




A Feature Selection and Association Rule Approach to Identify Genes Associated with Metastasis and Low Survival in Sarcoma

M. Lourdes Linares-Barrera¹, María Martínez-Ballesteros¹ (✉) ,
José M. García-Heredia² , and José C. Riquelme¹ 

¹ Departamento de Lenguajes y Sistemas Informáticos, Universidad de Sevilla,
41012 Seville, Spain

marlinbar@alum.us.es, {mariamartinez,riquelme}@us.es

² Departamento de Bioquímica Vegetal Biología Molecular, Universidad de Sevilla,
41012 Seville, Spain
jmgheredia@us.es

Abstract. Sarcomas are rare mesodermal tumors of heterogeneous nature and have a higher incidence in children. The relative 5-year survival rate for patients with metastatic sarcoma is usually low. Standard treatment for sarcomas involves surgical resection, and investigating the genetic basis of these tumors through genome-wide analysis is crucial due to their rarity and late diagnosis. This work proposes a methodology that combines preprocessing, feature selection and association rule mining to identify relevant genes and significant relationships in biological data from sarcoma patients. Our study aims to identify the relationships between metastasis-associated genes and patient survival of less than 5 years. The proposed approach was applied to a sarcoma dataset containing data on gene expression, metastasis occurrence, and survival time, revealing a set of biologically relevant gene interactions associated with sarcoma metastasis and low survival rates. The combined use of these techniques can facilitate the identification of biomarkers or gene signatures associated with the disease and provide insight into the underlying biological mechanisms involved in sarcomas.

Keywords: feature selection · association rules · gene expression · sarcoma

1 Introduction

Sarcomas are rare and heterogeneous mesodermal tumors that primarily affect children and adolescents, accounting for over 20% of all pediatric tumors. The standard treatment for sarcomas is surgical resection, and chemotherapy and/or radiation therapy are administered to only a subset of patients. The 5-year relative survival rate for patients with metastatic sarcoma is a mere 15% [18]. Given

https://doi.org/10.1007/978-3-031-40725-3_62

the rarity of the disease and the fact that it is often diagnosed in its later stages, investigating the genetic basis of sarcomas through genome-wide analysis is crucial. The identification of genetic patterns using machine learning techniques could have a significant impact on sarcoma research. These technologies allow for the analysis of large datasets, enabling the discovery of specific genetic patterns and relationships between genes that would not be detectable with traditional methods. This could lead to a better understanding of the molecular mechanisms underlying sarcomas and the identification of novel therapeutic targets [5].

Association rule mining is a powerful technique that could be useful for extracting patterns in gene expression data as it can identify interesting relationships between genes or gene expressions that may not be immediately apparent [13]. By analyzing co-occurrence patterns between genes or gene expressions, association rule mining can help identify groups of genes that are co-expressed, co-regulated, or functionally related. However, one of the main challenges in gene expression data is the high dimensionality in terms of the number of genes and the large imbalance between the number of genes and the number of samples. Feature selection techniques [8] can be useful in this case to reduce the number of features to a manageable subset that is relevant to the problem at hand. By selecting only the most informative genes, feature selection can improve the accuracy and interpretability of machine learning models, and reduce the computational complexity of downstream analyses.

In this study, we propose a methodology that combines preprocessing, feature selection and association rule mining to identify relevant genes in biological data from sarcoma patients and to determine significant relationships among them. Specifically, our analysis focused on selecting a subset of genes that are highly associated with the presence of metastasis and discovering relationships among those genes that contribute to low survival rates. To identify relevant genes, we employed a feature selection technique based on the conditional dependence of variables. This method allowed us to narrow down the subset of genes that are related to metastasis. We then used an association rule algorithm to discover association rules among the selected set of genes that are highly related to metastasis in patients with a survival time of less than 5 years. By combining both techniques, our methodology facilitates the identification of biomarkers or gene signatures associated with sarcoma and provides insights into the underlying biological mechanisms involved in the disease.

