ESANN 2022 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 5-7 October 2022, i6doc.com publ., ISBN 978287587084-1. Available from http://www.i6doc.com/en/.

Interactive dual projections for gene expression analysis

Ignacio Díaz[1], José M. Enguita[1], Diego García[1], Ana González[1], Abel A. Cuadrado[1], María D. Chiara[2] and Nuria Valdés[3] *

1- University of Oviedo - Dept of Electrical Engineering
Edificio Torres Quevedo, módulo 2, Campus de Gijón 33204 - SPAIN

2- Institute of Sanitary Research of the Principado de Asturias
Hospital Universitario Central de Asturias, Oviedo 33011 - SPAIN.

3- Department of Internal Medicine, Section of Endocrinology and Nutrition
Hospital Universitario de Cabueñes, Gijón 33204 - SPAIN.

Abstract. We present an application of interactive dimensionality reduction (DR) for exploratory analysis of gene expression data that produces two lively updated projections, a sample map and a gene map, by rendering intermediate results of a t-SNE. The user can condition the projections "on the fly" by subsets of genes or samples, so updated views reveal co-expression patterns for different cancer types or gene groups.

1 Introduction

Gene expression data analysis is one of the key tools in biomedical research. The underlying processes taking place in cancer involve complex interactions among the genes and other factors, with up or down regulations at different transcriptional stages. As a result, the levels of expressions in the genes form a transcriptomic signature that can be used to study, differentiate, detect and diagnose cancer. Typically, the levels of expressions of thousands of genes are measured for hundreds of samples taken under different biological conditions, such as types of cancers. A widely used approach consists in using DR techniques like the *t-distributed stochastic neighbor embedding* (t-SNE) to visualize the samples in two dimensions [1] so near points represent transcriptomically similar samples. While most papers focus on projecting samples, genes can also be projected, as in [2], where a dual t-SNE is used for transcriptome-wide and sample-wide exploration of gene expression data. However, existing approaches assume transcriptomic similarities in *all* genes and samples, potentially ignoring associations of groups of genes with groups of samples. In this paper, we propose a method based on interactive DR [3], to produce simultaneous free-running sample and gene maps, allowing *human in the loop* user selection of groups of genes or samples to update both maps. We present encouraging results of our approach through three case studies. See demo at `https://gsdpi.edv.uniovi.es/GEM-iDR`.

*This work is part of Grant PID2020-115401GB-I00 funded by MCIN/AEI/ 10.13039/501100011033. The results shown here are based upon data generated by the TCGA Research Network: `https://www.cancer.gov/tcga`.

ESANN 2022 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 5-7 October 2022, i6doc.com publ., ISBN 978287587084-1. Available from http://www.i6doc.com/en/.

2 Materials and Methods

2.1 Gene expression matrix

A *Gene Expression Matrix* (GEM) can be defined as a $n \times m$ matrix $\mathbf{G} = (g_{ij})$ for which the row index, i, represents biological samples under different conditions (e.g. cancer types or subtypes) the column index, j, represents genes and the value g_{ij} represents the expression level of gene j for sample i. We define the *sample vector* $\mathbf{x}_i^s = g_{i*} \in \mathbb{R}^m$ as a vector containing the gene expressions of sample i for all the m genes included in $\mathbf{G}$. Similarly, the *gene vector* $\mathbf{x}_j^g = g_{*j} \in \mathbb{R}^n$ contains the expressions of gene j across the n samples of $\mathbf{G}$.

2.2 DR projections of samples and genes

DR algorithms can be applied to visualize the samples according to their transcriptomic similarity. DR algorithms can produce a mapping $\varphi^s : \mathbb{R}^m \to \mathbb{R}^2$ where $\mathbf{y}_i^s = \varphi^s(\mathbf{x}_i^s)$, that tries to preserve the neighborhood, so sample vectors $\mathbf{x}_i^s, \mathbf{x}_j^s$ that show a similar transcriptomic profile (similarity in the input space), are mapped to close 2D positions $\mathbf{y}_i^s, \mathbf{y}_j^s$. Scatterplot visualizations of the projections $\mathbf{y}_i^s$ reveal a genetic map of the samples in the dataset, whereby clusters represent groups of samples with similar transcriptomic behavior, providing a valuable information for the biomedical scientist that can also be integrated in more sophisticated data analytics interfaces.

