Application of Biological Domain Knowledge

Subjects: Statistics & Probability
Contributor: Malik Yousef

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.

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 have been conducted which yielded massive volumes of biological data, including the genetic variations associated with specific phenotypes. Currently, we are encountering an -omics revolution in which genome, epigenome, transcriptome, and other -omics can be readily characterized. With advancements in various -omics approaches, it is 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]. Additionally, 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 noncoding 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.

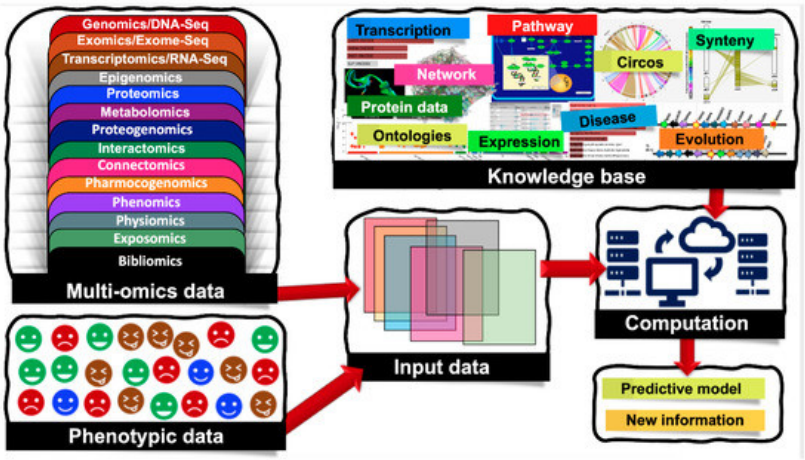

Figure 1. Machine learning (ML) applications that combine multi-omics and phenotypic data. Multi-omics data are classified into the following groups: genomics/DNA-Seq—the study of the genetic material for an organism, it assesses DNA sequence and structural variations including single-nucleotide polymorphisms (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 noncoding 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 behavior of an organism; exposomics—study of an organism's environment and bibliomics (the literature concerning a topic).

Traditional analyses attempted to untangle the molecular mechanisms of complex diseases using a single -omics dataset which contributes 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. As such, recent studies are shifting towards multi-omics data analysis, where each of these different -omics data types are critical for deciphering the 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 physiologically reasonable and 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 the large oncogenomic expression profiles hosted at The Cancer Genome Atlas (TCGA) [4]. Mutations are the core causative agents of diseases such as 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 the 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 such as lymph node negative [7].

Data analysis approaches to gene expression profiling 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 has become available, in the 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 (such as associations and classification) and implications (such as reverse-engineering of gene–gene networks and resulting phenotypes). Authors surveyed the clustering approaches that group the genes using different distance measures, such as 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 one to define semantic similarity between the genes via incorporating the GO [11]. An additional study by Kustra and Zagdanski [12] used the incorporation of GO annotation to expression data by inducing a correlation-based dissimilarity matrix to derive a 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 Ref. [13] is the first book on integrative data analysis and visualization in this area. It outlines essential techniques for the integration of data derived from multiple sources. It is one of the first systematic books that overviews 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.

2. Gene Selection Approaches for Gene Expression Datasets

Gene selection approaches for gene expression datasets can be mainly categorized into two classes, such 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 [15][16]. 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 the 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 Ref. [16] 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 Ref. [15]. I. Inza [16] also carried out a comparison between filter metrics and the wrapper sequential search procedure, which are both applied on gene expression datasets. Additionally, 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, 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 [17]. The main disadvantages 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][17], which improves interpretability and predictive performance. One of the widely used external ontology resources is the Gene Ontology (GO) [18], 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 a statistical analysis/ML analyses was to incorporate Gene Ontology (GO) [18] in clustering gene expression data [10]. Another widely used external ontology resource is the Kyoto Encyclopedia of Genes and Genomes (KEGG), which is a pathway knowledge-base providing manually curated pathways [19]. Yet another widely used external biological resource is DisGeNET, which is a meta knowledge-base on gene–disease–variant associations [19].

One example of the 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 [20]. 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 a 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 the integrative gene selection approach is SoFoCles [21], 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, such as χ2, ReliefF, or IG. The initial set of candidate genes is 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 into the gene selection process improves the results.

