interactions between microRNAs and their targets," *Bioinformatics*, vol. 35, no. 20, pp. 4020–4028, Oct. 2019, doi: 10.1093/bioinformatics/btz204.

[30] M. Yousef, M. Ketany, L. Manevitz, L. C. Showe, and M. K. Showe, "Classification and biomarker identification using gene network modules and support vector machines.," *BMC Bioinformatics*, vol. 10, p. 337, 2009, doi: 10.1186/1471-2105-10-337.

[31] D. Harris and A. V. Niekerk, "Feature clustering and ranking for selecting stable features from high dimensional remotely sensed data," *Int. J. Remote Sens.*, vol. 39, no. 23, pp. 8934–8949, 2018, doi: 10.1080/01431161.2018.1500730.

[32] N. Lazzarini and J. Bacardit, "RGIFE: A ranked guided iterative feature elimination heuristic for the identification of biomarkers," *BMC Bioinformatics*, 2017, doi: 10.1186/s12859-017-1729-2.

[33] G. Deshpande *et al.*, "Recursive cluster elimination based support vector machine for disease state prediction using resting state functional and effective brain connectivity," *PLoS ONE*, 2010, doi: 10.1371/journal.pone.0014277.

[34] X. Zhao, L. Wang, and G. Chen, "Joint Covariate Detection on Expression Profiles for Identifying MicroRNAs Related to Venous Metastasis in Hepatocellular Carcinoma," *Sci. Rep.*, vol. 7, no. 1, p. 5349, 2017, doi: 10.1038/s41598-017-05776-1.

[35] M. Johannes *et al.*, "Integration of pathway knowledge into a reweighted recursive feature elimination approach for risk stratification of cancer patients," *Bioinformatics*, 2010, doi: 10.1093/bioinformatics/btq345.

[36] M. Yousef, B. Bakir-Gungor, A. Jabeer, G. Goy, R. Qureshi, and L. C. Showe, "Recursive Cluster Elimination based Rank Function (SVM-RCE-R) implemented in KNIME," *F1000Research*, vol. 9, p. 1255, Oct. 2020, doi: 10.12688/f1000research.26880.1.

[37] M. R. Berthold *et al.*, "KNIME - The Konstanz Information Miner," *SIGKDD Explor.*, vol. 11, no. 1, pp. 26–31, 2009, doi: 10.1145/1656274.1656280.

[38] G. Zycinski, A. Barla, M. Squillario, T. Sanavia, B. Di Camillo, and A. Verri, "Knowledge Driven Variable Selection (KDVS) - a new approach to enrichment analysis of gene signatures obtained from high-throughput data," *Source Code Biol. Med.*, 2013, doi: 10.1186/1751-0473-8-2.

[39] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein, *Cluster analysis and display of genome-wide expression patterns*, vol. 95. 1998.

[40]  J. Wang *et al.*, "VISDA: An open-source caBIG<sup>TM</sup> analytical tool for data clustering and beyond," *Bioinformatics*, vol. 23, no. 15, 2007, doi: 10.1093/bioinformatics/btm290.

[41]  J. W. Isabelle Guyon, B. Stephen, and V. Vladimir, "Gene Selection for Cancer Classification using Support Vector Machines, Machine Learning," *Mach. Learn.*, vol. 46, no. 1–3, pp. 389–422, 2002.

[42]  S. Nacu, R. Critchley-Thorne, P. Lee, and S. Holmes, "Gene expression network analysis and applications to immunology," *Bioinformatics*, vol. 23, no. 7, pp. 850–858, 2007, doi: 10.1093/bioinformatics/btm019.

[43]  S. R. Sain and V. N. Vapnik, "The Nature of Statistical Learning Theory," *Technometrics*, 1996, doi: 10.2307/1271324.

[44]  K.-B. Duan, J. C. Rajapakse, H. Wang, and F. Azuaje, "Multiple SVM-RFE for Gene Selection in Cancer Classification With Expression Data," *IEEE Trans. Nanobioscience*, vol. 4, no. 3, pp. 228–234, Sep. 2005, doi: 10.1109/TNB.2005.853657.

[45]  P. Das, A. Roychowdhury, S. Das, S. Roychoudhury, and S. Tripathy, "sigFeature: Novel Significant Feature Selection Method for Classification of Gene Expression Data Using Support Vector Machine and t Statistic," *Front. Genet.*, vol. 11, p. 247, Apr. 2020, doi: 10.3389/fgene.2020.00247.

[46]  C. Xu and S. A. Jackson, "Machine learning and complex biological data," *Genome Biol.*, vol. 20, no. 1, pp. 76, s13059-019-1689–0, Dec. 2019, doi: 10.1186/s13059-019-1689-0.

[47]  Yousef Malik, Ülgen Ege, and Sezerman Osman Uğur, "CogNet: Classification of gene expression data based on ranked active-subnetwork-oriented KEGG pathway enrichment analysis," *PeerJ*, 2020.

[48]  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.

[49]  Yousef Malik, Goy Gokhan, Mitra Ramkrishna, Eischen Christine M., Jabeer Amhar, and Bakir-Gungor Burcu, "miRcorrNet:  Integrated microRNA Gene Expression and mRNA Expression Based Machine Learning combined with Features Grouping and Ranking," *Rev.*

[50]  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.