The structure of the paper is organized as follows. Section 2 provides a brief overview of related work focused on machine learning techniques based on feature selection and association rules applied to gene expression data. Section 3 describes the proposed methodology in detail. In Sect. 4, the experimental setting used is presented, followed by the reporting of the experimental results and analysis of the selected genes of the proposed methodology. Finally, Sect. 5 discusses the main conclusions of this work.

2 Related Work

Association rules and feature selection are widely used in gene expression data analysis. Association rules identify co-occurrences of genes in different conditions, while feature selection identifies the most informative genes relevant to a phenotype. These techniques have proven effective in identifying biomarkers in various gene expression datasets, including cancer and neurological disorders.

There is a wealth of literature in the field of gene expression analysis that has applied diverse machine learning techniques, particularly in cancer research [10]. In [16], authors present a methodology using AR to identify cancer-related genes, validated through hierarchical cluster analysis, fold-change, and literature review, which successfully characterizes colon cancer patients. In [12], ARs were used to identify genes highly linked to the neurodegenerative disease, based on changes in expression levels between control and patient samples on Alzheimer's disease. The authors in [2] propose a new explainable artificial intelligence strategy based on ARs, which involves pre-processing, knowledge extraction, and functional validation, to identify biologically relevant sequential patterns in longitudinal human gene expression data. In [7], ARs have been used to identify prognostic markers for selecting combined treatments in sarcoma.

Feature selection is a crucial step in analyzing gene expression data due to the challenge of extracting disease-related information from a large amount of redundant data and noise, and eliminating irrelevant genes can help address this issue. Authors in [9] proposed a method to effectively select feature genes from gene expression data by combining double RBF-kernels with weighted analysis. The method outperformed previous methods in terms of accuracy, true positive rate, false positive rate, and runtime when tested on four benchmark datasets. In [15], the authors introduced a methodology for identifying potential biomarkers at the DNA methylation level to distinguish different subtypes of sarcoma. They used a machine learning process to analyze sarcoma samples and employed feature selection and classification algorithms to construct models that could classify sarcoma subtypes based on DNA methylation patterns.

3 Methodology

The objective of the proposed study is to find relevant genetic patterns in sarcoma patients highly related to metastasis and involved in low survival rates using an integrated methodology based on feature selection and association rule mining.

The complete process of the methodology is drawn in Fig. 1. The data analysis methodology applied in this work includes the following steps, which will be described in the subsequent sections:

1. First phase: Preprocessing of data, including filtering, summarization and analysis of differential gene expression (Sect. 3.1).
2. Second phase: Feature selection to select a subset of relevant genes for analysis (Sect. 3.2).

3. Third phase: Application of association rule mining to analyze gene expression data and identify interesting relationships between genes, leading to the identification of potential biomarkers for diseases such as sarcoma (Sect. 3.3).

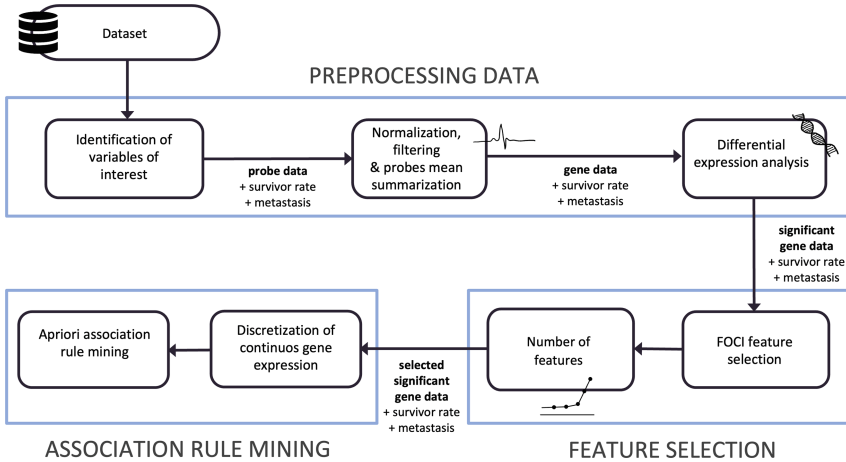


Fig. 1. Methodology proposed to find relevant gene expression patterns related with metastasis and a low survival time.

