Received: 27 January 2021 | Revised: 28 April 2021 | Accepted: 11 May 2021

DOI: 10.1002/int.22503

RESEARCH ARTICLE

WILEY

Improving the understanding of cancer in a descriptive way: An emerging pattern mining-based approach

Antonio Manuel Trasierras[1] | **José María Luna**[2] | **Sebastián Ventura**[2]

[1]Department of Computer Science and Numerical Analysis, University of Cordoba, Córdoba, Spain

[2]Department of Computer Science and Numerical Analysis, Andalusian Research Institute in Data Science and Computational Intelligence (DaSCI), University of Cordoba, Córdoba, Spain

Correspondence
Sebastián Ventura, Department of Computer Science and Numerical Analysis, Andalusian Research Institute in Data Science and Computational Intelligence (DaSCI), University of Cordoba, 14071 Córdoba, Spain.
Email: sventura@uco.es

Funding information
Consejería de Universidades, Junta de Andalucía, Grant/Award Number: UCO-FEDER 18 REF.1263116 MOD.A

Abstract

This paper presents an approach based on emerging pattern mining to analyse cancer through genomic data. Unlike existing approaches, mainly focused on predictive purposes, the proposal aims to improve the understanding of cancer descriptively, not requiring either any prior knowledge or hypothesis to be validated. Additionally, it enables to consider high-order relationships, so not only essential genes related to the disease are considered, but also the combined effect of various secondary genes that can influence different pathways directly or indirectly related to the disease. The prime hypothesis is that splitting genomic cancer data into two subsets, that is, cases and controls, will allow us to determine which genes, and their expressions, are associated with different cancer types. The possibilities of the proposal are demonstrated by analyzing RNA-Seq data for six different types of cancer: breast, colon, lung, thyroid, prostate, and kidney. Some of the extracted insights were already described in the related literature as good cancer bio-markers, while others have not been described yet mainly due to existing techniques are biased by prior knowledge provided by biological databases.

KEYWORDS
bioinformatics, cancer, emerging-pattern, gene-expression, RNA-Seq

1 | INTRODUCTION

Cancer is the second most common cause of death worldwide, and its incidence is estimated to be 20 million new cases per year by 2035.[1] This disease causes social and health costs that the least developed countries cannot afford. Although classic studies on the disease mainly focused on formulating a prior hypothesis to be contrasted,[2] the advent of new technologies and the possibility to perform profound analysis on genomic data have given rise to numerous research progresses to fight the disease.[3] Motivated by recent advances, new methodologies mainly based on a holistic approach have been developed,[4] not requiring the formulation of any previous hypothesis and avoiding, therefore, any bias in the analysis. However, the personalization of treatments in cancer is not still a straightforward issue and has to face different challenges. First, it includes similar phenotypes and, at the same time, distinct molecular profiles. Hence, two patients sharing the same cancer type may have completely different genetic profiles, mainly due to the dissimilarity of genes affecting different signalling pathways. Second, any statistical data analysis suffers from both thousands of genes to consider and class imbalance, adding complexity to the subsequent analytical process.

Current genomic cancer studies are generally carried out by a standardized workflow[5] that includes a first exploratory analysis with visualization methods to take a first look at the data distribution. Then, a feature selection procedure is carried out through traditional methods such as principal component analysis[6] followed by a differential expression analysis to test which genes are expressed differentially between cases and controls. Sometimes, subsequent studies of co-expression networks[7] or biclustering[8] are applied to such genes. Some authors[9] are replacing this differential expression analysis with new methodologies such as gene set analysis (GSA) to consider functional relationships at a gene level. GSA, however, requires previous knowledge usually taken from reputable biological databases such as *Gene Ontology*[10] and *Kyoto Encyclopedia of Genes and Genomes*.[11] Results produced through the standardized workflow are finally *enriched* through enrichment analysis methods to obtain insights about affected functional pathways. This widely accepted workflow, however, includes several downsides to be highlighted: (i) gene blocks produced by GSA techniques are based on the information provided by reputable biological databases; (ii) impossibility to obtain high-order relationships, so the information provided is not accurate at all.[12]

The boom of data mining has produced some studies for early cancer detection using machine learning methods.[13] In this regard, early melanoma diagnosis through image analysis has been performed recently through Deep Learning techniques.[14] Breast cancer prediction by considering risk factors also was done with alluring results.[15] The same type of cancer on Nigerian women[16] was analysed by considering classic classification methods such as Naive Bayes and J48. The above studies do not need any biological database as input, and they can consider high-order relationships. However, none of these studies focused on descriptive data mining techniques, which are essential to provide more actionable and understandable insights. Be able to denote which genes are responsible for producing each type of cancer is key to design the treatment accurately. Hence, this paper aims to propose a descriptive analysis based

on emerging pattern mining (EPM).[17] This data mining task searches discriminative patterns or elements in data which frequency increases significantly from a data subset to another. Our prime hypothesis is that splitting genomic cancer data into two subsets, that is, cases and controls, will allow us to determine which genes, and their expressions, are associated with each cancer type. The final aim is to address all the open challenges of the well-known and standardized workflow, producing understandable and actionable insights that help with the discovery of new and accurate forms of cancer treatment. The contribution of this study work can be summarized as follows:

- A methodology based on EPM to produce high-order relationships and considering the whole set of genes. This methodology enables to determine variables that are involved in the signalling pathways affected in the disease.
- Unlike the well-known standardized workflow, which requires information that is given by reputable biological databases, the proposed methodology starts from scratch. No previous knowledge or information is required, and therefore, it produces unbiased insights.
- A highly useful post-processing technique is considered for the specific problem of cancer. It enables filtering out redundant results typically produced by descriptive analysis, improving the actionability of the extracted knowledge as a result.

The proposed methodology is applied to different real scenarios related to varied cancer types. Six case studies based on RNA-Seq data taken from *The Cancer Genome Atlas* (TCGA)[18] has been considered. The obtained results demonstrate the usefulness of the proposal and how it faces all the existing challenges. Some of the extracted insights were already described in the related literature as good cancer bio-markers, while others have not been described yet mainly due to existing techniques are biased by prior knowledge provided by biological databases.

The rest of the article is structured as follows: Section 2 presents the EPM task and some studies about cancer. Section 3 describes the proposed methodology. Section 4 performs an exhaustive analysis on RNA-Seq data to describe several cancer types. Finally, some lesson learnt are outlined in Section 5, and some brief conclusions are pointed out in Section 6.

2 | PRELIMINARIES AND RELATED WORK

This section presents the EPM task and some primary definitions. Different cancer analyses considering classic biological workflows and new methods based on data mining are outlined.

2.1 | EPM

In data analysis, the major element to study is the pattern as it provides important knowledge to consider in the data domain. A pattern P is a set of items related somehow in a database Ω that describes the behaviour of the observed data. Formally speaking, a pattern P is denoted as $P \subseteq I \in \Omega$ where I is the set of n items $\{i_1, i_2, ..., i_n\}$ included in a database Ω. The pattern mining task[19] aims to find the whole set of patterns appearing at least in a fraction of database records. To this end, the metric most commonly used to set the fraction of data records to analyse is the support or frequency. In absolute terms, the support or frequency of a pattern P defines the number of data records in which P is present. This metric can also be defined in relative terms, considering the

percentage of data records in which P is present. Data records are commonly known as transactions, and T defines the collection of all transactions present in Ω. Formally, it is denoted as $T = \{t_1, t_2, ..., t_m\}$. Hence, the support of a pattern P in a database Ω is formally defined as follows:

$$Support\,(P, \Omega) = \frac{|\forall\, t \in T : P \subseteq t, t \subseteq I\,|}{|T|}. \tag{1}$$

The concept of supervised descriptive pattern mining (SDPM) appeared sometime later to retrieve patterns having a property of interest.[17] Many novel and interesting techniques appear under the umbrella of SDPM, including emerging patterns, contrast sets mining, subgroup discovery, and class association rules, among others. Focusing on the EPM task, it was described[17] as the seek for discriminative patterns which support or frequency, analysed on two different groups or data sets, significantly increases from one data set (group) to another. In other words, EPM aims at discovering patterns that clearly describe or distinguish two sets. Formally speaking, let I be the set of different items gathered from two data sets Ω_1 and Ω_2. Let these data sets to be represented in a binary tabular form (an item may be or not in data) and let $T = \{t_1^1, t_2^1, ..., t_n^1\}$ be the group of transactions or data records such as $\forall\, t^1 \in T : t^1 \subseteq I \in \Omega_1$. An emerging pattern is a pattern $P \subseteq I$ which frequency (support) vary considerably with respect to the frequency of the pattern P in data set Ω_2. The metric generally used to measure the interest of such type of patterns is the Growth Rate (GR), which defines the ratio between the frequency of a pattern P in a data set Ω_1 and its frequency in another data set Ω_2 (see Equation 2).

$$GR\,(P) = \frac{Support\,(P, \Omega_1)}{Support\,(P, \Omega_2)}. \tag{2}$$

Depending on the GR value, the EPM task groups the patterns on different categories[17]: minimal, maximal, essential, noise-tolerant, generalized noise-tolerant, chi, and so forth. From all the above, one of the most widely used is the jumping emerging pattern (JEP).[17] A JEP is a pattern that appears at least once in a data set Ω_1, but it does not appear in another data set Ω_2. As a result, a JEP presents a GR value of infinity, and it is extremely useful to describe the exclusive characteristics of each data set.

2.2 | Cancer studies

The study of cancer has been the main focus of research in recent years, and many research studies have produced a clear workflow in bioinformatics. This subsection describes classic and current standardized workflows in bioinformatics. Additionally, some key contributions using data mining techniques on cancer are outlined.

2.2.1 | Biomedical analysis

Classic research studies on any disease and, more specifically, in cancer, formulate an initial hypothesis to be contrasted. Thanks to the first microarray chips, many advances were made in cancer, and research analyses enabled to avoid any bias. Thus, existing methods were redefined, from the first models based on the Student's t tests[20] to more advanced linear models.[4]

Nevertheless, the use of microarrays has several key disadvantages when performing omic studies: (a) pre-designed probes are required, so new transcripts, gene fusions, single nucleotide variants, and small insertions or deletions could not be detected; (b) a microarray is restricted by light signal saturation in both low and high signal conditions; (c) there are problems in detecting rare or low expression genes.