Similarly, a *dual* projection can be carried out with the gene vectors $\varphi^g : \mathbb{R}^n \to \mathbb{R}^2$ where $\mathbf{y}_j^g = \varphi^g(\mathbf{x}_j^g)$. A scatterplot visualization of the $\mathbf{y}_j^g$, called *gene view*, complementary to the sample view, reveals the similarities among the genes regarding their expressions across all the samples in the GEM.

2.3 Dual interactive dimensionality reduction

The dynamics of an iterative DR algorithm for the sample view[1] can be conveniently described –see [4], section 5– in terms of a flattened *configuration vector* $\mathbf{y}^s$ containing all the coordinates of the sample projections $\mathbf{y}^s = (y_{11}^s, y_{12}^s, y_{21}^s, y_{22}^s, \ldots, y_{n1}^s, y_{n2}^s)^T$, so $\mathbf{y}^s(k)$ completely defines the current state of the sample view at iteration k. At each iteration, the DR algorithm can be defined in terms of a general nonlinear state function $\mathbf{f}()$ returning an updated configuration, based on the former configuration and an input configuration vector $\mathbf{u}^s(k)$ containing the input data $\mathbf{x}^s(k)$ and the algorithm parameters[2] $\mathbf{w}(k)$. Considering both the sample and the gene projections, the dynamics of convergence resulting from user interaction can be described as:

$$\mathbf{y}^s(k+1) = \mathbf{f}(\mathbf{y}^s(k), \mathbf{u}^s(k)), \qquad \mathbf{y}^g(k+1) = \mathbf{f}(\mathbf{y}^g(k), \mathbf{u}^g(k)) \tag{1}$$

The idea behind interactive DR is to render $\mathbf{y}^s(k)$ and $\mathbf{y}^g(k)$ for *every* iteration k, during convergence, in an infinite loop along which the user can modify or

[1] The gene view has an identical dual formulation.

[2] It may include hyperparameters of the algorithm (such as, perplexity in t-SNE), but may also include parameters of the distance metrics to compute similarities, as used here.

ESANN 2022 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 5-7 October 2022, i6doc.com publ., ISBN 978287587084-1. Available from http://www.i6doc.com/en/.

condition the analysis by acting in $\mathbf{w}(k)$, or even the input data might be changed during the analysis –see Fig.1. The main advantage of this approach is that it results in smooth transitions among the steady-state optima, thereby keeping the user's mental model of the data across the different conditions.

In the problem of analyzing a GEM, the input data $\mathbf{x}^s(k), \mathbf{x}^g(k)$ do not change. In our proposal the user can act on $\mathbf{w}(k)$, by alternatively, selecting a subset of genes to compute the similarities among the samples, or selecting a subset of the samples to compute the similarities among genes. In the first case (the second case is analogous), this can be achieved by changing the distance metric "on the fly" (i.e. during convergence), according to a selection $\mathcal{G}$ of the genes done at interaction time by the user, as $d^s_{\mathcal{G}}(\mathbf{x}^s_i, \mathbf{x}^s_j) = \sum_{k=1}^{n} w^g_k (x^s_{ik} - x^s_{jk})^2$ where $w^g_k = 1$ if k is a selected gene, and $w^g_k = 0$ in other case.