3.1 First Phase: Preprocessing Data

In the study conducted in this work, we will focus on microarray data. These data are obtained through molecular technology that enables the simultaneous measurement of the expression of thousands of genes from a biological sample. They contain measured values of multiple probes for each patient examined, along with a set of metadata.

Several preprocessing tasks have been conducted to prepare a suitably reduced dataset to be analyzed. The main tasks carried out in the preprocessing phase are detailed as follows.

1. **Identification of variables of interest:** Firstly, we need to select the variables that are of interest in our study. This work focuses on metastasis, survival time, and gene expression values from different probes. The first preprocessing step would involve removing null values and extracting the relevant data. This is crucial to ensure that subsequent analyzes are based on a clean and comprehensive dataset.
2. **Normalization, filtering and summarization:** The next step in the pipeline involves normalizing the data using the \log_2 technique, which has several benefits including transforming the data to a logarithmic scale, stabilizing variance, and facilitating comparison between experiments. Once normalized, we have applied a filtering step to remove data that did not meet certain criteria, such as expression level or variability, in order to

eliminate noisy or low-quality data. Specifically, we have removed all probes that exhibited low expression levels on average ($\log_2 < 5$) or displayed minimal variation among patients ($\log_2 \text{MAX} - \log_2 \text{MIN} < 1$).

After filtering probes, the data is summarized by collecting measurements from multiple probes associated with a gene. This summarization technique provides several advantages, including identifying gene patterns, reducing the dimensionality of the problem, and obtaining a more precise estimation. We used the mean summarization technique to calculate the average gene expression value from multiple probes associated with the same gene. The probes were grouped using a converter from Affymetrix probe to gene symbol. In particular, we have used the biological database The Database for Annotation, Visualization, and Integrated Discovery (DAVID) [6] to obtain the corresponding Affymetrix probe-to-gene symbol notation.

3. **Differential expression analysis:** The preprocessing phase's final stage aims to identify genes with significant changes in expression levels between different biological groups. To identify such genes between patients with and without metastasis, we used the limma package from the Bioconductor project [17], which employs Bayesian statistics to obtain the significance of differentially expressed genes. By fitting a linear model based on the contrast matrix, we estimated the disparity in gene expression between the two study groups, and then identified the differentially expressed genes based on Bayesian statistics. The extracted data represents the genes that are differentially expressed with respect to the two study groups related to metastasis.

3.2 Second Phase: Feature Selection

This works aims at identifying relationships between gene expression values and survival time in patients with metastasis. Therefore, it is necessary to identify the most relevant subset of genes for our study from all the differentially expressed genes that have been extracted in the previous stage.

Therefore, we have applied the FOCI method [3] for the selection of the most relevant subset of attributes for rule extraction. FOCI is a model-free algorithm that does not require tuning parameters, and its consistency is provable under sparsity assumptions. The algorithm has demonstrated excellent performance in both simulated and real datasets, making it a promising tool for variable selection in various applications. This attribute selector works as a forward step-wise feature selection algorithm. The algorithm starts with an empty subset of attributes and, at each iteration, adds the gene that best explains the variable we consider as a response (in our case, metastasis) based on a non-parametric statistic T_n that is based on the notions of conditional dependence (to take into account the relationship between genes as well as the variable's ability to explain metastasis) and nearest neighbors. The algorithm finishes when it reaches the selected number of attributes, returning the subset of genes of the indicated length that present a higher conditional dependence on metastasis. The selection of relevant genes will be used in addition to metastasis and survival data of patients in the extraction of gene patterns through association rules.

