Pattern-centric transformation of omics-data sources grounded on multi-wise gene associations aids predictive tasks in TCGA while ensuring interpretability

André Patrício,1,2,∗ Rafael S. Costa2 and Rui Henriques1

1INESC-ID, Instituto Superior Técnico, Universidade de Lisboa, R. Alves Redol 9, 1000-029, Lisboa, Portugal and 2LAQV-REQUIMTE, Department of Chemistry, NOVA School of Science and Technology, NOVA University Lisbon, Campus da Caparica, 2829-516, Caparica, Portugal

∗Corresponding author: andremppatricio@tecnico.ulisboa.pt

Abstract

Motivation: The increasing prevalence of omics data sources is pushing the study of regulatory mechanisms underlying complex diseases such as cancer. However, the vast quantities of features produced and the inherent interplay between them lead to a level of complexity that hampers both descriptive and predictive tasks, requiring custom-built algorithms that can extract relevant information from these sources of data.

Results: We propose a transformation that moves data centered on molecules (e.g. transcripts and proteins) to a new data space focused on putative regulatory modules given by statistically relevant patterns of co-expression. The proposed transformation extracts patterns from the data through biclustering and uses them to create new variables with guarantees of interpretability and discriminative power. The transformation is shown to achieve dimensionality reductions of up to 99% and to increase the predictive performance of various classifiers across multiple omics layers. Our results suggest that a transformation of omics data from gene-centric to pattern-centric data provides benefits to both prediction tasks and human interpretation. The proposed approach is expected to greatly support further bioinformatic analyses for precision medicine applications.

Availability: Software code and the raw results generated are available at github.com/Andrempp/Pattern-Centric-Transformation.

Contact: andremppatricio@tecnico.ulisboa.pt

Supplementary information: Supplementary data are available at Journal Name online.

Key words: omics-data, TCGA, dimensionality reduction, interpretability, systems biology, cancer, data integration, predictive models

1. Introduction

Cancer is a class of diseases characterized by the abnormal growth, replication, and survivability of cells, which in 2020 presented 19,292,800 new cases and caused 9,958,100 deaths worldwide (Ferlay et al., 2021). This convoluted disease can have its origin in various tissues and arise due to disruptions in multiple cellular processes. New profiling techniques enable an in-depth look into the biological mechanisms related to complex diseases. The information generated by these techniques, termed omics data, can be categorized according to the targeted molecular structures, whether genes (genomics), transcripts (transcriptomics), or proteins (proteomics), amongst others. Data from these omics layers have been widely used in cancer research for multiple purposes, such as molecular subtyping (Zhao et al., 2019; Sun et al., 2022; Doebbley et al., 2022), biomarker identification (Swan et al., 2013; Zhao et al., 2020; Ghosal et al., 2020), prognosis (Marquardt et al., 2012; Nault et al., 2013; Roessler et al., 2010), and drug development (Neary et al., 2021; Huang et al., 2019; Li et al., 2021). Still, the large quantities of data produced, hundreds or thousands of measurements per profiled sample, hinder the ability of humans and even algorithms to extract knowledge from it (Clarke et al., 2008).

Classic approaches to this problem focus on selecting Differentially Expressed Genes (DEGs) between the conditions under study (Zhang et al., 2015, 2018; Zhao et al., 2014). Although useful to reduce the high-dimensionality of omics data, these techniques are generally unable to capture dependencies among multiple genes. Since genes often work with each other in regulatory modules, this type of analysis can exclude genes that alone have a subtle differential profile, but when considered with other genes may have a decisive influence on one or more conditions, a problem shared by many other dimensionality reduction techniques. Procedures for reducing dimensionality focused on multivariate relationships (Ding and Peng, 2005; Gevaert et al., 2006; Yeung and Bumgarner, 2003) can be used to deal with this problem, but these are generally associated with the
loss of potentially relevant information and interpretability (Saëys et al., 2007). Complementary advances, including integration of domain knowledge (Fan et al., 2019; Rabgu et al., 2017; Yousef et al., 2021b), integration of multiple omics (Rokhart et al., 2017; Metselaar et al., 2021), and ensemble of feature selection techniques (Lu et al., 2017; Metselaar et al., 2021), have been proposed, with some showing capability of improving predictive performance (Chandra and Gupta, 2011; Ding and Peng, 2005; Acharya et al., 2017; Yousef et al., 2021a).