The rationale behind this design is rooted in several biomedical considerations [5]. An effective exploratory analysis of how samples of different types of tumors stay together, or remain apart, due to variations in expression of a single gene or sets of genes, can be used to dissect the complexity and heterogeneity of tumors, and to discover new biomarkers or therapeutic strategies. A common strategy is to classify the transcriptome data into sets of functional modules that are easy to understand; being able to explore novel clusters of genes with similar behaviors for given subsets of samples (e.g. cancer types or subtypes) enable generation of hypotheses that may improve the discovery processes.

3 Results and discussion

We implemented the proposed interactive approach to analyze a total amount of 449 samples with 571 gene expression measurements each, obtained from the TCGA database. The samples contain 157 pheochromocytoma and paraganglioma (PCPG), 221 kidney renal clear cell carcinoma (KIRC) and 71 renal normal tissue. Each sample was defined by the expression levels of 442 hypoxia-related genes and 129 microRNA (miRNA).

3.1 Case 1: behavior of cancer subtypes for specific gene clusters

At the initial state, the interface computes the t-SNE projections[3] of the sample view and the gene view considering all the genes and all the samples. Both views in the initial state are labeled in Fig. 2 with $t = 0s$. In the initial gene view, the user can already observe that genes become spatially organized according to relevant functional groups associated to known mechanisms taking place in different cancer types, such as angiogenesis, development of extracellular matrix (ECM) proteins, and hypoxia. The samples, in turn, are organized in three clearly differentiated clusters corresponding to kidney renal clear cell carcinoma (KIRC, yellow), pheochromocytoma and paraganglioma (PCPG, blue), and renal normal tissue (green). A special group of PCPG with a mutation in the VHL gene affecting its function in the hypoxia pathway has also been marked in red.

[3]We used perplexity=20 and early exaggeration = 12, to ensure a stable initial state.

ESANN 2022 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 5-7 October 2022, i6doc.com publ., ISBN 978287587084-1.
Available from http://www.i6doc.com/en/.

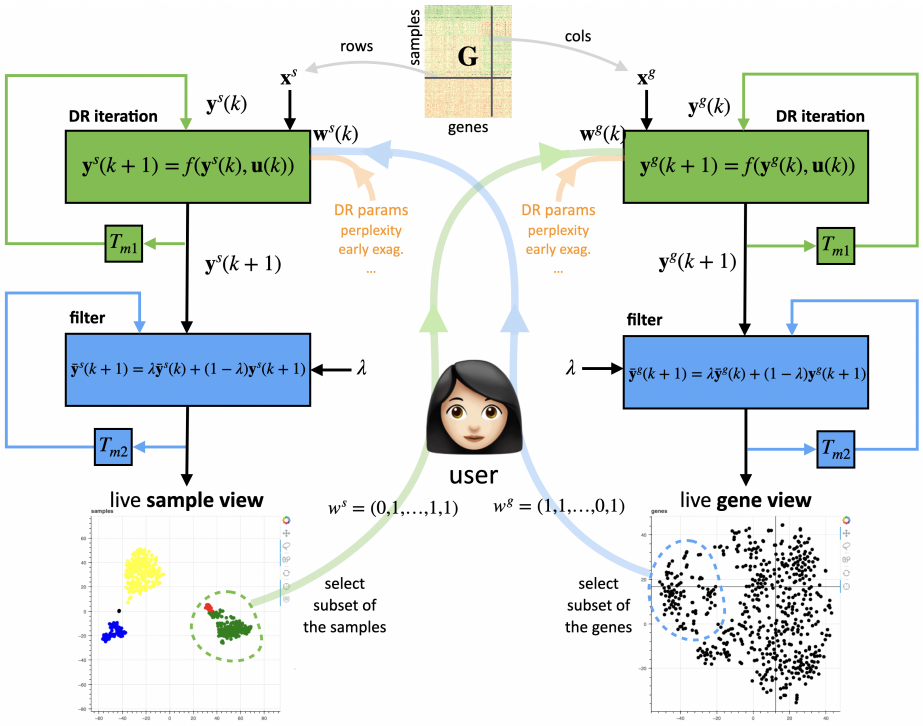

Fig. 1: Workflow of the approach (video: `https://youtu.be/n97DyuV1e24`)