The selected number of attributes is a parameter required by FOCI. In order to select the appropriate number of attributes for our analysis, we have considered the point at which the statistic T_n stabilizes, indicating that the addition of more attributes may not improve the accuracy or reliability of the analysis.

3.3 Third Phase: Association Rules Extraction

The well-known Apriori algorithm [1] was applied to discover the association rules in this study. This algorithm requires the discretization of numerical data and the parametrization of the minimum threshold of some quality measures such as support and confidence (Eqs. 1 and 2 described in Sect. 4.2).

The study proposed in this work aims at obtaining relationships among the selected genes in the previous phase and patient survival of less than 5 years. Furthermore, it is interesting to identify whether metastasis is also present in these relationships.

To fulfill this goal, several steps have been conducted:

- As first step, the numerical variables of dataset had to be discretized to convert the continuous variables into categorical variables suitable for association rule mining. On the one hand, the numerical variables associated to genes expression levels were discretized in two intervals to categorize the levels of expression in low (categorized as 0) or high (categorized as 1) using the equal interval width method. On the other hand, the numerical variable related to patient survival time was discretized using fixed interval boundaries to categorize into survival time less than 5 years (categorized as 0) or greater or equal to 5 years (categorized as 1). Metastasis has not been discretized because is already categorized into 0 (absence) or 1 (presence) in the dataset.
- As second step, the Apriori algorithm was applied in the discretized data using minimum support and confidence thresholds. Note that these measures range into the interval 0 to 1, then we are interested in obtaining specific association rules with the highest possible reliability. The minimum and maximum length (number of items) of the association rules have been set to a low number of items, to reduce the set of rules to be found by Apriori. Note that Apriori only creates rules with one item in the consequent of the rule, but the number of items in the antecedent could be higher than 1. Therefore, the selected genes categorized as 0 or 1 in addition to patient survival time categorized as 0 (less than 5 years) have been fixed to appear in the antecedent of the rules. Metastasis with value 1 has been restricted to belong to the consequent of the rules.
- In the final step, we have selected the strongest association rules that contain the survival time along with at least one gene in the antecedent. An example of such a rule that can be derived from our study is:

$$\{DSTN = 0 \wedge ZWINT = 1 \wedge t_survivor = 0\} \implies \{\text{metastasis} = 1\}$$

4 Experimentation

This section provides a description of the dataset selected used in our experimentation, the quality measures applied to evaluate the association rules obtained, and presents and discusses the results obtained in each phase of the proposed methodology detailed in Sect. 3.

4.1 Dataset

For the study, we have selected the gene expression dataset GSE21050 from the French Sarcoma Group (FSG) database, which is available in the public repository Gene Expression Omnibus (GEO) [4, 14].

This dataset contains sarcomas that were examined with the aim of characterizing gene expression profiles and identifying possible genetic markers to improve tumor diagnosis. In particular, soft tissue sarcomas with no recurrent chromosomal translocations and for which a frozen tissue of the untreated primary tumor was available. The sarcomas were split into two cohorts. In particular, the dataset contains 310 microarrays sarcoma expression data, survival time of patients, metastasis status, gene CINSARC signature, sample source and diagnosis, among other biological and clinical information. For the study conducted in this work, we have selected the expression data as well as the survival time and metastasis status. The expression values correspond to Affymetrix probes.

4.2 Association Rule Quality Measures

This section presents the most common measures to evaluate the quality of association rules [11].