The emergence of RNA-Seq technology[21] solved the above limitations. The main advantage is this technology can detect new transcripts that might serve as biomarkers, mainly because it does not require pre-designed probes for hybridization. Additionally, it does not present the signal saturation problems that microarrays have. Thanks to this alluring technology, many research studies have been performed following a standardized workflow.[5] Such a workflow begins with an initial exploratory data analysis. Subsequently, differential expression analysis is carried out to test which genes are expressed differentially between cases and controls. Those genes are sometimes analysed by means of co-expression networks[7] or biclustering[8] techniques. Some authors[9] are replacing this differential expression analysis with new methodologies such as GSA to additionally consider functional relationships at a gene level. However, it is important to remark that GSA requires previous knowledge usually taken from reputable biological databases such as *Gene Ontology*[10] and *Kyoto Encyclopedia of Genes and Genomes*.[11] Finally, results produced through the standardized workflow are *enriched* through enrichment analysis methods to obtain insights about affected functional pathways. Even when the previous is a widely accepted workflow, it suffers from two main limitations: (i) reputable biological databases are required by GSA techniques; (ii) high-order relationships cannot be obtained.

2.2.2 | Data mining approaches

Triggered by the boom of data mining, several approaches have been applied to cancer studies. Most of those applications focus on predictive models to detect cancer at early stages.[13] Thus, trending Deep Learning techniques are being used to effectively detect melanoma through image analyses.[14] Additionally, breast cancer prediction was addressed by classic techniques[16] such as Naive Bayes and J48. The accuracy of the final prediction for each type of cancer has been recently tested[15] through approaches based on support vector machines, k-nearest neighbour and best first trees. Oral cancer has also been the object of study, and Kalaiarasai et al.[22] compared the performance of several prediction models in this regard.

Additional research studies to predict different types of cancer paid attention to the way insights are provided to the users. Thus, they combined some descriptive techniques to produce accurate but understandable classifiers, a task known as associative classification.[23] Alagukumar et al.[24] proposed to build a classifier based on association rule mining. More recently, Vengateshkumar et al.[25] also proposed a methodology based on boolean association rules to build a classifier. In any of such algorithms, authors used a first feature selection process employing statistical t tests. Additionally, EPM was considered to form accurate classifiers in cancer prediction, considering acute lymphoblastic leukaemia.[26] Some authors analysed gene expression profiles associated with the stage of colon cancer in microarray data.[27] They discretized microarray data with an entropy-based method by partitioning the gene expression values into intervals of disjunctive real values for which the entropy value is minimal. Changes in expression levels have also been studied through emerging patterns to produce accurate classifiers.[28]

Even when predictive purposes have coped with most of the research studies on cancer through data mining techniques, descriptive tasks are becoming increasingly important. In a recent article, Vengateshkumar et al.[29] proposed an association rule mining approach to describe relationships in microarray data. In this approach, the authors considered statistical inference methodologies based on *t* tests, assuming a normal data distribution, and leading to the adding of bias in the study.

3 | AN APPROACH BASED ON EPM

The application of EPM to cancer analysis is of high interest due it enables to describe features that differentiate two groups of patients requiring no additional knowledge or previous hypothesis. Unlike existing research studies, mainly based on building accurate classifiers, producing useful descriptive information is vital to promote new and right cancer treatments. In this regard, this section presents an EPM approach (Figure 1 illustrates the general workflow) to describe properly genes and expression levels related to different cancer types. The proposal includes four main procedures as it is described below.

Algorithm 1 Pseudocode for gene selection procedure

Input RNA-Seq data including two groups of patients
Output Differentially expressed genes (*deg* set)
procedure DIFFERENTIAL GENE EXPRESSION ANALYSIS
  for each gene *g* in RNA-Seq data **do**
    if *CPM* < 1 in both groups **then**
      Remove *g* from RNA-Seq data
    end if
  end for
  nGenes ← calculate the number of genes in RNA-Seq data
  if *nGenes* > 0 **then**
    normData ← Perform a TMM normalization on RNA-Seq data using edgeR
    deg ← Take differentially expressed genes from *normData* using tests and *FDR* <= 0.05
  end if
return *deg*
end procedure

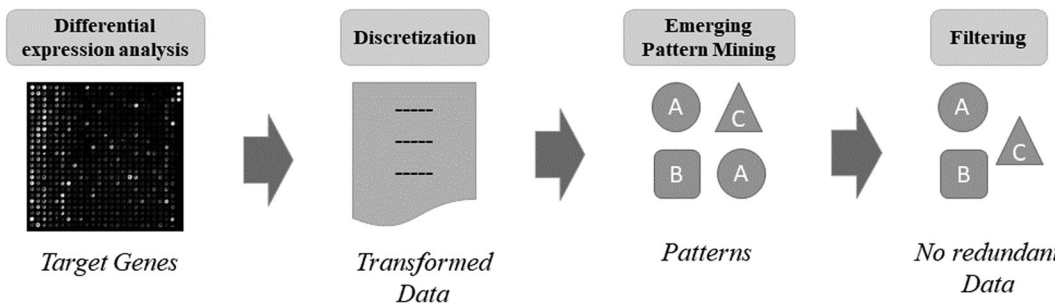

FIGURE 1 Proposed cancer analysis workflow

The first procedure (see Algorithm 1) is responsible for seeking genes with a significant differential expression between the two groups of patients (tumoral and healthy). In this first procedure, as it is a common practice in genomic analyses,[5] it is primordial to minimize the error made in the estimation of altered genes, avoiding the extraction of variables that do not contribute at all to cancer. There is no sense in analyzing genes that are expressed in none of the groups, so a predefined threshold based on counts for each million mapped reads (CPM) is considered (see Algorithm 1, Lines 3–5). A read is a raw RNA sequence obtained by sequencing machines and may be composed of multiple segments. We assume that a gene is relevant if it has at least 1 CPM across the samples of any of the conditions. Then, the proposed procedure performs a trimmed mean of M values (TMM) normalization[30] (see Algorithm 1, Line 9) to account for the variation in the total number of gene counts between samples of interest. A differential expression analysis, widely considered in RNA-Seq data analysis, is then performed (see Algorithm 1, Line 10). It takes a false discovery rate (FDR) of 0.05, a value widely accepted in bioinformatics.[31] This FDR value means a maximum of 5 false positives is allowed from 100 genes declared as differentially expressed. Finally, the set of differentially expressed genes is returned (see Algorithm 1, Line 12) and used as input data in the subsequent procedure.

It should be highlighted that the proposed gene selection procedure does not present the bias commonly considered in other approaches.[27,29] Unlike those approaches based on classic feature selection procedures[32] (entropy or correlation-feature selection), the proposed procedure considers any of the genes involved in secondary pathways and does not neglect secondary genes that contribute to the overall state of the main and/or secondary pathways. Additionally, the proposed procedure does not consider classic statistical *t* tests, which might lead to any bias due to the use of normal data distribution. On the contrary, the proposed method follows a negative binomial distribution since RNA-Seq data are mainly based on counts (negative or decimal values cannot exist).

Algorithm 2 Pseudocode for discretization procedure

```
Input Differentially expressed genes (deg set)
Output Discretized data (deg′ set)
1: procedure DISCRETIZATION
2:   deg′ ← Transposed deg data
3:   for each gene g in deg′ set do
4:     g ← Discretize g expression values into 5 bins
5:   end for
6: return deg′
7: end procedure
```

The second procedure (see Algorithm 2) performs a discretization on the expression count levels of the genes. First, the input data set (differentially expressed genes obtained in the previous procedure, see Algorithm 1) is transposed by taking genes as columns and samples as rows (see Algorithm 2, Line 2). The procedure discretizes the expression count level of each gene into five different labels corresponding to the equitable division of the interquartile range of their expression levels: EH (extremely high), H (high), M (medium), L (low), and EL (extremely low). This step is optional since data may be already discretized

beforehand (manually or following a completely different methodology). Finally, the discretized set of genes *deg′* is returned (see Algorithm 2, Line 6), so it is used as input data in the following procedure.

Algorithm 3 Pseudocode for the extraction of EPs

Input Discretized set of genes (*deg′* set)
Output All JEPs for each group G (JEP_G)
1: **procedure** FREQUENT ITEMSET MINING (LCM ALGORITHM)
2: **for** each group G in *deg′* set **do**
3: $X_G \leftarrow$ samples from *deg′* belonging to group G ▷ Generate a sub-data set X_G for each group G
4: $P \leftarrow \emptyset$ ▷ The set of all frequent patterns of the group G
5: $JEP_G \leftarrow \emptyset$ ▷ A set of all JEPs of the group G
6: $P \leftarrow$ LCM algorithm on X_G using α as minimum support value
7: **for** each frequent pattern p in P **do**
8: **if** p is JEP **then**
9: $JEP_G \leftarrow JEP_G \cup p$
10: **end if**
11: **end for**
12: **end for**
13: **return** JEP_G, X_G
14: **end procedure**

The third procedure corresponds to an EPM process. It takes the discretized gene expression counts as input and divides them into two sub-data sets: cancer and healthy patients. Then, the procedure extracts all the frequent patterns from each sub-data set by considering a minimum support threshold α, which is predefined by the user (see Algorithm 3, Line 4). Once all patterns are obtained, the procedure seeks JEP on each group or sub-data set (see Algorithm 3, Lines 5–9). Thus, the algorithm analyses intrinsic patterns related to cancer as well as intrinsic patterns for healthy patients. In other words, the procedure obtains descriptive insights from both groups, thus reinforcing the obtained results.