At this point the user can explore how the samples are distributed according to their similarity in specific groups of genes. This can give useful knowledge about whether known mechanisms involved in the selected genes differentially affect some sample groups over others. Selecting the genes in the area related to angiogenesis on the gene view, the user constrains the distance metric to angiogenesis, disregarding the other functions. The three rightmost pictures in Fig. 2 show frames of the animated sample view at three different times $t = \{0s, 6s, 12s\}$. It can be observed that the three initial clusters corresponding to PCPG, KIRC and normal renal tissue, smoothly change, so PCPG and KIRC are merged into a single cluster, while normal tissue remains apart. This suggests a specific role of angiogenesis in cancer processes that takes place in a different way in normal tissue. Obviously, further exploration selecting other groups of genes can be done along the analytics session.

3.2 Case 2: similarities of genes conditioned to cancer type

The proposed approach allows other different yet complementary analyses. In this case, a cluster of microRNA (miRNA) was located by the user in the gene view (red dots in Fig. 3, A). Such cluster corresponds to miRNA that are simi-

ESANN 2022 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 5-7 October 2022, i6doc.com publ., ISBN 978287587084-1.
Available from http://www.i6doc.com/en/.

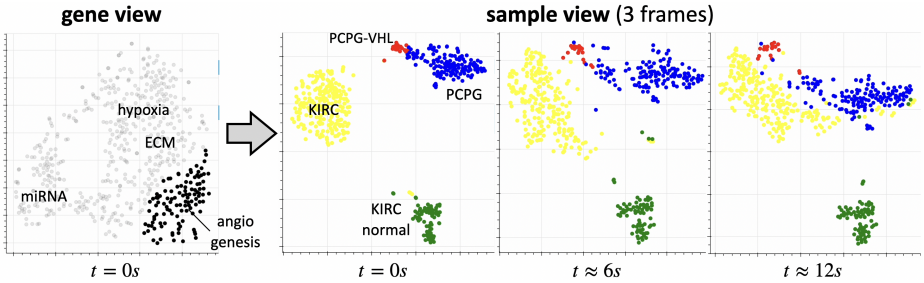

Fig. 2: Case 1. Impact of angiogenesis genes in characterization of KIRC/PCPG

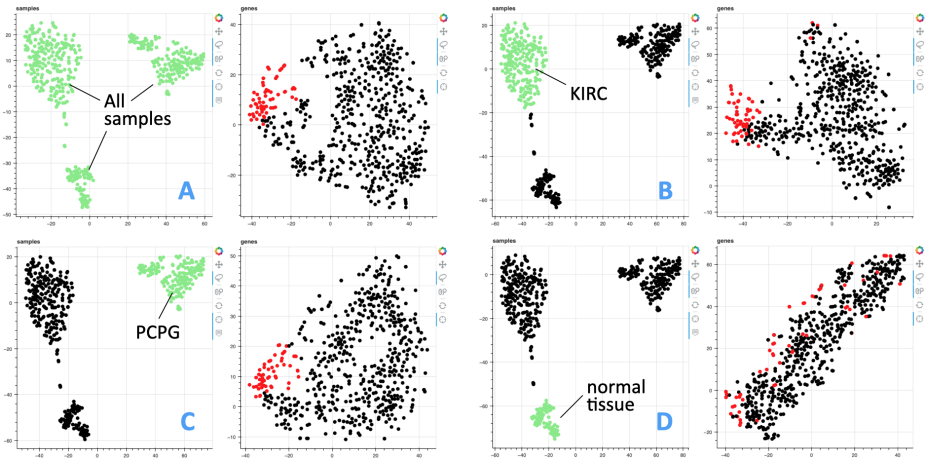

Fig. 3: Different behavior of miRNA group for cancer subtypes

larly expressed across all samples, presumably sharing some common functionality or relationship. The user wants to check if the behavior of these elements is similar for all the cancer subtypes. Selecting samples from one cancer type only, reveals that the miRNA still form a distinct cluster (Fig 3, B and C). However, selecting normal samples causes a spread of the group of miRNAs along the whole gene map, revealing a remarkably different behavior of these miRNA for normal samples (Fig. 3, D).