The support for the rule $(A \implies C)$, where A and C denote the antecedents and consequents respectively, measures the percentage of instances in the dataset that contain A and B , simultaneously. $n(A, C)$ denotes the number of instances satisfying both conditions, and N is the total number of instances in the dataset. The support values lie between 0 and 1.

$$\text{sup}(A \implies C) = \frac{n(A, C)}{N} \quad (1)$$

The confidence is the probability that instances in the dataset that satisfy the antecedent condition (A) also satisfy the consequent condition (C). Like support, confidence also ranges from 0 to 1.

$$\text{conf}(A \implies C) = \frac{\text{sup}(A \implies C)}{\text{sup}(A)} \quad (2)$$

Lift measures how the co-occurrence of A and C in the dataset exceeds what would be expected if A and C were statistically independent. A lift value greater than one indicates a positive dependence between the two items.

$$\text{lift}(A \implies C) = \frac{\text{sup}(A \implies C)}{\text{sup}(A)\text{sup}(C)} \quad (3)$$

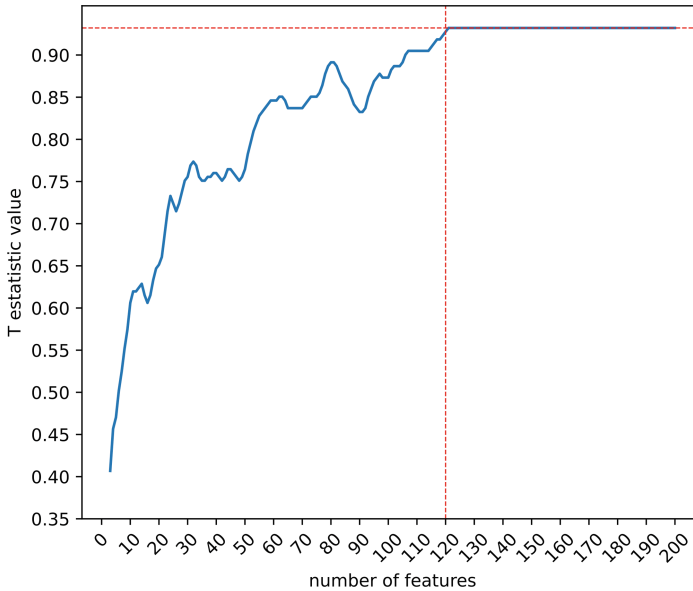


Fig. 2. Number of genes selected using FOCI.

4.3 Preprocessing Results

After completing the pre-processing steps outlined in the first phase of the proposed methodology (Sect. 3.1), we obtained a dataset from 309 patients that contains 1,516 gene expression data points, along with information regarding the metastasis status and the corresponding survival time. In particular, after the filtering and summarization process, we achieved a reduction of the dimensionality from 54,612 probes to 11,901 genes. Additionally, following the differential expression analysis, we obtained a filtering step resulting in 1,516 genes out.

4.4 Feature Selection Results

As described in Sect. 3.2, the number of attributes to be obtained by the FOCI selection feature method in the second phase of our proposal has been determined using the methodology based on the stabilization of the T_n statistic value. As shown in Fig. 2, it can be seen that the T_n value is stabilized when the number of genes reaches 120. Therefore, we have selected this value as the number of genes to be selected by the FOCI method.

Figure 3 displays the volcano plot generated for the genes selected by FOCI. In this plot, the log-fold change of gene expression levels for patients with metastasis is plotted on the x-axis, while the negative log₁₀ p-value is plotted on the y-axis. The most significantly differentially expressed genes are represented by red dots, with a false discovery rate of less than 0.05 used as the cut-off for