This work provides a data transformation that can be applied to different types of omics data to increase the prediction capability of machine learning models while solving the previously mentioned challenges: dimensionality reduction; presence of putative regulatory modules with synergistic gene interplay on specific conditions; and preserved interpretability of the transformed data space.

We hypothesize that the defined goal can be achieved through pattern mining techniques, more precisely, biclustering. Biclustering algorithms search the data for subsets of observations following a coherent pattern across a subset of features, which in gene expression data corresponds to a group of co-expressed genes with a statistically significant regulatory pattern. The discriminative power of this pattern can then be calculated through various metrics. In accordance, biclustering has been widely used for putative regulatory module discovery (Madeira and Oliveira, 2004), with recent contributions improving the classic stances to further ensure functional relevance (Henriques and Madeira, 2014), statistical significance (Henriques and Madeira, 2018), and discriminative power (Henriques and Madeira, 2021).

The data transformation proposed in this work uses biclustering to extract relevant patterns from the data and subsequently use them to create new variables encoding the activation of said putative regulatory pattern. Contrary to other feature extraction methods, information loss is minimal as each pattern encompasses relevant information on a subset of the original molecular variables. Furthermore, interpretability is preserved in the reduced data space (arguably augmented given its organization in putative functions) and well suited for subsequent predictive ends as each pattern shows notable guarantees of discriminative power.

To validate the aforementioned qualities of the proposed pattern-centric data transformation, we study its impact on dimensionality reduction and classification performance across different predictive models. This analysis is performed with data from The Cancer Genome Atlas (TCGA) program (Weinstein et al., 2013), encompassing multiple omics layers from the Brain Lower Grade Glioma (TCGA-LGG) and Colon Adenocarcinoma (TCGA-COAD) projects.

2. Materials and Methods

The proposed data transformation converts a feature space based on molecules (genes, transcripts, proteins, etc.) to a space based on patterns of expression of these molecules. The transformation identifies relevant combinations of molecules and corresponding expression values and groups them into a single feature, resulting in new features that encapsulate information about patterns that are relevant to the problem at hand. The first step of the transformation is the mining of putative regulatory patterns in the data through pattern mining based biclustering (Henriques et al., 2015), followed by the creation of new features for each found pattern, where the corresponding value represents the presence of that regulatory pattern in the sample.

2.1. Datasets and Preprocessing

The data used for assessing the impact of the proposed transformation originates from The Cancer Genome Atlas (TCGA) program. Two distinct TCGA projects were selected based on heterogeneity and coverage of transcriptome- and proteome-wide data: Brain Lower Grade Glioma (TCGA-LGG)1 and Colon Adenocarcinoma (TCGA-COAD)2. For each project, we explore regulatory activity pertaining to miRNA, mRNA, and protein expression, and further access patients’ clinical variables. Vital Status is selected as the primary target in this work, encoding the survivability of the patient. The data was obtained through the R package TCGABioblinks (Colaprico et al., 2016) and corresponds to the Genomic Data Commons (GDC) harmonized database, which uses the human genome version GRCh38 as reference. The number of samples and variables per project and omics layer are shown in Table 1.

From the available measurements in each layer, we select Reads Per Million Mapped Reads (RPD) for miRNA, Transcripts Per Kilobase Million (TPM) for mRNA, and relative levels of expression for proteins. Features with a high percentage of 0’s (≥95%) were removed, log transformation was applied to the miRNA and mRNA layers, and the filtering function from edgeR (Robinson et al., 2010; Chen et al., 2016) was used to perform preliminary filtering of genes not significantly expressed.