3.3 Case 3: over/under expression of genes across cancer subtypes

Another simple yet insightful use is to reveal singular levels of expression for a gene. We explore the extreme expressions of miRNA `-210-3p`, which has been shown in last years to be more consistently over-expressed in hypoxia tumors. Arranging the samples using only this miRNA, leads to a 1D snake-shaped projection with all the samples sorted by `-210-3p` expression value. The user may

ESANN 2022 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 5-7 October 2022, i6doc.com publ., ISBN 978287587084-1. Available from http://www.i6doc.com/en/.

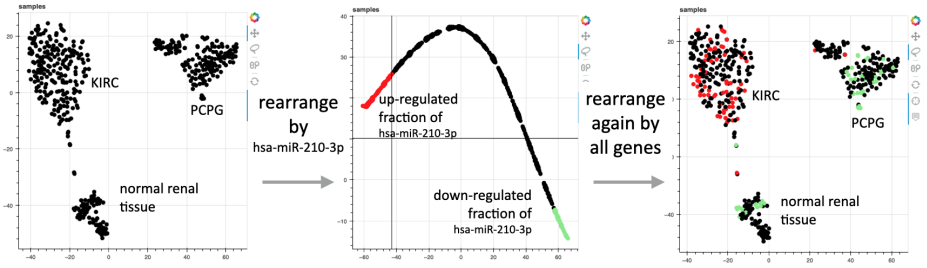

Fig. 4: Case 3. Revealing location of up/down regulated `hsa-miR-210-3p`.

select and mark the head and tail sections with red and green colors. Rearranging again the view using all genes, it smoothly returns to its original configuration, revealing that this miRNA is overexpressed in KIRC [6].

4 Conclusion and future work

We have presented an interactive DR method able to produce two lively updated projections of samples and genes, that can be conditioned by subsets of the samples or the genes. The results presented through three case studies on TCGA data, show how biomedically relevant patterns and relationships can be found, and suggest this may be powerful approach for exploration of gene expression data in combination with gene expression matrix visualization techniques.

References

[1] Dmitry Kobak and Philipp Berens. The art of using t-sne for single-cell transcriptomics. *Nature communications*, 10(1):1–14, 2019.

[2] Sjoerd MH Huisman, Baldur Van Lew, Ahmed Mahfouz, Nicola Pezzotti, Thomas Höllt, Lieke Michielsen, Anna Vilanova, Marcel JT Reinders, and Boudewijn PF Lelieveldt. Brainscope: interactive visual exploration of the spatial and temporal human brain transcriptome. *Nucleic acids research*, 45(10):e83–e83, 2017.

[3] Ignacio Díaz, Abel A Cuadrado, Daniel Pérez, Francisco J Garcia, and Michel Verleysen. Interactive dimensionality reduction for visual analytics. In *2014 European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN 2014)*, 2014.

[4] Alex Endert, William Ribarsky, Cagatay Turkay, BL William Wong, Ian Nabney, Ignacio Díaz, and Fabrice Rossi. The state of the art in integrating machine learning into visual analytics. In *Computer Graphics Forum*, volume 36, pages 458–486. Wiley Online Library, 2017.

[5] Marcin Cieślik and Arul M Chinnaiyan. Cancer transcriptome profiling at the juncture of clinical translation. *Nature Reviews Genetics*, 19(2):93–109, 2018.

[6] Michelle E Watts, Sarah M Williams, Jess Nithianantharajah, and Charles Claudianos. Hypoxia-induced microrna-210 targets neurodegenerative pathways. *Non-coding RNA*, 4(2):10, 2018.