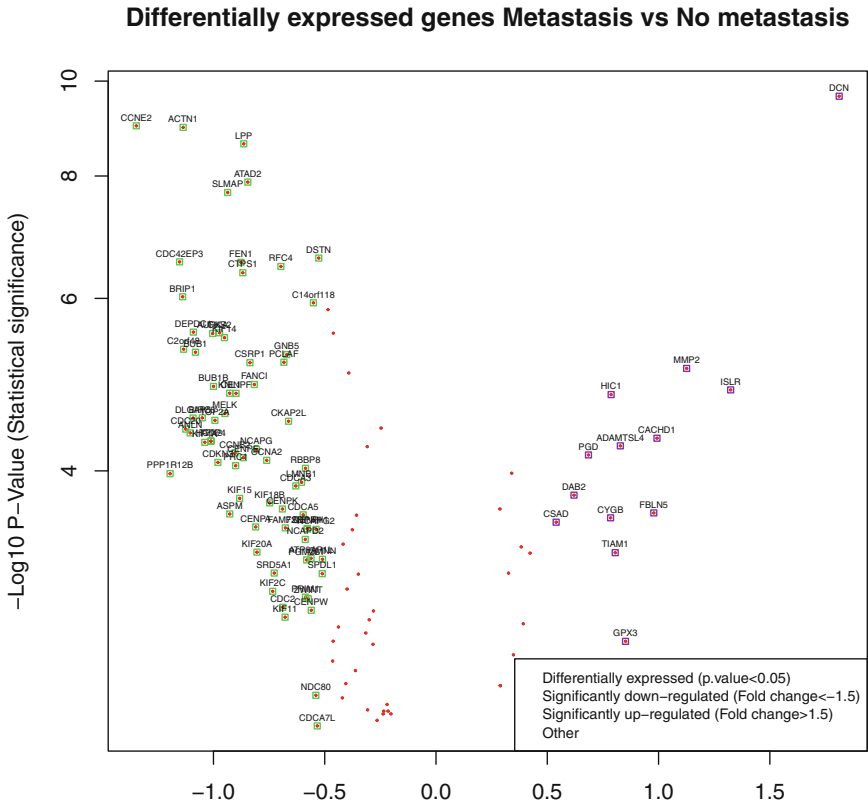


Fig. 3. Volcano plot of the expression levels of the genes selected by FOCI.

statistical significance. Notably, all 120 genes selected by FOCI exhibit p-values lower than 0.05, indicating that they are all statistically significant.

Genes that are significantly up-regulated in patients with metastasis are located towards the right-hand side of the graph and are depicted as purple boxes (fold-change > 1.5 or log₂-fold threshold > 0.58). Genes that are significantly down-regulated in metastasis patients are located towards the left-hand side of the graph and are represented as green circles (fold-change < 0.66 or log₂-fold threshold < -0.58). Both up-regulated and down-regulated genes are labeled with their corresponding Gene Symbol Identification.

4.5 Association Rules Mining Results

Furthermore, the Apriori algorithm was applied in the third phase of the proposed methodology as described Sect. 3.3 using the discretized and filtered data obtained in the first and second phase (Sects. 3.1 and 3.2) using a minimum support threshold of 0.05, a minimum confidence threshold of 0.9. Note that these measures range into the interval 0 to 1, then we are interested in obtaining

specific association rules with the highest possible reliability. The minimum and maximum length (number of items) of the association rules have been set to 3 and 5, respectively, to reduce the set of rules to be found by Apriori.

Our aim is to find potential and relevant genes highly related with sarcoma in survival time less than 5 years. As described Sect. 3.3, the consequent of the rules has been fixed to metastasis status and the rules have been filtered to obtain those that also satisfy the condition of survival time less than 5 years.

In total, we have obtained 57 non redundant and reliable rules that are satisfied by patients with presence of metastasis (1) and survival time < 5 years (0). Note that genes were discretized in two intervals to categorize the levels of expression in low (0) or high (1). Survival time was discretized as less than 5 years (0) or greater or equal to 5 years (1).

Table 1. Example of some association rules that relate relevant relationships among genes with low (0) and high (1) expression levels in patients with survival time less than 5 years ($t_survivor = 0$) and presence of metastasis.