2.2. Biclustering

Given a multivariate dataset described by a set of observations \( X = \{x_1, ..., x_n\} \) and a set of variables \( Y = \{y_1, ..., y_m\} \), biclustering is a data mining technique that identifies subspaces, each described by a subset of observations and variables \( B = \{I \subset X, J \subset Y\} \) with pattern \( p = \{y_j = c_j | y_j \in J\} \), satisfying specific homogeneity and statistical significance criteria (e.g., coherently expressed genes on an unexpectedly high number of samples).

The homogeneity criteria of the patterns can be adjusted according to the addressed problem. Patterns can be described by constant values across observations or variables, shifting and scaling factors across rows or columns, and coherent order of variables. In this work, we use patterns of constant values across observations. Biclustering, therefore, has the advantage over clustering of only comparing observations according to a subset of variables. This characteristic is fundamental when considering gene expression data since genes work in subsets to form regulatory modules that control biological mechanisms.

Biclustering approaches based on pattern mining further show unique benefits, including efficient exhaustive searches, robustness to noise and missing values, flexible bicluster structure, and the possibility to customize the desirable pattern coherence, amongst others (Henriques et al., 2015). These characteristics make BicPAMS (Henriques et al., 2017), an algorithm that combines state-of-the-art contributions from pattern mining-based biclustering for biological data analysis, a viable candidate for the defined goal: the identification of relevant patterns of expression in omics data to transform gene-centric data spaces in pattern-centric data spaces.

2.3. Pattern-centric transformation

The proposed transformation works by extracting putative regulatory patterns from data and using them to generate new dimensions, moving from a gene-centric to a pattern-centric space. We start by applying the BicPAMS algorithm (Henriques et al., 2017) to identify patterns that yield statistical significance and are additionally discriminative of a selected target variable. Each pattern is defined by a set of observations, a set of variables and corresponding values, and descriptive metrics such as p-value and lift. The columns of the new dimension space will correspond to the found patterns. For each select pattern, the process followed is described in Figure 1: i) select a pattern and identify the corresponding values (Pattern 1 and values 3, 4, and 8); ii) create a column in the new dataset corresponding to this pattern (column $p_1$); and iii) for each row, calculate the distance between its values and the values of the pattern (Euclidean distance between 6, 8, 5 and 3, 4, 8) and use it as the new value;

$$\text{lift} = \frac{P(p,c)}{P(p)P(c)}$$  

where $P(p)$ is the probability of pattern $p$ occurring in the data, $P(c)$ is the probability of class $c$ occurring in the data, and $P(p,c)$ is the joint probability of $p$ and $c$ in the data.

In order to evaluate the expression level of a regulatory pattern in a sample, we further need to select a distance metric. According to sensitivity analysis over various similarity and correlation measures, the Euclidean distance is selected for this end. A final possible step is to perform a feature selection post-transformation, reducing the number of new features according to some criteria. However, in order to preserve the analysis of the proposed transformation, postprocessing options were excluded.

The motivations for the proposed transformation are varied: i) the transformed data doesn’t lose interpretability since the relation between the new features and the corresponding gene expression values is known; ii) the data stays in tabular form, allowing the use of classical ML models; iii) the new features encode relevant relations between the genes, simplifying the prediction task for classifiers; and iv) the identified putative regulatory modules can be used in a descriptive analysis of the data to strengthen the understanding of the condition under study.

2.4. Hyperparemterization

The proposed transformation has multiple hyperparameters that can be adjusted to best suit the data in question. These can be divided into two groups, the parameters of the biclustering algorithm and the parameters specific to the subsequent transformation.

The biclustering parameters subjected to optimization are: number of iterations, the number of repetitions of the mining process, with each new iteration masking the previously found biclusters to force the discovery of less trivial ones; minimum lift, a threshold that every bicluster must reach to be selected; coherence strength, the number of intervals used for the discretization of the real-valued data; and minimum number of biclusters, the number of biclusters per iteration at which the search stops.

The remaining hyperparameters are used directly in the transformation, comprising the filtering of the found patterns and the distance calculus, as previously explained.