Yet another study by Fang et al. [17] 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. [22] makes use of KEGG, DisGeNET, and further genetic meta information [19]. 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 a 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 [23]. 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 are 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 thus for the classification task the dimensionality is reduced from 22,283 to 17. Even though this approach was not tested on multiclass problems such as cancer (sub-) type classification, the experiments on binary classification problems showed an improved performance over different traditional approaches.

Mitra et al. adopted the clustering large applications based upon randomized search (CLARANS) method to the feature (gene) selection problem via utilizing biological knowledge [24]. 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 [25].

In Ref. [26], the authors proposed an integrative gene (feature) selection approach based on the 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 are integrated, such traditional filter approaches become compatible with more complex machine learning approaches at very similar classification accuracies, but far lower 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. However, 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 by which -omics data types can be integrated and information on 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. As such, 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 [27]. 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 [28], which uses the same approach based on the interactions of microRNAs (miRNA) and their gene targets. Additionally, in the literature, the biological information buried in genetic interaction networks is utilized for classification studies. For example, SVM-RNE (SVM with recursive network elimination) integrates network information with recursive feature elimination based on SVM [29]. It is shown that SVM-RNE has a 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 [30][31], which indicates the importance and the merit of the SVM-RCE approach. The study of Ref. [32] has a slightly modified SVM-RCE algorithm in the disease state prediction step. Additionally, they used the already invented term of “recursive cluster elimination”.

The study of Zhao, X. et al. [33] 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. [34] for integration of pathway knowledge into a reweighted recursive feature elimination approach for the risk stratification of cancer patients. A recent tool, SVM-RCE-R [35], is an updated version of SVM-RCE, which is implemented in Knime [36], and uses a random forest classifier with additional important features such as suggesting a new approach of ranking the clusters.

The term “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 Ref. [37], in which the authors proposed a framework that uses a priori biological knowledge in high-throughput data analysis.

The RCE algorithm [27] 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 [28][29]. 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 are 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 1, we summarize the specifications, advantages and disadvantages of the presented integrative gene selection approaches.

Table 1. Summary table of the presented methodologies that integrate biological knowledge. While “A” refers to the advantages, “D” refers to the disadvantages of the methods.

Tool Name	Incorporated Biological Knowledge	Methodology	Advantage/Disadvantage	Ref
N/A	GO	Rank the genes uses information gain (IF) incorporated with Gene Ontology 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.	[20]
N/A	GO	$\chi 2$, ReliefF, or IG	**A:** Including biological knowledge in the gene selection process improves results.	[21]
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.	[17]
N/A	KEGG, DisGeNET, and further genetic meta information	Gene–disease association score from DisGeNET Gene distance metrics		[22]
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.	[23]
N/A	Gene ontology (GO)	Randomized search (CLARANS)	**A:** Reducing the dimension dramatically.	[24]
SVM-RCE	Genes related are correlated	SVM and K-means	**A:** Discover significant of clusters. **D:** Might lose important genes because they were in lower-ranked clusters.	[27] [35]
SVM-RNE	GXNA for creating subnetworks from gene expression	SVM, GXNA	**A:** Reducing the dimension of the data by considering subnetworks. **D:** The subnetworks are created as a prediction of the gene expressions data.	[29]
maTE	microRNA genes targets	Random forest groups the genes that associated with microRNA	**A:** A novel approach of integrating microRNA into gene expression. **D:** The size of the groups might be large and might rank these groups highly as a result of that.	[27]
CogNet		Random forest, based on pathFindR tool	**A:** Improve the results of the pathFindR tool by ranking its groups.	[38]
miRcorrNet		Random forest based on the correlation with miRN expressions	**A:** Novel approach for integrating miRNE and mRNA expressions using machine learning.	[39]

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 [40]. Therefore, one potential way of discovering gene function is to group genes with a similar expression profile. Thus, different clustering algorithms [41] 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 to 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 for and determine significant groups/clusters of features based on one or more biological grouping function (will be referred as *bioF()* throughout the rest of this paper) that are integrated with the 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, it 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 (such as SVM-RFE [42]), where the algorithm takes as input the data of 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 their interest and their research aims. The outcome of the algorithm will be more specific to the researcher's interest.