Algorithm 4 Pseudocode for reduce redundant data

Input JEP_G, X_G, β
Output *Res* ▷ Set of representative JEPs of a given group G.
1: **procedure** FILTER REDUNDANT DATA
2: $JPE_G \leftarrow$ Sort JPE_G by support
3: **for** each set of solutions $S \subseteq JPE_G$ having the same support **do**
4: $JPE_G \leftarrow JPE_G \setminus S$
5: $S \leftarrow$ Sort S by length (from larger to shorter)
6: $JPE_G \leftarrow JPE_G \cup S$ ▷ Add S at the same position it was originally (support ranking)
7: **end for**
8: *Res* $\leftarrow \emptyset$
9: $p \leftarrow$ Take the first element from JEP_G

(Continues)

```
10:   X_G ← mark records covered by p
11:   Res ← Res ∪ p
12:   for each p in JEP_G do
13:     if p covers at least β% new unmarked records then
14:       Res ← Res ∪ p
15:       X_G ← mark records covered by p
16:     end if
17:   end for
18: return Res
19: end procedure
```

The frequent itemset mining approach included in this third procedure (see Algorithm 3, Line 6) is the well-known LCM (linear closed itemset mining) algorithm.[19] This algorithm is the one that best performs in the frequent itemset mining task, and it was the winner of the widely recognized FIMI 2004 competition. It owes its efficiency and speed to the combination of several ideas that greatly optimize the searching procedure, especially in dense data. The algorithm performs transaction projections, database reduction methods, an efficient frequency counting structure and a hypercube decomposition method. Frequency counting is carried out through a scheme called occurrence deliver, in which the algorithm uses a list of buckets where each bucket stands for a frequent singleton in the database. After a first sweep on the database, the algorithm establishes the order of the singletons in an increasing support way. Once the projections are performed, the buckets are restarted and re-scan the conditional database produced by the projections to fill the buckets. This frequency counting method allows the algorithm to analyse databases efficiently because the database becomes smaller and smaller with each projection, and resetting and refilling the buckets is not expensive. Additionally, the use of the hypercube decomposition method, also known as perfect extension pruning, provides even more advantage over other approaches belonging to the state-of-the-art. More specifically, given a itemset G and an item $g, g \notin G$, if G and $G \cup g$ have the same frequency in the database, then g is a perfect extension of the itemset G. This method is a great gain in performance because once the perfect extensions are detected, they are reported and excluded from the recursion.

The fourth and final procedure (see Algorithm 4) aims to reduce the number of redundant results, typically provided by descriptive tasks. Due to obtained solutions can be rather large, mainly because of the number of genes to be analysed, this procedure searches for redundancy to return a representative set of solutions (patterns). The resulting set should include features describing any patient (desirably all of them) with a small and, therefore, understandable set of solutions. In this regard, the procedure first ranks solutions according to the support or frequency quality measure (see Algorithm 4, Line 2) and, if there is a tie, then solutions are ranked by length (from larger to shorter) as it is shown in Lines 3–7, Algorithm 4. The top pattern from this ranking is taken, and data records covered by it are marked (see Algorithm 4, Lines 9–11). This process is repeated, pattern by pattern, till all data records are covered, or there is no pattern to select (see Algorithm 4, Lines 12–17). The procedure considers an overlapping threshold β, previously defined by the user. This threshold denotes that at least β% of new unmarked records should be covered. The resulting set of patterns depends, to a large extent, on the data distribution. Groups of patients with dissimilar or uncommon characteristics will require a large number

of patterns to represent the observed data set. On the other hand, as usually happens in this type of diseases, patients are expected to share phenotype or similar characteristics, not requiring a large set of solutions to represent the whole set of patterns.

4 | CASE STUDIES BASED ON DIFFERENT CANCER TYPES

This section aims to illustrate the use and usefulness of the proposed approach based on emerging patterns. Six case studies are analysed to discover new biomarkers, gene associations and biological functions most affected in different cancer types, some of them already studied by dissimilar methodologies.[13,14,16,22] The hypothesis is the feasibility to discover novel insights that help with a better understanding of the disease as well as knowledge already described in the literature. RNA-Seq data for six different types of cancer are analysed: breast (BRCA), colon adenocarcinoma (COAD), lung adenocarcinoma (LUAD), thyroid carcinoma (THCA), prostate adenocarcinoma (PRAD) and kidney (KICA). Data are taken from *The Cancer Genome Atlas* (TCGA)[18] repository (data are publicly available at https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga). Data are in RSEM format and only contain paired samples (cancer vs healthy samples). Results are finally analysed through a functional enrichment analysis, considering GO terms and KEGG pathways from the Gene Ontology[10] and KEGG[11] biological databases. A further literature review has been carried out.

4.1 | BRCA analysis

Breast cancer is the most common type of cancer among women (men may also suffer from it but in an extremely low percentage).[1] It is, therefore, one of the most important types of cancer nowadays, presenting a terrific impact on society. In this case study, a real data set comprising 224 paired samples and 20,531 genes is analysed through the proposed approach.

The first procedure reduces the number of genes to 11,738 (those with a significant differential expression between the two groups of patients). This set of genes are then discretized into five labels (already described in the second procedure of the proposed approach), and an EPM algorithm is finally applied (third procedure). We considered a minimum threshold support value equal to 30% and 35% for each database condition, obtaining a total of 167,239 and 486,708 interesting patterns for cancer and healthy groups, respectively. Since this large number of solutions is hardly understandable for an expert, a filtering process is then applied (the fourth procedure of the proposed approach). As a result, the resulting set is reduced to 15 in cancerous group (see Table 1), and 15 in healthy group (see Table 2). These two sets of solutions can represent 87.5% and 83.93% of the whole set of patients. Some of such patterns or solutions are described below through a functional enrichment analysis, confirming the importance of the extracted insights.

The analysis of the first pattern (ID 1, Table 1), and focusing on the GO terms, revealed that all the genes except for C12orf48, play a crucial role in the cell cycle and usually act as oncogenes in the breast or related cancers.[33] The thorough revision of the specialized literature in the field revealed that C12orf48 generally encodes a PARP-1 binding protein called PARPBP, which function is the inhibition of the DNA repair mechanism through homologous

TABLE 1 Final set of patterns obtained from the cancerous group of patients

ID	Patterns	Coverage (% of data)
1	CCNB1 = EH, CDK1 = EH, KIF23 = EH, NUSAP1 = EH, CCNB2 = EH, MELK = EH, NCAPG = EH, BUB1B = EH, NCAPH = EH, C12orf48 = EH	30.36
2	TPX2 = EH, CDK1 = EH, BUB1 = EH, KIF23 = EH, FANCI = EH, CCNB2 = EH, NCAPG = EH, SGOL1 = EH, BUB1B = EH, C12orf48 = EH	32.14
3	CDK1 = EH, SPC25 = EH, ERCC6L = EH, BUB1 = EH, SMC4 = EH, SGOL1 = EH, CENPI = EH, BUB1B = EH, NCAPH = EH, ARHGAP11A = EH	34.82
4	GPD1 = EL, FABP4 = EL, CIDEC = EL, ADH1A = EL, TUSC5 = EL, ADH1B = EL, ADIPOQ = EL, CIDEA = EL	52.68
5	CCNB1 = EH, KIF11 = EH, KIF23 = EH, MELK = EH, NCAPG = EH, SGOL1 = EH, NCAPH = EH, CDC7 = EH	54.46
6	ANLN = EH, CENPE = EH, EXO1 = EH, CCNA2 = EH, ASPM = EH, MCM10 = EH, PLK4 = EH, MND1 = EH	57.14
7	FABP4 = EL, ADH1A = EL, ADH1B = EL, ADIPOQ = EL, CIDEA = EL, MEOX1 = EL	58.93
8	LDB2 = EL, CXorf36 = EL, MMRN2 = EL, ECSCR = EL, MYCT1 = EL, CLEC14A = EL	61.61
9	LDB2 = EL, GNG11 = EL, CXorf36 = EL, CD34 = EL, MYCT1 = EL	64.29
10	AQP7 = EL, FABP4 = EL, C14orf180 = EL, TUSC5 = EL, CIDEA = EL	66.07
11	FAP = EH, COL5A2 = EH, VCAN = EH, COL1A2 = EH, ADAM12 = EH	80.36
12	C1QTNF6 = EH, COL5A2 = EH, COL5A1 = EH, COL1A1 = EH, COL1A2 = EH	82.14
13	ERCC6L = EH, RACGAP1 = EH, ECT2 = EH, SGOL1 = EH, CASC5 = EH	83.93
14	LDB2 = EL, CXorf36 = EL, CD34 = EL, ERG = EL	85.71
15	C1QTNF6 = EH, COL5A1 = EH, COL1A1 = EH, OLFML2B = EH	87.50

Note: The labels next to the genes refer to gene expression values: extremely low (EL), low (L), medium (M), high (H) and extremely high (EH).

recombination. Several studies pointed out the upregulation of C12orf48 in several types of cancer such as hepatocellular cancer.[34] Such studies also suggested its possible role as an oncogene as well as within the cell cycle. As a result, it is possible to assert that C12orf48 participates in the cell cycle in breast cancer and could share a common regulatory framework with the genes that belong to the pattern with ID 1. This is a clear example of the potential of using pattern mining techniques within the proposed framework, not requiring any previous knowledge or hypothesis to reach excellent and unknown insights.

Pattern with ID 8 (see Table 1) is another interesting result to be considered. Genes included in this pattern were already detected in a recent study,[35] denoting a relationship with endothelial functions. Interestingly, the biological function of CXorf36 remains understudied, while that of CLEC14A appears to be the promotion of angiogenesis activities and cancer migration.[36] This gene expression signature should be considered in future

TABLE 2 Final set of patterns obtained from the adjacent tissue group of samples

ID	Patterns	Coverage (% of data)
16	TNS1 = EH, SORBS1 = EH, HSPB6 = EH, KCNIP2 = EH, AOC3 = EH, C7orf41 = EH, CAT = EH, PLXNA4 = EH, PLIN1 = EH, AQP7 = EH, FZD4 = EH, TMEM37 = EH, TLN2 = EH, HEPN1 = EH, AIFM2 = EH, MOSC1 = EH, PCDH9 = EH, ACO1 = EH	35.71
17	HSPB6 = EH, KCNIP2 = EH, AOC3 = EH, LIPE = EH, GYG2 = EH, AQP7 = EH, GPD1 = EH, TMEM37 = EH, HEPN1 = EH, AIFM2 = EH, CIDEC = EH, PNPLA2 = EH, TYRO3 = EH, SGCG = EH, ALDH2 = EH	37.50
18	TNS1 = EH, MYOM1 = EH, ACACB = EH, C7orf41 = EH, SIK2 = EH, AQPEP = EH, CAT = EH, ACVR1C = EH, PLIN1 = EH, AIFM2 = EH, FERMT2 = EH, LOC283392 = EH	39.29
19	LPL = EH, AQP7 = EH, MLXIPL = EH, SLC7A10 = EH, MOSC1 = EH, CIDEC = EH, LGALS12 = EH, ADIPOQ = EH	41.07
20	C7orf41 = EH, AKAP12 = EH, MMD = EH, FERMT2 = EH, ECM2 = EH, ITSN1 = EH, C2CD2 = EH	42.86
21	TNS1 = EH, ACACB = EH, PLXNA4 = EH, C14orf49 = EH, HEPACAM = EH	44.64
22	FRMD1 = EH, PNPLA2 = EH, NMB = EH, VEGFB = EH	46.43
23	LIPE = EH, GPD1 = EH, TUSC5 = EH, CIDEA = EH	48.21
24	KIF4A = EL, CEP55 = EL, IQGAP3 = EL	62.5
25	CAV1 = EH, CAV2 = EH, KLB = EH	64.29
26	IQGAP3 = EL, KIF18B = EL	66.96
27	CHRDL1 = EH, PCOLCE2 = EH	69.64
28	SLC25A39 = EL, TIMM17B = EL	79.46
29	FAM13A = EH, CDC14B = EH	82.14
30	GPAM = EH, DGAT2 = EH	83.93

Note: The labels next to the genes refer to gene expression values: extremely low (EL), low (L), medium (M), high (H) and extremely high (EH).

research studies as a possible mechanism for developing antiangiogenic therapies in breast cancer.