Regarding biclustering, its parametrization was fixed according to sensitivity analysis over multiple cancers and omics layers. In a biological system, the underlying molecular mechanisms are complex, a result of the many interactions between a lot of molecules, be it transcripts, proteins, or others. Consequently, it is desirable to capture patterns across the whole dataset, being careful not to extract patterns focused only on a part of the data. Still, these patterns should be discriminative of the target variable, which requires a balance between the breadth of the search and the identification of patterns that are indeed relevant.

Finally, we also aim for a significant compression of the existing information, in the form of dimensionality reduction. To achieve all this, the bicluster parameters are fixed as follows: a relatively high number of iterations (10) to force a broad search, a minimum lift of 1.2 to ensure the discriminative power of the patterns, a low minimum number of biclusters (10) to help reduce dimensionality.
and maintain the search feasible, and a heightened coherence strength (11) to ensure discrimination of the different values.

2.5. Implementation details

The proposed transformation is developed in Python 3.10, and the ML models used to assess classification performance are implemented in the scikit-learn package (Pedregosa et al.). The biclustering algorithm, BicPAMS (Henriques et al., 2017), is implemented in Java and the Differential Gene Expression (DEG) analysis is performed through the R package edgeR (Robinson et al., 2010). The code used to obtain the presented results is available in GitHub (github.com/Andrempyp/Pattern-Centric-Transformation) and was run on a machine with an Intel Core i7-11700K 8-Core 3.6GHz and 32 GB of RAM.

3. Results

3.1. Evaluation and Metrics

To evaluate the impact of the proposed transformation on classification performance we train and test three ML classifiers: a Logistic Regression (LR), a Support Vector Machine (SVM), and a Random Forest (RF). This set of classifiers enables a broader understanding of the impacts on performance across different types of models, namely, linear (LR), kernel-based (SVM), and ensemble (RF). The evaluation is performed with a cross-validation setting with 10 folds, assuring that the patterns used for the transformation are discovered only on the training data.

The classifiers are evaluated through the following metrics: accuracy, to provide an overall indicator of performance; and recall, precision, and F1-score, to obtain a balanced view of how well the models identify the positive class (Vital Status).

To provide a reference of comparison for the results obtained in the transformed data, we also display the results obtained in the non-transformed dataset that maintains the transcripts/proteins as features. These results are presented in two settings. First, the results obtained in the miRNA and protein layers, where the gene-centric dataset goes through the preprocessing steps previously explained. Second, the results obtained in the mRNA layer, where due to its high dimensionality the gene-centric dataset is further reduced through Differential Expression Gene (DEG) analysis with a p-value of 0.05 adjusted with Benjamini-Hochberg procedure multiple testing correction (Benjamini and Hochberg, 1995).

3.2. Prediction Results

The prediction results for Low Grade Glioma (LGG) can be seen in Figure 2 for miRNA and Figure 3 for protein.

In the miRNA layer, averaging over the 10 folds of cross-validation, there are 351 features in the gene-centric dataset and 26 features in the pattern-centric dataset. This corresponds to a reduction in the dimensionality of 92.6% while improving the predictive performance across all metrics and classifiers except one (accuracy in Random Forest). In the protein layer, there is a reduction of 457 features to 29, a 93.7% dimensionality reduction while improving the predictive performance of SVMs and Random Forests across all metrics.

Advancing to the mRNA layer, the dataset has a total of 60664 features before processing, and 24027 features postprocessing. In order to obtain a better comparison of results, the gene-centric data is further reduced through feature selection with DEG analysis, selecting the features that have a p-value of 0.05 or lower. The pattern-centric data is obtained by applying the transformation to the original preprocessed data, as in the previous results. Figure 4 shows the obtained results.

In the original dataset, the average dimensionality across folds after DEG analysis is 12140. The pattern-centric transformation reduces it to an average of 30 variables, corresponding to an increased reduction in the dimensionality of 99%. This is accompanied by an increase in performance in all metrics in the SVM and Random Forest classifiers. The Logistic Regression shows decreases in recall and F1-score.