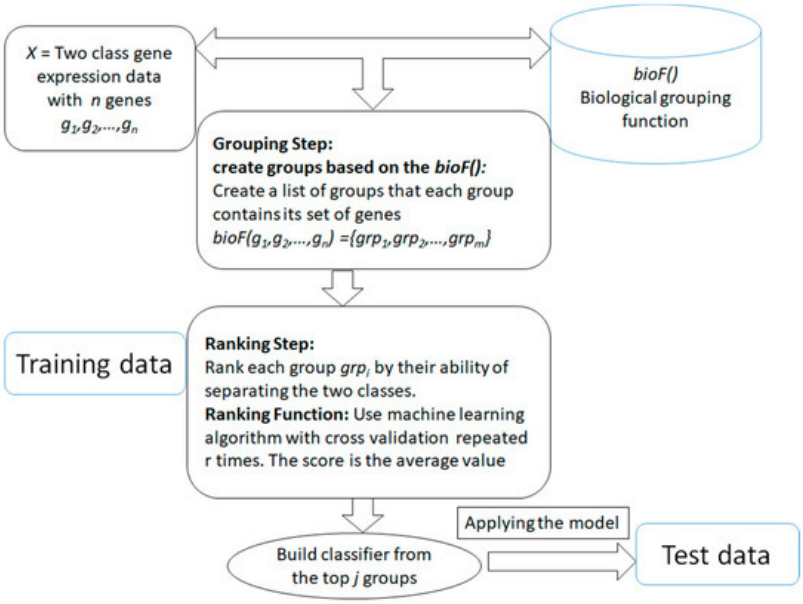

Figure 2. The generic framework of the algorithm that is based on biological integration for grouping, ranking and classification.

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 [28] tool. One interesting use of *bioF()* is that it allows one 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 [27] for grouping correlated genes. Similarly, SVM-RNE [29] incorporates another tool, GXNA [43], to create the groups. GXNA utilizes gene expression profiles and prior biological information to suggest differentially expressed pathways or gene networks.

References

1. Hasin, Y.; Seldin, M.; Lusis, A. Multi-omics approaches to disease. Genome Biol. 2017, 18, 83, doi:10.1186/s13059-017-1215-1.