Finally, it is also important to highlight the pattern with ID 17 (see Table 2), which generally represents the specific functions of the healthy tissue group. Some of the genes belonging to this pattern (LIPE, GYG2, AQP7, GPD1, CIDEC and PNPLA2) implies the deregulation of lipid metabolism in cancer (these genes are found with a low level of expression in breast cancer samples compared to healthy samples). The lack of lipases in breast cancer is associated with alternative means of obtaining energy from tumour cells such as glucose and glutamine metabolism. These results are similar to some found in literature[37]: *the downregulation of lipid metabolism through the absence of lipases and free fatty acids as well as low activity of PPAR-γ*

signalling are key events in the formation of various types of cancer such as liposarcomas. Note that some genes within this pattern, for example, HSPB6, have been already linked to tumour suppressors due to their low expression in breast cancer.[38] However, it is quite interesting to note that no interaction nexus has been described yet between the genes related to lipid metabolism and the rest of the genes in the pattern.

4.2 | COAD analysis

Colon adenocarcinoma is a type of cancer that comes into origin in those cells that are responsible for producing the mucus that lubricates both the colon and rectum. It is the most frequent type of colon cancer, accounting for 95% of cases and the third most common type of cancer in both men and women (producing around 9% of all cancer deaths).[1] More than 90% of colorectal cancer occurs from the age of 50 onwards, whereas the average age of diagnosis is 72. This case study analyses a real data set comprising 52 paired samples and 20,531 genes through the proposed approach.

The first procedure reduces the number of genes to a total of 8,642 differentially expressed genes. Then, the gene values are discretized (second procedure), and the EPM algorithm is finally run on the discretized data (third procedure). In this case study, a pattern is frequent if it appears in at least 40% of the records in the cancer group, and 42% in the healthy group. A total of 132,638 and 9,013,927 patterns were obtained for the cancerous and healthy group, respectively. The filtering process described in the fourth procedure reduced the set of patterns to 5 for the cancerous group (see Table 3), and 5 for the healthy group (see Table 4). As a result, considering just 5 patterns, it is feasible to describe any patient belonging to the cancerous group. 5 additional patterns are taken to describe 80.77% of the samples in the healthy group. Some of these patterns are described below through a functional enrichment analysis.

TABLE 3 Final set of patterns obtained from the cancerous group of patients

ID	Patterns	Coverage (% of data)
1	NBLA00301 = EL, LMOD1 = EL, GPM6A = EL, SYNM = EL, KCNMB1 = EL, CNN1 = EL, UCHL1 = EL, HAND2 = EL, DACT3 = EL, C7 = EL, PDLIM3 = EL, SLITRK3 = EL, RERG = EL, CHRDL1 = EL, NAP1L3 = EL, HSPB7 = EL, MGP = EL	42.31
2	PHF14 = EH, UTP20 = EH, BAZ1B = EH, RBAK = EH, AVL9 = EH, TNPO3 = EH, TRRAP = EH, TGFBRAP1 = EH, RNASEN = EH	69.23
3	CDK5RAP1 = EH, DDX27 = EH, ASXL1 = EH, TM9SF4 = EH, PHF20 = EH, POFUT1 = EH, NCOA6 = EH, SULF2 = EH	80.77
4	CDH11 = EH, ZNF469 = EH, ITGA11 = EH, MMP14 = EH, MFRP = EH, COL8A1 = EH, PLXDC1 = EH	88.46
5	ZBTB33 = EH, RBMX2 = EH, EMD = EH, ENOX2 = EH, VBP1 = EH, TMEM187 = EH	100.00

Note: The labels next to the genes refer to gene expression values: extremely low (EL), low (L), medium (M), high (H) and extremely high (EH).

TABLE 4 Final set of patterns obtained from the healthy group of samples

ID	Patterns	Coverage (% of data)
6	PDZD4 = EH, FAM129A = EH, PYGM = EH, ARL4D = EH, GPM6A = EH, HSPB6 = EH, KCNIP3 = EH	42.31
7	ATP11A = EL, PLXNA1 = EL, XPO4 = EL, RRP1B = EL, RBM19 = EL, PFAS = EL, AHCTF1 = EL	65.38
8	CIRH1A = EL, GARS = EL, NOP58 = EL, LYAR = EL, DCUN1D5 = EL, METTL1 = EL, NAA15 = EL	73.08
9	RAD18 = EL, PYCR1 = EL, NUF2 = EL, CDCA4 = EL, KIF18A = EL, SHCBP1 = EL, TRAIP = EL	76.92
10	EXOSC8 = EL, CCDC59 = EL, SNRPE = EL, SNRPD2 = EL, TOMM6 = EL, GEMIN6 = EL, POLR2G = EL	80.77

Note: The labels next to the genes refer to gene expression values: extremely low (EL), low (L), medium (M), high (H) and extremely high (EH).

Let us first focus on the pattern with ID 1 (see Table 3), which comprises 17 genes with a low level of expression in cancer samples. The enrichment analysis and the subsequent bibliographical review[39] described that some of these genes are related to altered functions in cancer such as poor cell differentiation (CHRDL1, MGP and HAND2) and migration (GPM6A, SYNM, CNN1, HAND2, PDLIM3, RERG, CHRDL1 and MGP). Both functions are intimately related since undifferentiated cancer cells tend to have a greater capacity for growth and migration.[39]

Another interesting pattern is the one with ID 4 (see Table 3), which is composed of structural genes related to angiogenesis and cell adhesion processes. Alterations of these genes have been described in disturbing cell adhesion between cancer cells and the extracellular matrix. They are also known for their interaction with tyrosine kinase receptors that promote the proliferation of cancer cells and their differentiation.[40] MFRP, the only gene that does not have an annotated structural function, is upregulated in several cancers and could activate the Wnt pathway by increasing signalling of cell proliferation, migration and angiogenesis.[41]

Finally, it is also interesting to highlight the pattern with ID 10 (see Table 4), most of its genes (SNRPE, SNRPD2, GEMIN6 and POLR2G) presenting functions in the splicing of the mRNA through the spliceosome. Mutations in the spliceosome components may cause aberrant splicing of mRNAs with oncogenic properties.[42] Moreover, deregulation in the levels of expression of its structural components is related to the most aggressive phenotypes of cancer.

4.3 | LUAD analysis

Lung cancer is the most commonly diagnosed type of cancer and the one with the highest mortality rate.[1] Among their subtypes, lung adenocarcinoma is the most common one, and it represents around 40% of all lung cancers, affecting even nonsmoking people. The present case study analyses a real data set comprising 114 paired samples and 20,531 genes.

The gene set was first reduced to a total of 10,074 genes with significant differential expression between the two groups of patients. Then, all these genes were discretized into five labels, and the EPM algorithm (third procedure) was applied to obtain patterns that appear in at least 35% (cancer group) and 36% (healthy group) of the records. As a result, we obtained 66,423 and 43,730 solutions for the cancerous and healthy group, respectively. To overcome the high number of feasible solutions, we applied the proposed filtering procedure, reducing the set of patterns to 11 for the cancerous samples (see Table 5) and 23 for the healthy group (see Table 6). The extracted patterns were able to represent 100% of all the samples for both groups. Finally, a functional enrichment analysis is carried out on the resulting set to increase the interpretability of the results.

The analysis of the extracted patterns, and especially patterns with ID 1, 2 and 3, Table 5, reveals upregulation of genes belonging to the cell cycle, mitosis, and structural functions related to microtubules. Focusing on the largest pattern (ID 1), the functional enrichment analysis and a subsequent literature review revealed that any of the genes included in this pattern are essential in the cell cycle and mitosis. Additionally, genes within the pattern with ID 4 (see Table 5) belong to the focal-adhesion-PI3K-Akt-mTOR signalling pathway. It relates many crucial aspects of cell growth and survival that are constitutively activated.[43]

On the other hand, genes included in patterns belonging to the samples with healthy tissue also revealed altered functions in cancer. For example, pattern with ID 13 (see Table 6) includes genes that encode for proteasome components (PSMB4, PSMC4 PSMB3) as well as genes

TABLE 5 Final set of patterns obtained from the cancerous group of patients

ID	Patterns	Coverage (% of data)
1	DLGAP5 = EH, TPX2 = EH, KIF20A = EH, KIF2C = EH, CDCA8 = EH, HMMR = EH, KIF23 = EH, TTK = EH, SKA3 = EH, CDCA5 = EH, KIFC1 = EH, SKA1 = EH, C15orf42 = EH, AURKB = EH	36.84
2	TOP2A = EH, CDC6 = EH, STIL = EH, CKAP2L = EH, KIF15 = EH, KIFC1 = EH, SKA1 = EH, NCAPG2 = EH, LMNB1 = EH, ATAD5 = EH, BRCA1 = EH	40.35
3	KIF23 = EH, CCNA2 = EH, SPC25 = EH, PLK4 = EH, RAD51 = EH, RACGAP1 = EH, FAM54A = EH, C15orf23 = EH	43.86
4	THBS2 = EH, COL5A2 = EH, COL1A1 = EH, COL3A1 = EH, COL5A1 = EH, COL1A2 = EH	57.89
5	CCNB1 = EH, BUB1B = EH, KIF11 = EH, PRC1 = EH, ARHGAP11A = EH	61.40
6	RBM28 = EH, NOL10 = EH, NUP85 = EH, XPO1 = EH	64.91
7	GIMAP6 = EL, STARD8 = EL, TLR4 = EL, RAB8B = EL	85.96
8	C1QA = EL, C1QB = EL, IFI30 = EL	92.98
9	ITLN2 = EL, SLC6A4 = EL	94.74
10	MGC29506 = EH, TNFRSF17 = EH	98.25
11	EPM2A = EL, FAM13B = EL	100.00

Note: The labels next to the genes refer to gene expression values: extremely low (EL), low (L), medium (M), high (H) and extremely high (EH).