Antecedent	Consequent	Sup	Lift	Conf
$CKAP2 = 1 \wedge GART = 0 \wedge H2AZ1 = 0 \wedge t_survivor = 0$	metastasis	0.055	1.64	1.00
$ADAMTSL4 = 1 \wedge CDC42EP2 = 0 \wedge DSTN = 0 \wedge t_survivor = 0$	metastasis	0.058	1.64	1.00
$ADAMTSL4 = 1 \wedge DSTN = 0 \wedge GMNN = 1 \wedge t_survivor = 0$	metastasis	0.055	1.64	1.00
$ADAMTSL4 = 1 \wedge ISLR = 1 \wedge TOP2A = 1 \wedge t_survivor = 0$	metastasis	0.052	1.55	0.94
$ACADVL = 1 \wedge C2orf48 = 1 \wedge FBLN5 = 1 \wedge t_survivor = 0$	metastasis	0.058	1.48	0.90
$ACADVL = 1 \wedge SPDL1 = 1 \wedge TIAM1 = 1 \wedge t_survivor = 0$	metastasis	0.061	1.49	0.90
$ATP6AP1L = 0 \wedge DPYSL3 = 0 \wedge NCAPG2 = 0 \wedge t_survivor = 0$	metastasis	0.061	1.56	0.95
$DPYSL3 = 0 \wedge NCAPG2 = 0 \wedge POLA2 = 0 \wedge t_survivor = 0$	metastasis	0.074	1.51	0.92
$CDC42EP2 = 0 \wedge DPYSL3 = 0 \wedge NCAPG2 = 0 \wedge t_survivor = 0$	metastasis	0.074	1.51	0.92

Table 1 presents a subset of the rules obtained that relate relevant relationships among genes in patients with survival time less than 5 years and presence of metastasis divided in three parts. In the first part, we present an example of rules that combines high and low gene expression levels in the antecedent. In the second part, we show an example of rules that present high gene expression levels in the antecedent. In the last part, we display an example of rules that have low gene expression levels in the antecedent. It can be seen that all the expression levels of the genes that appear in the expression levels are consistent with the results obtained in the volcano plot (up or down-regulated) in Fig. 3.

As can be observed, the confidence values of the rules are close to 1, and the lift values are higher than 1. These results indicate that the proposed study was successful in identifying relevant, reliable, and strong relationships among genes with both high and low expression levels in patients with survival times of less than 5 years and the presence of metastasis.

These findings are encouraging, as they may offer potential insights into previously unknown patterns of gene expression that could be relevant to the

development and treatment of sarcoma. They could serve as a valuable starting point for further analysis by domain experts and potentially contribute to advancements in our understanding of this disease.

5 Conclusions and Future Works

In this study, we propose a methodology for discovering relevant genes highly associated with metastasis and low survival rates in sarcoma microarrays, based on data processing, feature selection, and association rule mining. Initially, we filtered, grouped, and summarized probes corresponding to the same gene. Subsequently, we applied the FOCI and Apriori algorithms to select potential genes highly associated with metastasis and discover association rules linking high or low expression levels of these selected genes in patients with low survival rates.

The results obtained show the potential and usefulness of the proposed study for expert analysis in the field, providing insights into the complex relationship between gene expression, metastasis, and low survival rates in sarcoma patients. This methodology can also serve as a starting point for future research in the field of sarcoma and other related diseases, paving the way for more targeted and effective treatments.

As future work, we aim to analyze different discretization methods for gene expression data, apply the proposed methodology to various types of omic data, and develop more advanced algorithms for identifying relevant association rules. These efforts could lead to a deeper understanding of the mechanisms underlying sarcoma and other diseases.

Acknowledgements. The authors would like to thank the Spanish Ministry of Science and Innovation for the support under the projects PID2020-117954RB-C22 and TED2021-131311B-C21, and the Junta de Andalucía for projects PYC20 RE 078 USE.