In the TCGA-COAD project, increases in performance are observed, although less prominent. Furthermore, the reductions in dimensionality are above 80%. The complete prediction results can be consulted in Tables 2 and 3, and the corresponding dimensionality reductions are available in Table 4.

3.3. Dimensionality Reduction

To better understand the capabilities of the proposed transformation for reducing dimensionality while preserving relevant information, we now consider the mRNA layer of the LGG project. Using the same evaluation setting as before, we compare the differences in performance in relation to a continuous decrease in the number of variables. We present the study of three techniques: DEG analysis, Principal Component Analysis (PCA), and the proposed pattern-centric transformation. For the DEG analysis, the variables are ordered according to their Benjamini-Hochberg adjusted p-value and the top n variables selected. This dimensionality reduction maintains the data in the same type of space, with the mRNA transcripts as the variables. The PCA transformation identifies the n top principal components, extracting information about the variability of the data and using it to transform it. This results in an alteration of the variable space to one that hinders human interpretation. Regarding the pattern-centric transformation, it is applied as previously explained, but it further filters the found patterns according to their lift, enabling a further reduction of dimensionality. This process transforms the variable space into a different one, where each variable represents the interactions between the original variables, whilst maintaining the possibility of human interpretation. The obtained results can be seen in Figure 5.

The results show that although the pattern-centric transformation doesn’t start as the best-performing technique, as the dimensionality is reduced, its performance steadily increases, making it the best performer at the lowest dimensionalities.

3.4. Feature Space

In addition to the impact on performance, it is also useful to analyze how the proposed transformation affects other characteristics of the feature space, namely, class separability.

Both the gene- and pattern-centric data are reduced to two dimensions through the UMAP (McInnes et al., 2020) technique and their data points are plotted using color to differentiate the two classes (Alive and Deceased), as seen in Figure 6. Both plots indicate an inherently difficult classification task since there is no clear separation between the two classes. Still, it is visible that the transformation causes a significant alteration in the way data is distributed.

Complementary to the study of the data distribution as a whole, the distribution of individual variables is also analyzed.
Fig. 2: Comparison of dimensionality (left plot) and classification performance between gene-centric and pattern-centric data in the miRNA layer of TCGA-LGG project.

Fig. 3: Comparison of dimensionality (left plot) and classification performance between gene-centric and pattern-centric data in the protein layer of TCGA-LGG project.

Fig. 4: Comparison of dimensionality (left plot) and classification performance between gene-centric and pattern-centric data in the mRNA layer.

We compare the distributions conditioned on the target variable of the most discriminative features, both in the gene- and pattern-centric datasets. The five top discriminative features for each dataset are selected through their Mutual Information (MI). Figure 7 shows a Kernel Density Estimate (KDE) plot for the conditioned distributions of each of these variables, with the gene-centric features on top and the pattern-centric features on the bottom. The overlap of the conditioned distributions in the gene-centric dataset indicates, once again, an innate difficulty in this prediction task. However, the pattern-centric features show a clear improvement in the separability of the two distributions, demonstrating that the new feature space yields higher guarantees of discriminative power.

The same analysis for the data from the project COAD can be found in Supplementary Figures S1 and S2.

4. Discussion

Omics data contains fundamental information about regulatory biological mechanisms, however, capturing the inherent complexity of molecular interplay in the various pathways is a challenging task. This complicates the use of omics data in clinical settings since it is important to maintain some level of interpretability of both the data and the obtained results. We have proposed here a transformation based on the mining of putative regulatory patterns with discriminative guarantees to create a new feature space with unique properties of interest.
Table 2. Comparison of classification results for different omics layers and classifiers between gene-centric (GC) and pattern-centric (PC) data in the TCGA-LGG project (best results in bold and ± are the standard deviations).