TABLE 6 Final set of patterns obtained from the adjacent tissue group of samples

ID	Patterns	Coverage (% of data)
12	PPIB = EL, SEC13 = EL, METTL11A = EL, PSMC4 = EL, PFDN2 = EL, HM13 = EL, P4HB = EL, YIF1A = EL, ROMO1 = EL, EIF5A = EL, JTB = EL, CALR = EL, PMM2 = EL, TMED9 = EL	35.09
13	TPI1 = EL, PSMB4 = EL, CCT3 = EL, METTL11A = EL, PSMC4 = EL, RAG1AP1 = EL, NHP2 = EL, PSMB3 = EL, ERGIC3 = EL, CYC1 = EL, ALDOA = EL, SEPX1 = EL, EIF4A3 = EL	38.60
14	PSMB4 = EL, CCT3 = EL, ILF2 = EL, METTL11A = EL, MRPL3 = EL, SNRPD2 = EL, ERGIC3 = EL, EIF3I = EL, CYC1 = EL, EIF4A3 = EL	40.35
15	PSMB4 = EL, KRTCAP2 = EL, SNRPG = EL, RAG1AP1 = EL, HSPE1 = EL, SSBP1 = EL, CHCHD8 = EL, TIMM10 = EL, TMED9 = EL	42.11
16	HMGA1 = EL, HM13 = EL, GAPDH = EL, DOLPP1 = EL, SRPRB = EL, ECE2 = EL, MRTO4 = EL	43.86
17	FARSB = EL, CCT3 = L, C14orf166 = EL, MRPL3 = EL, DAP3 = EL, BTF3L4 = EL	45.61
18	PDIA6 = EL, ILF2 = EL, HM13 = EL, DDOST = EL, HSPE1 = EL, MRPL24 = EL	47.37
19	MRPS12 = EL, DOLPP1 = EL, ALG3 = EL, SLC25A39 = EL, FLAD1 = EL	49.12
20	CLNS1A = EL, GPR89A = EL, DAP3 = EL, PRMT3 = EL, MRPL30 = EL	50.88
21	APEX1 = EL, PPP1R14B = EL, NPM3 = EL, PYCR1 = EL	52.63
22	IMP4 = EL, ALDOA = EL, GPI = EL, P4HB = EL	54.39
23	FARSB = EL, MRPL9 = EL, AAMP = EL, VPS72 = EL	56.14
24	FGFR4 = EH, RNF182 = EH, LRRN3 = EH, CCDC48 = EH	78.95
25	PAICS = EL, ALG8 = EL, CCT6A = EL	80.70
26	COPG = EL, HSP90B1 = EL, STK39 = EL	84.21
27	LIMS2 = EH, SH2D3C = EH, USHBP1 = EH	87.72
28	FIGF = EH, SGCA = EH, C1QTNF7 = EH	89.47
29	GFPT1 = EL, XPR1 = EL	91.23
30	SLC35F2 = EL, CTHRC1 = EL	92.98
31	GPR172A = EL, SLC1A5 = EL	94.74
32	POLG2 = EL, DDX12 = EL	96.49
33	ADRB2 = EH, STX11 = EH	98.25
34	RAMP3 = EH, RGS9 = EH	100.00

Note: The labels next to the genes refer to gene expression values: extremely low (EL), low (L), medium (M), high (H) and extremely high (EH).

related to glucose metabolism (TPI1, RAG1AP1, and ALDOA) with a low level of expression in healthy samples. According to our data samples, these genes are upregulated in lung adenocarcinoma. CCT3 is a coding gene for a molecular chaperone with a role in protein folding that is upregulated in several cancers and promotes greater proliferation.[44] Several cooperation functions have been reported between the chaperones and the proteasome, which may be linked to the degradation of tumour suppressor proteins in cancer.

4.4 | THCA analysis

Thyroid cancer is the eleventh most commonly diagnosed cancer in 2018.[1] Although the incidence of this cancer has increased in recent years, its death rate remains practically the same, and its prognosis is, in general, really favourable. This case study analyses a real data set comprising 118 paired samples and 20,531 genes through the proposed approach.

The total amount of genes is first reduced to 9244 genes, considering a differential expression analysis between the two groups. A discretization process and the EPM algorithm are carried out (second and third procedures), extracting a total of 258,536 patterns from the cancer group and 6370 patterns from the healthy group. Patterns obtained in both groups had to overcome a minimum support threshold of 30% to consider them as frequent. The filtering procedure is then applied to all these patterns to keep the most representative ones. As a result, 14 patterns can cover 100% of the samples that suffer from cancer (see Table 7), whereas 17 patterns can explain 96.61% of the healthy samples (see Table 8). Finally, a functional enrichment analysis is carried out on the whole set of discovered patterns (31 in total).

Focusing on the genes belonging to pattern with ID 7 (see Table 7), it is remarkable that many of the genes, such as CD47,[45] have been already studied in the specialized literature and were described as oncogenes. The enrichment analysis also revealed that several genes play a crucial role during the MAPK signalling cascade (NOD1, PDE5A and TGFA). As for the pattern with ID 10 (see Table 7), the functional enrichment analysis and the literature revision proof that its genes (RHOBTB3, RNF144A and USP31) are involved in ubiquitination processes, which deregulation can cause tumorigenesis.[46]

Finally, the pattern with ID 21 (see Table 8) gathers a set of genes that encode proteins with zinc finger domains that have been involved in transcription regulation. Interestingly, KIAA1383, also known as MAP10, is a cell cycle regulator that promotes microtubule stability. Several studies have already reported its low level of expression in cancer.[47] Many proteins that play crucial functions during the cell cycle have expression patterns that are usually regulated at the transcriptional level.[48]

4.5 | PRAD analysis

Prostate cancer is the most common cancer in men (especially in older men), existing 1,276,106 new cases and 358,989 deaths in 2018.[1] More specifically, the adenocarcinoma of the prostate is the most common type of prostate cancer (90% of prostate cancers). This case study analyses a real data set comprising 100 paired samples and 20,531 genes through the proposed approach.

First, the set of genes is reduced to 7899 genes, that is, those having a significant differential expression between the two groups. Then, the set of genes is discretized into five different

TABLE 7 Final set of patterns obtained from the cancer group of patients

ID	Patterns	Coverage (% of data)
1	CDH6 = EH, ITGA2 = EH, RUNX1 = EH, TUBB3 = EH	35.59
2	CD55 = EH, SLPI = EH, ANXA2 = EH, ANXA2P2 = EH	45.76
3	TCF7L1 = EL, PKNOX2 = EL, FLRT1 = EL	54.24
4	DOCK9 = EH, CDC42BPG = EH, PTPRF = EH	64.41
5	EPS8 = EH, IGF2BP2 = EH, LLGL1 = EH	67.80
6	MGAT4B = EH, CTSA = EH, PI4K2A = EH, MAPKAPK3 = EH, ARL6IP5 = EH, RABGGTB = EH, LACTB2 = EH, GLRX3 = EH	69.49
7	TGFA = EH, SHROOM4 = EH, EPS8 = EH, NOD1 = EH, PDE5A = EH, NRP2 = EH, MED13 = EH, CD47 = EH	71.19
8	S100A13 = EH, C19orf53 = EH, GNPTG = EH, ARF5 = EH, POLR2I = EH, CHMP2A = EH, RPS24 = EH	84.75
9	DNAH7 = EL, POLI = EL, THAP9 = EL, RSPH4A = EL, GPAM = EL	86.44
10	TTL = EH, RHOBTB3 = EH, RNF144A = EH, SNX30 = EH, USP31 = EH	88.14
11	DOCK9 = EH, ZMAT3 = EH, COL8A1 = EH, ASCC3 = EH, IGF2R = EH	89.83
12	OSR1 = EL, PODN = EL, DCN = EL, TMEM119 = EL, WISP2 = EL	96.61
13	LCN2 = EH, ST14 = EH, FTH1 = EH, LACTB2 = EH, GLRX3 = EH	98.31
14	DCN = EL, FBLN2 = EL, PRELP = EL	100.00

Note: The labels next to the genes refer to gene expression values: extremely low (EL), low (L), medium (M), high (H) and extremely high (EH).

labels, and the EPM algorithm is run. In this case, the minimum support thresholds considered were 32% for the cancer samples and 34% for the healthy ones. After the third step, 56,988 and 40,854 patterns were obtained for both groups. Then, to reduce the number of solutions and keep the most representative ones, the filtering procedure was applied. In this way, the set of cancer samples was reduced to 21 patterns (see Table 9), whereas the set of patterns from the healthy samples was reduced to 15 representative patterns (see Table 10). These solutions explain 90% of both groups. For a better functional interpretation, a functional enrichment analysis is applied to these sets of patterns.