References

1. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules in large databases. In: Proceedings of the International Conference on Very Large Databases, pp. 478–499 (1994)
2. Anguita-Ruiz, A., Segura-Delgado, A., Alcalá, R., Aguilera, C.M., Alcalá-Fdez, J.: eXplainable Artificial Intelligence (XAI) for the identification of biologically relevant gene expression patterns in longitudinal human studies, insights from obesity research. *PLoS Comput. Biol.* **16**(4), 1–34 (2020)
3. Azadkia, M., Chatterjee, S.: A simple measure of conditional dependence. *Ann. Stat.* **49**(6), 3070–3102 (2021)
4. Chibon, F., et al.: Validated prediction of clinical outcome in sarcomas and multiple types of cancer on the basis of a gene expression signature related to genome complexity. *Nat. Med.* **16**(7), 781–787 (2010)
5. Dancsok, A.R., Asleh-Aburaya, K., Nielsen, T.O.: Advances in sarcoma diagnostics and treatment. *Oncotarget* **8**, 7068–7093 (2016)
6. Dennis, G., et al.: DAVID: database for annotation, visualization, and integrated discovery. *Genome Biol.* **4**, R60 (2003)

7. García-Heredia, J.M., Pérez, M., Verdugo-Sivianes, E.M., Martínez-Ballesteros, M., Ortega-Campos, S.M., Carnero, A.: A new treatment for sarcoma extracted from combination of miRNA deregulation and gene association rules. *Sig. Transduct. Target. Ther.* **8**(1), 231 (2023)
8. Jiménez-Navarro, M.J., Martínez-Ballesteros, M., Sousa, I.S., Martínez-Álvarez, F., Asencio-Cortés, G.: Feature-Aware Drop Layer (FADL): a nonparametric neural network layer for feature selection. In: 17th International Conference on Soft Computing Models in Industrial and Environmental Applications (SOCO 2022), pp. 557–566 (2023)
9. Liu, S., et al.: Feature selection of gene expression data for cancer classification using double RBF-kernels. *BMC Bioinform.* **19**(1), 396 (2018)
10. Macías-García, L., Martínez-Ballesteros, M., Luna-Romera, J., García-Heredia, J., García-Gutiérrez, J., Riquelme-Santos, J.: Autoencoded DNA methylation data to predict breast cancer recurrence: machine learning models and gene-weight significance. *Artif. Intell. Med.* **110**, 101976 (2020)
11. Martínez-Ballesteros, M., Riquelme, J.C.: Analysis of measures of quantitative association rules. In: Corchado, E., Kurzyński, M., Woźniak, M. (eds.) HAIS 2011. LNCS (LNAI), vol. 6679, pp. 319–326. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-21222-2_39
12. Martínez-Ballesteros, M., García-Heredia, J., Nepomuceno-Chamorro, I., Riquelme-Santos, J.: Machine learning techniques to discover genes with potential prognosis role in Alzheimer’s disease using different biological sources. *Inf. Fusion* **36**, 114–129 (2017)
13. Martínez-Ballesteros, M., Nepomuceno-Chamorro, I., Riquelme, J.C.: Inferring gene-gene associations from quantitative association rules. In: 11th International Conference on Intelligent Systems Design and Applications, pp. 1241–1246 (2011)
14. Peille, A.L., et al.: Prognostic value of *PLAGL1*-specific CpG site methylation in soft-tissue sarcomas. *PLoS ONE* **8**(11), e80741 (2013)
15. Ren, J., Zhou, X., Guo, W., Feng, K., Huang, T., Cai, Y.D.: Identification of methylation signatures and rules for sarcoma subtypes by machine learning methods. *Genet. Res.* **2022** (2022)
16. Medina, A.S., Pichardo, A.G., García-Heredia, J.M., Martínez-Ballesteros, M.: Discovery of genes implied in cancer by genetic algorithms and association rules. In: Martínez-Álvarez, F., Troncoso, A., Quintián, H., Corchado, E. (eds.) HAIS 2016. LNCS (LNAI), vol. 9648, pp. 694–705. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-32034-2_58
17. Smyth, G.K.: Limma: Linear models for microarray data. R package version 3.48.3 (2021). <https://bioconductor.org/packages/release/bioc/html/limma.html>
18. Strassmann, D., et al.: Impact of sarcopenia in advanced and metastatic soft tissue sarcoma. *Int. J. Clin. Oncol.* **26**(11), 2151–2160 (2021). <https://doi.org/10.1007/s10147-021-01997-7>