<table>
<thead>
<tr>
<th>Layer</th>
<th>Classifier</th>
<th>GC</th>
<th>PC</th>
<th>GC</th>
<th>PC</th>
<th>GC</th>
<th>PC</th>
<th>F1-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>miRNA</td>
<td>RF</td>
<td>0.782 ± 0.05</td>
<td>0.775 ± 0.044</td>
<td>0.601 ± 0.062</td>
<td>0.602 ± 0.061</td>
<td>0.723 ± 0.181</td>
<td>0.723 ± 0.134</td>
<td>0.608 ± 0.090</td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>0.769 ± 0.039</td>
<td>0.778 ± 0.038</td>
<td>0.566 ± 0.046</td>
<td>0.586 ± 0.057</td>
<td>0.701 ± 0.180</td>
<td>0.717 ± 0.166</td>
<td>0.559 ± 0.069</td>
</tr>
<tr>
<td>Protein</td>
<td>RF</td>
<td>0.745 ± 0.044</td>
<td>0.755 ± 0.022</td>
<td>0.633 ± 0.074</td>
<td>0.652 ± 0.040</td>
<td>0.636 ± 0.063</td>
<td>0.619 ± 0.127</td>
<td>0.632 ± 0.069</td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>0.778 ± 0.031</td>
<td>0.799 ± 0.030</td>
<td>0.556 ± 0.063</td>
<td>0.612 ± 0.054</td>
<td>0.641 ± 0.215</td>
<td>0.723 ± 0.154</td>
<td>0.539 ± 0.097</td>
</tr>
</tbody>
</table>

Table 3. Comparison of classification results for different omics layers and classifiers between gene-centric (GC) and pattern-centric (PC) data in the TCGA-COAD project (best results in bold and ± are the standard deviations).

<table>
<thead>
<tr>
<th>Layer</th>
<th>Classifier</th>
<th>GC</th>
<th>PC</th>
<th>GC</th>
<th>PC</th>
<th>GC</th>
<th>PC</th>
<th>F1-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>miRNA</td>
<td>RF</td>
<td>0.776 ± 0.049</td>
<td>0.780 ± 0.035</td>
<td>0.758 ± 0.061</td>
<td>0.627 ± 0.054</td>
<td>0.720 ± 0.053</td>
<td>0.719 ± 0.095</td>
<td>0.727 ± 0.053</td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>0.772 ± 0.031</td>
<td>0.796 ± 0.032</td>
<td>0.609 ± 0.052</td>
<td>0.627 ± 0.066</td>
<td>0.685 ± 0.091</td>
<td>0.769 ± 0.094</td>
<td>0.620 ± 0.064</td>
</tr>
<tr>
<td>mRNA</td>
<td>RF</td>
<td>0.782 ± 0.028</td>
<td>0.790 ± 0.025</td>
<td>0.612 ± 0.030</td>
<td>0.619 ± 0.040</td>
<td>0.742 ± 0.099</td>
<td>0.760 ± 0.091</td>
<td>0.625 ± 0.039</td>
</tr>
</tbody>
</table>

Table 4. Dimensionalities before transformation (BT) and after transformation (AT) across different layers and projects.

<table>
<thead>
<tr>
<th>Project</th>
<th>miRNA</th>
<th>Protein</th>
<th>mRNA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Accuracy</td>
<td>Recall</td>
<td>Reduction</td>
</tr>
<tr>
<td>TCGA-LGG</td>
<td>BT</td>
<td>AT</td>
<td>Reduction</td>
</tr>
<tr>
<td></td>
<td>351</td>
<td>26</td>
<td>84.2%</td>
</tr>
<tr>
<td>TCGA-COAD</td>
<td>292</td>
<td>46</td>
<td>84.2%</td>
</tr>
</tbody>
</table>

Through the extraction of relevant relations from the data, this transformation can improve prediction capability, not only maintaining the interpretability according to the original variables but also increasing it. The underlying hypothesis is that such a transformation facilitates both predictive learning and the interpretation of biological mechanisms by humans.

The present results confirm that the transformation can indeed result in increased classification performance and additionally serve as an efficient feature extraction and dimensionality reduction technique.

The prediction results (Figures 2, 3 and 4 and Tables 2 and 3) show an improvement in performance across different layers and cancer types. In addition, we can further see a considerable compression of the information scattered throughout thousands of variables into the new features (Table 4), reaching in some cases, a reduction in dimensionality of 99%.