Among all the obtained patterns, it is interesting to focus first on the pattern with ID 3 (see Table 9). In this pattern, based on the functional enrichment analysis, 9 of the 14 genes it includes are related to cell proliferation. Among them, ID1, SMARCD3, CTF1, TP63 and ALDH3A1 enrich the term positive regulation of cell regulation and the regulation of cell proliferation. These genes have appeared with a low level of expression in tumour samples from our prostate cancer study. In particular, TP63 is an already known tumour suppressor which expression is undetectable in prostate adenocarcinoma.[49] Additionally, ID1 and SMARCD3 have been found upregulated in prostate cancers.[50] However, the obtained results present these genes with a low-level expression in tumour cells. These results could be due to the combined activity of the rest of the genes in the pattern, giving rise to cancer. It is interesting to note the

TABLE 8 Final set of patterns obtained from the adjacent tissue group of samples

ID	Patterns	Coverage (% of data)
15	FGFR1OP = EH, RPS6KA5 = EH, IRF6 = EH, WDR47 = EH, RBMXL1 = EH, SLC38A2 = EH	30.51
16	NARG2 = EH, CPNE3 = EH, SNURF = EH, IPO8 = EH, SMARCA2 = EH, SLMAP = EH	45.76
17	C11orf30 = EH, TAF3 = EH, SLC10A7 = EH, RANBP10 = EH, RG9MTD3 = EH, SHPRH = EH	50.85
18	CPNE3 = EH, FAM82B = EH, FAM35A = EH, SMARCA2 = EH, RABL3 = EH	54.24
19	FAM126B = EH, FBXL3 = EH, ASPH = EH, GABPA = EH, DCUN1D1 = EH	55.93
20	ZFP62 = EH, LOC653501 = EH, TTC30B = EH, C8orf48 = EH, ZNF285 = EH	61.02
21	LOC653501 = EH, ZNF658 = EH, ZNF214 = EH, ZNF230 = EH, KIAA1383 = EH	62.71
22	ZNF2 = EH, SENP8 = EH, ZNF396 = EH, ZNF346 = EH, LOC158572 = EH	64.41
23	C4orf38 = EH, ZKSCAN3 = EH, ZNF658 = EH, FLRT2 = EH, ZNF396 = EH 66.1	
24	TNFRSF1B = EL, GPRIN1 = EL, RUNX1 = EL, ST8SIA4 = EL	76.27
25	TGFBI = EL, TNFRSF1B = EL, RUNX1 = EL, ST8SIA4 = EL	77.97
26	PODN = EH, MFAP4 = EH, PID1 = EH, ANKRD35 = EH	81.36
27	MFAP4 = EH, FEZ1 = EH, PID1 = EH	83.05
28	QDPR = EH, FAM165B = EH, NGFRAP1 = EH	84.75
29	RHOC = EL, UBTD1 = EL, MPG = EL	93.22
30	CLTA = EL, CD63 = EL, PLA2G16 = EL	94.92
31	LRRN4CL = EH, F10 = EH, ANKRD35 = EH	96.61

Note: The labels next to the genes refer to gene expression values: extremely low (EL), low (L), medium (M), high (H) and extremely high (EH).

presence of GSTP1 in the pattern, which downregulation is very common in prostate cancer (about 80% of cases[51]). Hence, this pattern suggests an interaction that could require further research studies.

Another interesting pattern is the one with ID 12 (see Table 9), which involves genes related to the pathway Wnt. Among them, RNF213, CHD8, and UBE2O have been already studied[52] and marked as related to such a pathway. The rest of the genes (TMEM48 and PPFIA1) might reveal any kind of functional implication in the pathway.

The pattern with ID 14 (see Table 9) contains altered genes in erythroleukemia cells, a variety of acute leukaemia. Also, two of its genes, ARID1A and NRIP1 are associated with signalling carried out by the androgen receptor pathway, one of the main mutated signalling pathways in prostate cancer. This signalling pathway regulates multiple cellular events such as proliferation, apoptosis, migration, invasion and differentiation.[53]

Finally, the genes included in the pattern with ID 23 (see Table 10) are related to dysfunctions in muscle contraction. Muscle contraction dysfunction is a prevalent phenomenon in a wide variety of cancers regardless of stage or nutritional status,[54] and the following genes

TABLE 9 Final set of patterns obtained from the cancer group of patients

ID	Patterns	Coverage (% of data)
1	CTF1 = EL, RND2 = EL, ST6GALNAC2 = EL	36.00
2	CTF1 = EL, RND2 = EL, PCDH7 = EL	38.00
3	GSTP1 = EL, CTF1 = EL, SLC2A9 = EL, GJA1 = EL, S100A16 = EL, ALDH3A1 = EL, SMARCD3 = EL, ID1 = EL, NGFR = EL, ITGB4 = EL, TP63 = EL, TPRG1 = EL, GPX2 = EL, PDPN = EL	40.00
4	XYLB = EH, UBE2O = EH, ZNF253 = EH, DNAJC16 = EH, ZNF828 = EH, ARID1A = EH, SMCR8 = EH, TMEM48 = EH, CKAP5 = EH, ZBTB44 = EH, CELF1 = EH, SPTLC2 = EH, C10orf46 = EH	56.00
5	GSTP1 = EL, SLC2A9 = EL, S100A14 = EL, ALDH3A1 = EL, SMARCD3 = EL, WFDC2 = EL, UBE2QL1 = EL, ID1 = EL, PHYHIP = EL, ST6GALNAC2 = EL, GPX2 = EL, NIPAL4 = EL, ITGB4 = EL	58.00
6	ZNF664 = EH, UBE2O = EH, PAXIP1 = EH, UBAP2 = EH, SYNJ2 = EH, ZNF253 = EH, GRSF1 = EH, CKAP5 = EH, ZBTB33 = EH, ADNP = EH	60.00
7	EFNB1 = EL, ETNK2 = EL, CAV2 = EL, EFEMP2 = EL, TCF7L1 = EL, ANO1 = EL, SOX15 = EL, FBXO17 = EL, GAS6 = EL, DKK3 = EL	62.00
8	SLC2A9 = EL, ALDH3A1 = EL, TPRG1 = EL, GPX2 = EL, GJB5 = EL, GJB4 = EL, CSTA = EL, TP63 = EL, NTN4 = EL	64.00
9	CSRP1 = EL, ACTC1 = EL, MPP2 = EL, JPH4 = EL, SRD5A2 = EL, TMEM35 = EL, HRNBP3 = EL	66.00
10	HSPB8 = EL, MPP2 = EL, KY = EL, LDB3 = EL, PGM5P2 = EL, ATP1A2 = EL	68.00
11	LAMB3 = EL, GPR87 = EL, CHST9 = EL, TRIM29 = EL, KRT5 = EL, LOC642587 = EL	70.00
12	UBE2O = EH, TMEM48 = EH, RNF213 = EH, CHD8 = EH, PPFIA1 = EH	72.00
13	SERPINB1 = EL, ID1 = EL, TPRG1 = EL, FHL2 = EL, NGFR = EL	74.00
14	RFFL = EH, UBE2O = EH, NRIP1 = EH, ARID1A = EH, CELF1 = EH	76.00
15	ANGPT1 = EL, TPM1 = EL, POPDC2 = EL, PRICKLE2 = EL	78.00
16	MYO10 = EH, ZNF609 = EH, RREB1 = EH	80.00
17	SNAP25 = EL, ALDH1A2 = EL, CCBE1 = EL	82.00
18	PLS3 = EL, ELF4 = EL, ANTXR1 = EL	84.00
19	KY = EL, LDB3 = EL, CCBE1 = EL	86.00
20	SERPINB1 = EL, MECOM = EL, AMIGO2 = EL	88.00
21	LARS = EH, POLR1A = EH, ANKHD1-EIF4EBP3 = EH	90.00

Note: The labels next to the genes refer to gene expression values: extremely low (EL), low (L), medium (M), high (H) and extremely high (EH).

TABLE 10 Final set of patterns obtained from the adjacent tissue group of samples

ID	Patterns	Coverage (% of data)
22	TMSB15A = EL, SYNGR2 = EL, RABEP2 = EL, C19orf46 = EL, MPST = EL, STRA13 = EL, ATP6V1G1 = EL, SPDEF = EL	36.00
23	MYL9 = EH, CNN1 = EH, KCNMB1 = EH, CSRP1 = EH, TPM1 = EH, TAGLN = EH, TPM2 = EH	58.00
24	CNN1 = EH, ACTG2 = EH, CSRP1 = EH, TPM1 = EH, DBNDD2 = EH	60.00
25	MYL9 = EH, GEFT = EH, TPM1 = EH, TAGLN = EH, TPM2 = EH	62.00
26	CCDC82 = EH, IL15 = EH, SEPT10 = EH, HMGN4 = EH, CHURC1 = EH	70.00
27	CNN1 = EH, MYL9 = EH, GCOM1 = EH, CSRP1 = EH	72.00
28	CCDC82 = EH, VAMP3 = EH, CHURC1 = EH, HMGN4 = EH	74.00
29	FAM76A = EH, IL15 = EH, KLHDC1 = EH, CCDC82 = EH	76.00
30	HEBP2 = EL, C2orf79 = EL, TMSB15A = EL	78.00
31	GSTZ1 = EL, POLD4 = EL, SLC35C2 = EL, PRKCZ = EL, BAIAP2 = EL	80.00
32	HEBP2 = EL, STRA13 = EL, H2AFJ = EL, NECAB3 = EL	82.00
33	TGFB1I1 = EH, PDLIM7 = EH, HAPLN2 = EH, TAGLN = EH	84.00
34	DDAH1 = EL, ADM2 = EL, GNPNAT1 = EL	86.00
35	RPS2 = EL, EEF1G = EL, CENPM = EL	88.00
36	PLBD1 = EH, UNC5B = EH, KSR1 = EH	90.00

Note: The labels next to the genes refer to gene expression values: extremely low (EL), low (L), medium (M), high (H) and extremely high (EH).

have been related to that: MYL9, KCNMB1, TPM1, TAGLN, TPM2 and CNN1. CNN1 is also a tumour suppressor that encodes an actin-binding protein that stabilizes the actin filamentous systems.[55] Also, in String,[56] the protein encoded by the last gene in the pattern (CSRP1) appears as having a predicted interaction with all the other genes in the pattern.

4.6 | KICA analysis

Renal cell cancer is the most common type of kidney cancer. The incidence of kidney cancer was recently estimated to be 403,262 new cases and 175,098 deaths in 2018.[1] Age, race, and gender play a role in the disease. This kind of cancer is more common in men between the ages of 60 and 80. This case study analyses a real data set comprising 72 paired samples and 20,531 genes through the proposed approach.