2. Zitnik, M.; Nguyen, F.; Wang, B.; Leskovec, J.; Goldenberg, A.; Hoffman, M.M. Machine learning for integrating data in biology and medicine: Principles, practice, and opportunities. Inf. Fusion 2019, 50, 71–91, doi:10.1016/j.inffus.2018.09.012.
3. Wang, Z.; Gerstein, M.; Snyder, M. RNA-Seq: A revolutionary tool for transcriptomics. Nat. Rev. Genet. 2009, 10, 57–63, doi:10.1038/nrg2484.
4. Tomczak, K.; Czerwińska, P.; Wiznerowicz, M. The Cancer Genome Atlas (TCGA): An immeasurable source of knowledge. Contemp. Oncol. Poznan Pol. 2015, 19, A68–A77, doi:10.5114/wo.2014.47136.
5. Fiala, C.; Diamandis, E.P. Mutations in normal tissues—some diagnostic and clinical implications. BMC Med. 2020, 18, 283, doi:10.1186/s12916-020-01763-y.
6. Sheng, Q.; Zhao, S.; Li, C.-I.; Shyr, Y.; Guo, Y. Practicability of detecting somatic point mutation from RNA high throughput sequencing data. Genomics 2016, 107, 163–169, doi:10.1016/j.ygeno.2016.03.006.
7. Veer, L.J.V.; Laura, J.; Dai, H.; van de Vijver, M.J.; He, Y.D.; Hart, A.A.M.; Mao, M.; Peterse, H.L.; van der Kooy, K.; Marton, M.J.; Witteveen, A.T.; et al. Gene expression profiling predicts clinical outcome of breast cancer. Nature 2002, doi:10.1038/415530a.
8. Chou, C.; Chang, N.; Shrestha, S.; Hsu, S.; Lin, Y.; Lee, W.; Yang, C.; Hong, H.; Wei, T.; Tu, S.; et al. miRTarBase 2016: Updates to the experimentally validated miRNA-target interactions database. Nucleic Acids Res. 2016, 44, doi:10.1093/nar/gkv1258.
9. Bellazzi, R.; Zupan, B. Towards knowledge-based gene expression data mining. J. Biomed. Inform. 2007, 40, 787–802, doi:10.1016/j.jbi.2007.06.005.
10. Falcon, S.; Gentleman, R. Using GOstats to test gene lists for GO term association. Bioinformatics 2007, 23, 257–258, doi:10.1093/bioinformatics/btl567.
11. Consortium, T.G.O. Gene ontology: Tool for the unification of biology. Gene Ontol. Consort. 2000, 25, 25–29.
12. Kustra, R.; Zagdanski, A. Incorporating Gene Ontology in Clustering Gene Expression Data. In Proceedings of the 19th IEEE Symposium on Computer-Based Medical Systems (CBMS'06), Salt Lake City, UT, USA, 22–23 June 2006; pp. 555–563, doi:10.1109/CBMS.2006.100.
13. Data Analysis and Visualization in Genomics and Proteomics; Azuaje, F., Dopazo, J., Eds.; John Wiley: Hoboken, NJ, USA, 2005.
14. Perscheid, C.; Grasnick, B.; Uflacker, M. Integrative Gene Selection on Gene Expression Data: Providing Biological Context to Traditional Approaches. J. Integr. Bioinforma. 2019, 16, doi:10.1515/jib-2018-0064.
15. Lazar, C.; Taminau, J.; Meganck, S.; Steenhoff, D.; Coletta, A.; Molter, C.; de Schaetzen, V.; Duque, R.; Bersini, H.; Nowe, A. A survey on filter techniques for feature selection in gene expression microarray analysis. IEEEACM Trans. Comput. Biol. Bioinform. IEEE ACM 2012, 9, 1106–1119, doi:10.1109/TCBB.2012.33.
16. Inza, I.; Larrañaga, P.; Blanco, R.; Cerrolaza, A.J. Filter versus wrapper gene selection approaches in DNA microarray domains. Artif. Intell. Med. 2004, 31, 91–103, doi:10.1016/j.artmed.2004.01.007.
17. Fang, O.H.; Mustapha, N.; Sulaiman, M.N. An integrative gene selection with association analysis for microarray data classification. Intell. Data Anal. 2014, 18, 739–758, doi:10.3233/IDA-140666.
18. Ashburner, M.; Ball, C.A.; Blake, J.A.; Botstein, D.; Butler, H.; Cherry, J.M.; Davis, A.P.; Dolinski, K.; Dwight, S.S.; Eppig, J.T.; et al. Gene Ontology: Tool for the unification of biology. Nat. Genet. 2000, 25, 25–29, doi:10.1038/75556.
19. Janet Piñero, Àlex Bravo, Núria Queralt-Rosinach, Alba Gutiérrez-Sacristán, Jordi Deu-Pons, Emilio Centeno, Javier García-García, Ferran Sanz, Laura I. Furlong, DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants, Nucleic Acids Research, Volume 45, Issue D1, January 2017, Pages D833–D839, https://doi.org/10.1093/nar/gkw943
20. Qi, J.; Tang, J. 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, 11–15 March 2007; p. 430, doi:10.1145/1244002.1244101.
21. Papachristoudis, G.; Diplaris, S.; Mitkas, P.A. SoFoCles: Feature filtering for microarray classification based on Gene Ontology. J. Biomed. Inform. 2010, 43, 1–14, doi:10.1016/j.jbi.2009.06.002.
22. Raghu, V.K.; Ge, X.; Chrysanthis, P.K.; Benos, P.V. Integrated Theory-and Data-Driven Feature Selection in Gene Expression Data Analysis. In Proceedings of the 2017 IEEE 33rd International Conference on Data Engineering (ICDE), San Diego, CA, USA, 19–22 April 2017; pp. 1525–1532, doi:10.1109/ICDE.2017.223.
23. Quanz, B.; Park, M.; Huan, J. Biological pathways as features for microarray data classification. In 2nd International Workshop on Data and Text Mining in Bioinformatics—DTMBIO '08; ACM Press: Napa Valley, CA, USA, 2008; p. 5,

doi:10.1145/1458449.1458455.