The impact of the data compression on classification performance can be further seen in Figure 5, where is shown that the transformation actually performs better with an increasingly reduced dimensionality. This can be explained by the way the biclustering algorithm works. Biclustering performs a defined number of iterations of its search, returning all patterns found above a defined lift, some considerably more discriminative than others. It is through the subsequent filtering step that we can obtain the patterns with higher hypothesized discriminative power whilst removing patterns that mainly contributed with noise, resulting in increased performance with lower dimensionalities.

We further show that this transformation creates variables that can more easily distinguish between the conditions under study, contrary to the original features that present highly overlapping conditional distributions (Figure 7). However, the discriminative power of the highest discriminative patterns is diluted when considering all patterns (Figure 6). Still, it is noted that the transformation has a strong impact on the distribution of the data points.

In addition to the listed benefits in prediction capability and dimensionality reduction, the proposed transformation can further be useful for a descriptive analysis of the found patterns, possibly leading to a deeper understanding of the conditions under study. As an example, we look at the pattern found to be the most discriminative in the mRNA layer of the LGG project, denominated as p29 and whose conditional distribution is shown in Figure 7. This pattern is found in 52 out of the total 534 samples, corresponds to high expression levels of genes EMP3, CHISL1, and SH2D4A, and presents a lift of 2.51. Functional enrichment analysis performed through the Enrichr tool (Chen et al., 2013) shows a relation between this gene set and a pathway for Glioblastoma according to the Elsevier Pathway Collection.
Pattern-centric transformation of omics-data sources

(a) Accuracy

(b) Recall

(c) Precision

(d) F1-Score

Fig. 5: Classification performance of distinct dimensionality reduction methods across different dimensionalities (better performance by Pattern-Centric Transformation highlighted in orange).

Furthermore, patterns \textit{p15} (genes \textit{TAGLN2}, \textit{TIMP1}, and \textit{EMP3}) and \textit{p330} (genes \textit{HOXA11-AS}, \textit{HOXA7}, and \textit{NTS}), also present in Figure 7, show relations to pathways concerning proteins involved in Glioblastoma and telomerase reverse transcriptase (TERT) activation in cancer, respectively.

The results here presented open the door to pattern-centric transformations with patterns obtained through methods other than biclustering since as long as the found patterns are significantly discriminative, some benefits will arguably hold. Even in the context of biclustering, there are other alternatives that should be tested, namely, the use of non-constant patterns. Order-preserving biclusters, for example, would capture just the relations of order between genes, adjusting for different base levels of expression between individuals, while coherent-valued biclusters would further enforce that the found patterns follow coherent variations across features. The proposed methodology can be further extended in multiple ways, namely: by changing the method to assess the level of expression of a given pattern in a sample or by filtering the found patterns by other characteristics such as statistical significance and putative biological relevance.

**Competing interests**

No competing interest is declared.

**Author contributions statement**

Conceptualization: all authors; Methodology: A.P., R.H.; Formal analysis: A.P.; Writing - original draft: A.P.; Funding Acquisition: R.H., R.S.C.; Writing - review and editing: all authors; Supervision: R.S.C., R.H. All authors have read and agreed to the published version of the manuscript.
Fig. 6: Class separability in two dimensions for the gene-centric (left) and pattern-centric (right) data.

Fig. 7: Normalized conditional distribution of the most discriminative variables in gene-centric (top) and pattern-centric (bottom) data.

Funding
This work was supported by the Associate Laboratory for Green Chemistry (LAQV), financed by national funds from FCT/MCTES (UIDB/50006/2020 and UIDP/50006/2020), INESC-ID plurianual (UIDB/50021/2020), the contract CEECIND/01399/2017 to RSC and the FCT individual PhD grant to AP (2022.14201.BD).

Data availability
The data underlying this article are available in GDC Data Portal at portal.gdc.cancer.gov, and can be accessed with the project codes TCGA-LGG and TCGA-COAD. The processed data is available at github.

References