First, the number of genes is reduced to 10,815 genes with a significant differential expression among the two groups. The set of genes is then discretized into 5 labels, and the EPM algorithm is run by considering minimum support thresholds of 32% and 38% for cancer and healthy data sets, respectively. After that, the total number of frequent patterns obtained from

TABLE 11 Final set of patterns obtained from the cancer group of patients

ID	Patterns	Coverage (% of data)
1	FXYD4 = EL, BSND = EL, ATP6V1G3 = EL	51.39
2	PLA2G4F = EL, BSND = EL, ATP6V1G3 = EL	56.94
3	FXYD4 = EL, BSND = EL, LOC389493 = EL	59.72
4	FXYD4 = EL, RAB25 = EL, ATP6V1G3 = EL	68.06
5	SEPT1 = EH, CD2 = EH, CD3E = EH, SIT1 = EH, IL2RG = EH, KIAA0748 = EH	77.78
6	RALYL = EL, HEPACAM2 = EL, FOXI1 = EL, ATP6V1G3 = EL	80.56
7	CA10 = EL, AGR2 = EL, LOC389493 = EL	83.33
8	EPB41L5 = EL, MSI2 = EL, KIAA0564 = EL, ZNF844 = EL, C9orf80 = EL	86.11
9	ECSCR = EH, CDH5 = EH, PCDH12 = EH, CD34 = EH	93.06
10	LRBA = EL, TMEM184C = EL, C9orf80 = EL	95.83

Note: The labels next to the genes refer to gene expression values: extremely low (EL), low (L), medium (M), high (H) and extremely high (EH).

the cancerous group was 81,299, and 43,380 from the healthy group. Then, to improve the interpretation of experts, this set of patterns was reduced by the proposed filtering procedure, resulting in a set of 10 patterns (see Table 11) from the cancer group, and 5 from the healthy group (see Table 12). These representative patterns can represent 95.83% and 90.28% of the

TABLE 12 Final set of patterns obtained from the adjacent tissue group of samples

ID	Patterns	Coverage (% of data)
11	TMEM30B = EH, ERMP1 = EH, SRGAP3 = EH, COBLL1 = EH, MPP7 = EH, EHF = EH, KLHL14 = EH, ILDR1 = EH, TSPYL5 = EH, GRHL2 = EH, CDH1 = EH, CDKL2 = EH, ESRP1 = EH, C1orf116 = EH, TLR5 = EH	38.89
12	HSPA2 = EH, AIF1L = EH, C14orf4 = EH, MAL = EH, HOXB9 = EH, UXS1 = EH, HOXB5 = EH, SLC29A2 = EH, HOXB8 = EH, CNP = EH, GLTP = EH, MRPS6 = EH, RECQL5 = EH	41.67
13	SLC7A8 = EH, ENPP6 = EH, XPNPEP2 = EH, SLC22A8 = EH, ALDOB = EH, DAO = EH, SUSD2 = EH	79.17
14	ESRRB = EH, SIM2 = EH, SYT7 = EH, NAT8L = EH, FLJ42875 = EH, LOC25845 = EH	81.94
15	CLIC5 = EH, PTPRO = EH, WT1 = EH	90.28

Note: The labels next to the genes refer to gene expression values: extremely low (EL), low (L), medium (M), high (H) and extremely high (EH).

whole set of patients. Some of these patterns are described below after the application of a functional enrichment analysis.

Taking the aforementioned results, the functional enrichment analysis and literature review into account, it is possible to assert that several patterns are highly related to functions involved in cancer. For example, the patterns with ID 5 and 9 (see Table 11) refer to enriched biological processes related to the development of the immune response and angiogenesis. Specifically, in the pattern with ID 5, we only obtain a single gene which function is not directly linked to the development of T lymphocytes, that is, SEPT1.[57] SEPT1 is used in cytokinesis and the maintenance of cell morphology, and it promotes the migration of cancer cells when it is upregulated. Additionally, the deregulation of the immune system is known in several types of cancer.

Another interesting pattern, the one with ID 11 (see Table 12), is composed of genes involved in processes of regulation of the transport of proteins and organization of cellular unions. These genes are found with a higher level of expression in healthy samples indicating that they are downregulated in tumour samples. Further research studies[58] revealed that genes contained in the pattern such as EHF, C1orf116, ILDR1, and ESRP1 are related to migration and invasion processes in cancer. At the same time, ESRP1 may be intimately related to those genes that mediate cell binding functions. Some splice variants regulated by ESRPs genes are related to cytoskeleton reorganization and cell adhesion.[59]

5 | LESSON LEARNT

Emerging patterns is a powerful descriptive analysis technique that can be used in the study of cancer. Most existing cancer studies are based on binary comparisons (cancer samples vs. healthy samples) which is the aim of this technique. According to the study carried out, the discriminatory power of this technique enables to search for simple gene relationships that can be easily detected within the use of computers. More interesting is the feasibility to find high-order relationships that produce interactions not described yet. The proposed approach, which is based on EPM, has enabled alluring results to be obtained when it was applied to several types of cancer. Some discovered genes have been described individually in cancer, while others have a known interaction or participate in the same biological processes. Other genes, however, have not been associated with cancer yet, no study demonstrates their interaction with other genes of the discovered pattern, or they are poorly studied.

A really interesting insight was discovered with pattern {LDB2 = EL, CXorf36 = EL, MMRN2 = EL, ECSCR = EL, MYCT1 = EL, CLEC14A = EL}, belonging to the breast cancer analysis. Most of the genes involved with a low expression level in this pattern are well-known for their relationship with cell death avoidance processes. The gene LDB2, however, does not have a recorded interaction with the rest of the genes in the pattern. Something similar happens with gene CXorf36 (also called DIPK2B), which function remains poorly studied. It is, therefore, necessary to perform further research studies on these genes, which might be of vital interest to the function and possible interaction of these genes as supposed tumour suppressors.

Another alluring solution is determined by the pattern PSMB4 = EL, KRTCAP2 = EL, SNRPG = EL, RAG1AP1 = EL, HSPE1 = EL, SSBP1 = EL, CHCHD8 = EL, TIMM10 = EL, TMED9 = EL, discovered in the lung adenocarcinoma analysis. Some of these genes such as the components of the proteasome have been reported to mediate functions that promote the development of cancer. In contrast, the role of genes such as KRTCAP2, RAG1AP1 and TMED9

remains unexplored in lung cancer. Therefore, it is of high interest to perform a further investigation on these genes to reveal the existence of some nexus of interaction between them that would allow the development of new therapies in cancer.

Another interesting pattern that may require further study is LOC653501 = EH, ZNF658 = EH, ZNF214 = EH, ZNF230 = EH, KIAA1383 = EH, obtained in thyroid cancer study. This pattern groups together several genes corresponding to zinc finger proteins that are known to act as transcriptional activators. These proteins were reported in cancer studies, but the underlying mechanisms by which it influences cancer need further studies. Depending on their binding to different partners, zinc finger proteins may mediate the opposite function. In this case, those genes that encode these proteins are upregulated in thyroid cancer, and they are also related to MAP10. The functions of MAP10 include cell division and positive regulation of cytokinesis, both of which are involved in cancer progression. The obtained results suggested a possible interaction nexus between zinc finger protein genes and MAP10 that could be intimately linked to the development and progression of thyroid cancer. However, the mechanisms by which this interaction occurs should be reviewed in future research.

These are just some examples of insights extracted by the proposed workflow. Really interesting set of genes were described in the case studies. Not only unknown genes were obtained from these case studies but also well-known interactions among genes that demonstrate that the proposed methodology works well. For example, some of the genes belonging to the set {LIPE, GYG2, AQP7, GPD1, CIDEC, PNPLA2} are known by their implication in the deregulation of lipid metabolism in cancer. Deregulation in lipid metabolism is commonly associated with alternative means of obtaining energy from tumour cells, such as glucose metabolism.

6 | CONCLUSIONS

In this study, a new approach for cancer analysis based on the use of emerging patterns is proposed. The proposal includes four different procedures that are specifically designed to deal with RNA-Seq data on cancer. The approach enables the extraction of useful and highly interpretable knowledge from data belonging to several types of cancer without requiring any kind of starting hypothesis or previous knowledge. Also, the proposed approach has the advantage of using high order relationships without being restricted to pair-wise correlations or restricted by assuming homoscedasticity, allowing a more powerful descriptive analysis without the introduction of biases commonly assumed by other techniques. The proposed approach has been evaluated on different real scenarios related to varied cancer types. Six different case studies based on RNA-Seq data taken from The Cancer Genome Atlas (TCGA) were analysed. The obtained results have demonstrated the usefulness of the proposal. Some of the extracted insights were already described in the specialized literature as good cancer bio-markers, while other results have not been described yet due to some existing techniques are biased by biological databases. More specifically, among the results, we found some genes relate to the deregulation of the cell cycle in breast cancer, a phenomenon described in the literature and that certifies the adequacy of this approach. On the other hand, not widely studied genes have been discovered to be involved in the disease. Associations of genes such as KRTCAP2, RAG1AP1 and TMED9 have been detected along with other genes related to the development of cancer through the proteasome in lung adenocarcinoma while CXorf36 and LDB2 could be new biomarkers related to cell death avoidance functions. For all the above reasons, EPM, and more specifically, the proposed approach should be considered in genomic data studies, not

only by its potential for the detection of new biomarkers but also because it can be really useful in the data exploratory analysis.

Finally, it is mandatory to highlight that future research works may extend the use of EPM to multiple diseases such as liver, kidney, heart, mental, or any medical application where two groups need to be compared. This methodology is also feasible to apply to the analysis of additional diseases such as tropical diseases (malaria) since it compares healthy with unhealthy groups. The present research work can also be extended to consider not only two groups to be compared but multiple ones through a methodology named contrast set mining.

ACKNOWLEDGEMENTS

This study was supported by the Spanish Ministry of Science and Innovation, project TIN2017-83445-P, and the University of Cordoba, project UCO-FEDER 18 REF.1263116 MOD.A. Both projects were also supported by the European Fund of Regional Development.