24. Mitra, S.; Ghosh, S. Feature Selection and Clustering of Gene Expression Profiles Using Biological Knowledge. IEEE Trans. Syst. Man Cybern. Part. C Appl. Rev. 2012, 42, 1590–1599, doi:10.1109/TSMCC.2012.2209416.
25. Ghosh, S.; Mitra, S. Gene selection using biological knowledge and fuzzy clustering. In Proceedings of the 2012 IEEE International Conference on Fuzzy Systems, Brisbane, Australia, 10–15 June 2012; pp. 1–9, doi:10.1109/FUZZ-IEEE.2012.6250797.
26. Acharya, S.; Saha, S.; Nikhil, N. Unsupervised gene selection using biological knowledge : Application in sample clustering. BMC Bioinform. 2017, 18, 513, doi:10.1186/s12859-017-1933-0.
27. Yousef, M.; Jung, S.; Showe, L.C.; Showe, M.K. Recursive Cluster Elimination (RCE) for classification and feature selection from gene expression data. BMC Bioinform. 2007, 8, doi:10.1186/1471-2105-8-144.
28. Yousef, M.; Abdallah, L.; Allmer, J. maTE: Discovering expressed interactions between microRNAs and their targets. Bioinformatics 2019, 35, 4020–4028, doi:10.1093/bioinformatics/btz204.
29. Yousef, M.; Ketany, M.; Manevitz, L.; Showe, L.C.; Showe, M.K. Classification and biomarker identification using gene network modules and support vector machines.. BMC Bioinform. 2009, 10, 337, doi:10.1186/1471-2105-10-337.
30. Harris, D.; Niekerk, A.V. Feature clustering and ranking for selecting stable features from high dimensional remotely sensed data. Int. J. Remote Sens. 2018, 39, 8934–8949, doi:10.1080/01431161.2018.1500730.
31. Lazzarini, N.; Bacardit, J. RGIFE: A ranked guided iterative feature elimination heuristic for the identification of biomarkers. BMC Bioinform. 2017, doi:10.1186/s12859-017-1729-2.
32. Deshpande, G.; Li, Z.; Santhanam, P.; Coles, C.D.; Lynch, M.E.; Hamann, S.; Hu, X. 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.
33. Zhao, X.; Wang, L.; Chen, G. Joint Covariate Detection on Expression Profiles for Identifying MicroRNAs Related to Venous Metastasis in Hepatocellular Carcinoma. Sci. Rep. 2017, 7, 5349, doi:10.1038/s41598-017-05776-1.
34. Johannes, M.; Brase, j.; Fröhlich, H.; Gade, S.; Gehrmann, M.; Fälth, M.; Sültmann, H.; Beißbarth, T. Integration of pathway knowledge into a reweighted recursive feature elimination approach for risk stratification of cancer patients. Bioinformatics 2010, doi:10.1093/bioinformatics/btq345.
35. Yousef, M.; Bakir-Gungor, B.; Jabeer, A.; Goy, G.; Qureshi, R.; Showe, L.C. Recursive Cluster Elimination based Rank Function (SVM-RCE-R) implemented in KNIME. F1000Research 2020, 9, 1255, doi:10.12688/f1000research.26880.1.
36. Berthold, M.R.; Cebron, N.; Dill, F.; Gabriel, T.R.; Kötter, T.; Meinl, T.; Ohl, P.; Thiel, K.; Wiswedel, B. KNIME—The Konstanz Information Miner. SIGKDD Explor. 2009, 11, 26–31, doi:10.1145/1656274.1656280.
37. Zycinski, G.; Barla, A.; Squillario, M.; Sanavia, T.; di Camillo, B.; Verri, A. 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.
38. Malik, Y.; Ege, Ü.; Uğur, S.O. CogNet: Classification of gene expression data based on ranked active-subnetwork-oriented KEGG pathway enrichment analysis. PeerJ 2020.
39. Malik, Y.; Gokhan, G.; Ramkrishna, M.; Eischen, C.M.; Amhar, J.; Burcu, B. miRcorrNet: Integrated microRNA Gene Expression and mRNA Expression Based Machine Learning combined with Features Grouping and Ranking.
40. Eisen, M.B.; Spellman, P.T.; Brown, P.O.; Botstein, D. Cluster Analysis and Display of Genome-Wide Expression Patterns; National Academy of Sciences: Washington, D.C., WA, USA 1998; Volume 95.
41. Wang, J.; Li, H.; Zhu, Y.; Yousef, M.; Nebozhyn, M.; Showe, M.; Showe, L.; Xuan, J.; Clarke, R.; Wang, Y. VISDA: An open-source caBIGTM analytical tool for data clustering and beyond. Bioinformatics 2007, 23, doi:10.1093/bioinformatics/btm290.
42. Guyon, J.W.I.; Stephen, B.; Vladimir, V. Gene Selection for Cancer Classification using Support Vector Machines, Machine Learning. Mach. Learn. 2002, 46, 389–422.
43. Nacu, S.; Critchley-Thorne, R.; Lee, P.; Holmes, S. Gene expression network analysis and applications to immunology. Bioinformatics 2007, 23, 850–858, doi:10.1093/bioinformatics/btm019.