ORCID

Antonio Manuel Trasierras https://orcid.org/0000-0001-7564-4159
José María Luna https://orcid.org/0000-0003-3537-2931
Sebastián Ventura https://orcid.org/0000-0003-4216-6378

REFERENCES

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*. 2018; 68(6):394-424.
2. Bruggeman FJ, Westerhoff HV. The nature of systems biology. *Trends Microbiol*. 2007;15(1):45-50.
3. Azim FS, Houri H, Ghalavand Z, Nikmanesh B. generation sequencing in clinical oncology: applications, challenges and promises: a review article. *Iran J Public Health*. 2018;47(10):1453.
4. Ritchie ME, Phipson B, Wu D, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res*. 2015;43(7):e47.
5. Love MI, Anders S, Kim V, Huber W. RNA-Seq workflow: gene-level exploratory analysis and differential expression. *F1000Research*. 2015;4:1070.
6. Hsu YL, Huang PY, Chen DT. Sparse principal component analysis in cancer research. *Transl Cancer Res*. 2014;3(3):182.
7. Stuart JM, Segal E, Koller D, Kim SK. A gene-coexpression network for global discovery of conserved genetic modules. *Science*. 2003;302(5643):249-255.
8. Perscheid C, Uflacker M. Integrating biological context into the analysis of gene expression data. In: Rodríguez S, Prieto J, Faria P, et al., eds. *Distributed Computing and Artificial Intelligence, Special Sessions, 15th International Conference*. Cham: Springer International Publishing; 2019:339-343.
9. Mathur R, Rotroff D, Ma J, Shojaie A, Motsinger-Reif A. Gene set analysis methods: a systematic comparison. *BioData Min*. 2018;11(1):8.
10. Ashburner M, Ball CA, Blake JA, et al. Gene ontology: tool for the unification of biology. *Nat Genet*. 2000; 25(1):25.
11. Kanehisa M, Goto S, Sato Y, Kawashima M, Furumichi M, Tanabe M. Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res*. 2013;42(D1):D199-D205.
12. Gaudelet T, Malod-Dognin N, Pržulj N. Higher-order molecular organization as a source of biological function. *Bioinformatics*. 2018;34(17):i944-i953.
13. Shastry KA, Sanjay H. *Machine Learning for Bioinformatics*. Singapore: Springer; 2020:25-39.
14. Pérez E, Reyes O, Ventura S. Convolutional neural networks for the automatic diagnosis of melanoma: an extensive experimental study. *Med Image Anal*. 2020;67(22):101858.
15. Chaurasia V, Pal S. A novel approach for breast cancer detection using data mining techniques. *Int J Innovat Res Comput Commun Eng*. 2017;2.

16. Williams K, Idowu PA, Balogun JA, Oluwaranti AI. Breast cancer risk prediction using data mining classification techniques. *Trans Netw Commun*. 2015;3(2):01.
17. Ventura S, Luna JM. *Supervised Descriptive Pattern Mining*. Switzerland: Springer; 2018.
18. Tomczak K, Czerwińska P, Wiznerowicz M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemp Oncol*. 2015;19(1A):A68.
19. Luna JM, Fournier-Viger P, Ventura S. Frequent itemset mining: a 25 years review. *Wiley Interdiscip Rev Data Min Knowl Discov*. 2019;9(6):e1329.
20. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci*. 2001;98(9):5116-5121.
21. Zhao S, Fung-Leung WP, Bittner A, Ngo K, Liu X. Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells. *PlOS One*. 2014;9(1):e78644.
22. Kalaiarasai A, Amanulla KM. Unconscious oral cancer detection using data mining classification approaches. *Int J Adv Res Comput Eng Technol*. 2015;4(7):3177-3184.
23. Padillo F, Luna JM, Ventura S. LAC: Library for associative classification. *Knowl Based Syst*. 2020;193:105432. https://doi.org/10.1016/j.knosys.2019.105432
24. Alagukumar S, Lawrance R. *Classification of Microarray Gene Expression Data using Associative Classification*. International Conference on Computing Technologies and Intelligent Data Engineering (ICCTIDE). Kovilpatti, India: IEEE; 2016:1-8.
25. Vengateshkumar R, Alagukumar S, Lawrance R. Analysis of microarray gene expression data using Boolean association rule mining. *Int J Innov Technol Creat Eng*. 2017;7(5):412-416.
26. Boulesteix AL, Tutz G, Strimmer K. A CART-based approach to discover emerging patterns in microarray data. *Bioinformatics*. 2003;19(18):2465-2472.
27. Li J, Wong L. Emerging patterns and gene expression data. *Genome Inform*. 2001;12:3-13.
28. Li J, Wong L. Identifying good diagnostic gene groups from gene expression profiles using the concept of emerging patterns. *Bioinformatics*. 2002;18(5):725-734.
29. Vengateshkumar R, Alagukumar S, Lawrance R. *Boolean Association Rule Mining on Microarray Gene Expression Data*. Singapore: Springer; 2020:99-111.
30. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*. 2009;26(1):139-140.
31. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Ser B Methodol*. 1995;57(1):289-300.
32. Dong G, Li J, Wong L. The use of emerging patterns in the analysis of gene expression profiles for the diagnosis and understanding of diseases. In: *New Generation of Data Mining Applications*. New York: IEEE Press/Wiley; 2004:331-354.
33. Kang J, Sergio CM, Sutherland RL, Musgrove EA. Targeting cyclin-dependent kinase 1 (CDK1) but not CDK4/6 or CDK2 is selectively lethal to MYC-dependent human breast cancer cells. *BMC Cancer*. 2014;14(1):32.
34. Yu B, Ding Y, Liao X, Wang C, Wang B, Chen X. Overexpression of PARPBP correlates with tumor progression and poor prognosis in hepatocellular carcinoma. *Dig Dis Sci*. 2019;64(10):2878-2892.
35. Winslow S, Lindquist KE, Edsjö A, Larsson C. The expression pattern of matrix-producing tumor stroma is of prognostic importance in breast cancer. *BMC Cancer*. 2016;16(1):841.
36. Mura M, Swain R, Zhuang X, et al. Identification and angiogenic role of the novel tumor endothelial marker CLEC14A. *Oncogene*. 2012;31(3):293.
37. Bi P, Yue F, Karki A, et al. Notch activation drives adipocyte dedifferentiation and tumorigenic transformation in mice. *J Exp Med*. 2016;213(10):2019-2037.
38. Zoppino FCM, Guerrero-Gimenez ME, Castro GN, Ciocca DR. Comprehensive transcriptomic analysis of heat shock proteins in the molecular subtypes of human breast cancer. *BMC Cancer*. 2018;18(1):700.
39. Jögi A, Vaapil M, Johansson M, PÅhlman S. Cancer cell differentiation heterogeneity and aggressive behavior in solid tumors. *Upsala J Med Sci*. 2012;117(2):217-224.
40. Pan Y, Liu G, Yuan Y, Zhao J, Yang Y, Li Y. Analysis of differential gene expression profile identifies novel biomarkers for breast cancer. *Oncotarget*. 2017;8(70):114613.
41. Katoh M. Molecular cloning and characterization of MFRP, a novel gene encoding a membrane-type Frizzled-related protein. *Biochem Biophys Res Commun*. 2001;282(1):116-123.

42. ElMarabti E, Younis I. The Cancer Spliceome: reprograming of alternative splicing in cancer. *Front Mol Biosci*. 2018;5:80.
43. Porta C, Paglino C, Mosca A. Targeting PI3K/Akt/mTOR signaling in cancer. *Front Oncol*. 2014;4:64.
44. Li LJ, Zhang LS, Han ZJ, He ZY, Chen H, Li YM. Chaperonin containing TCP-1 subunit 3 is critical for gastric cancer growth. *Oncotarget*. 2017;8(67):111470.
45. Schürch CM, Roelli MA, Forster S, et al. Targeting CD47 in anaplastic thyroid carcinoma enhances tumor phagocytosis by macrophages and is a promising therapeutic strategy. *Thyroid*. 2019;29(7):979-992.
46. Ho SR, Lin WC. RNF144A sustains EGFR signaling to promote EGF-dependent cell proliferation. *J Biol Chem*. 2018;293(42):16307-16323.
47. Fong K-w, Leung JW-c, Li Y. MTR120/KIAA1383, a novel microtubule-associated protein, promotes microtubule stability and ensures cytokinesis. *J Cell Sci*. 2013;126(3):825-837.
48. Dynlacht BD. Regulation of transcription by proteins that control the cell cycle. *Nature*. 1997;389(6647):149.
49. Signoretti S, Waltregny D, Dilks J, et al. p63 is a prostate basal cell marker and is required for prostate development. *Am J Pathol*. 2000;157(6):1769-1775.
50. Jordan NV, Prat A, Abell AN, et al. SWI/SNF chromatin-remodeling factor Smarcd3/Baf60c controls epithelial-mesenchymal transition by inducing Wnt5a signaling. *Mol Cell Biol*. 2013;33(15):3011-3025.
51. Martignano F, Gurioli G, Salvi S, et al. GSTP1 methylation and protein expression in prostate cancer: diagnostic implications. *Dis Markers*. 2016;2016:4358292.
52. Li X, Xu W, Kang W, et al. Genomic analysis of liver cancer unveils novel driver genes and distinct prognostic features. *Theranostics*. 2018;8(6):1740.
53. Culig Z, Santer FR. Androgen receptor signaling in prostate cancer. *Cancer Metastasis Rev*. 2014;33(2-3):413-427.
54. Christensen JF, Jones L, Andersen J, Daugaard G, Rorth M, Hojman P. Muscle dysfunction in cancer patients. *Ann Oncol*. 2014;25(5):947-958.
55. Yanagisawa Y, Takeoka M, Ehara T, Itano N, Miyagawa S, Taniguchi S. Reduction of Calponin h1 expression in human colon cancer blood vessels. *Eur J Surg Oncol*. 2008;34(5):531-537.
56. Szklarczyk D, Gable AL, Lyon D, et al. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res*. 2018;47(D1):D607-D613.
57. Mizutani Y, Ito H, Iwamoto I, et al. Possible role of a septin, SEPT1, in spreading in squamous cell carcinoma DJM-1 cells. *Biol Chem*. 2013;394(2):281-290.
58. Parsana P, Amend SR, Hernandez J, Pienta KJ, Battle A. Identifying global expression patterns and key regulators in epithelial to mesenchymal transition through multi-study integration. *BMC cancer* 2017;17(1):1-14.
59. Warzecha CC, Shen S, Xing Y, Carstens RP. The epithelial splicing factors ESRP1 and ESRP2 positively and negatively regulate diverse types of alternative splicing events. *RNA Biol*. 2009;6(5):546-562.

How to cite this article: Trasierras AM, Luna JM, Ventura S. Improving the understanding of cancer in a descriptive way: an emerging pattern mining-based approach. *Int J Intell Syst*. 2021;1-27. https://doi.org/10.1002/int.22503