Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Manuscript Category

Subject Section

Morphing Projections: a new visual technique for fast and interactive large-scale analysis of biomedical datasets

Ignacio Díaz [1]*, José M. Enguita [1], Ana González [1], Diego García [1], Abel A. Cuadrado [1], María D. Chiara [2,3] and Nuria Valdés [4]

[1] Department of Electrical Engineering, University of Oviedo, Gijón, 33204, Spain
[2] Institute of Sanitary Research of the Principado de Asturias, Hospital Universitario Central de Asturias, 33011 Oviedo, Spain.
[3] CIBERONC (Network of Biomedical Research in Cancer), 28029 Madrid, Spain.
[4] Section of Endocrinology and Nutrition, Hospital Universitario de Cabueñes, Gijón, Spain.

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation. Biomedical research entails analyzing high dimensional records of biomedical features with hundreds or thousands of samples each. This often involves using also complementary clinical metadata, as well as a broad user domain knowledge. Common data analytics software makes use of machine learning algorithms or data visualization tools. However, they are frequently *one-way* analyses, providing little room for the user to reconfigure the steps in light of the observed results. In other cases, reconfigurations involve large latencies, requiring a retraining of algorithms or a large pipeline of actions. The complex and multiway nature of the problem, nonetheless, suggests that user interaction feedback is a key element to boost the cognitive process of analysis, and must be both *broad* and *fluid*.
Results In this paper we present a technique for biomedical data analytics, based on blending meaningful views in an efficient manner, allowing to provide a natural smooth way to transition among different but complementary representations of data and knowledge. Our hypothesis is that the confluence of diverse complementary information from different domains on a highly interactive interface allows the user to discover relevant relationships or generate new hypotheses to be investigated by other means. We illustrate the potential of this approach with two case studies involving gene expression data and clinical metadata, as representative examples of high dimensional, multidomain, biomedical data.
Availability and implementation Code and demo app to reproduce the results available at https://gitlab.com/idiazblanco/morphing-projections-demo-and-dataset-preparation
Supplementary information Supplementary data is available at Bioinformatics online.

1 Introduction

Biomedical data is growing at astonishing rates with the broadening of access to massive analyses, including laboratory tests, medical image, or gene expression data, bearing precise information about the underlying biological state of the subject. However, analysis of high dimensional biomedical datasets is rarely directly managed by doctors, even though they encompass a unique set of challenges, hard important problems, and huge potential sanitary impact.

At the same time as quality and availability of biomedical data increased, accordingly did its application in biomedical research. Most interestingly, there has also been a huge research activity in the field of visualization of this high-dimensional data, as a very powerful tool for researchers, providing better data interpretation, easier detection of patterns, and generation of new hypotheses. Related surveys about the use of these techniques can be found in (Kamal *et al.* (2014); O'Donoghue *et al.* (2018)).

In this paper, we set the focus on the analysis of genomic data for cancer research, which is a paradigmatic example of such problems. Many processes related to the onset, spreading and evolution of cancer, are strongly related to biological pathways that involve complex chains of regulation/deregulation mechanisms acting on different stages of the gene expression path, and resulting in specific footprints in the gene expression pattern that coexist with many others coming from normal biological processes. The large amount of coexisting processes, including normal and aberrant cancer-related ones, along with the large number (tens of thousands) of involved genes in the human genome, make the analysis process of gene expression data a challenging task, even for modern machine learning techniques, that requires a joint approach that combines data analysis and expert knowledge from other domains. Detecting genes with differential expression between groups of samples, finding relationships between the expression levels for groups of genes (co-expressed genes), identifying prognostic genes or finding genetic disorders, are just a few examples. Classical analysis methods, however, were mainly based on clustering, regression, correlation and other similar statistical analysis techniques (Trevino *et al.* (2007)), which soon started showing limitations to handle this complexity.

Moreover, in recent years, micro-RNA (miRNA) data has also attracted a lot of attention, as some miRNAs were identified as prognostic markers or associated with overall survival for specific type of cancers (Di Leva and Croce (2013); Akcakaya *et al.* (2011); Võsa *et al.* (2011); Eyking *et al.* (2016); Xu *et al.* (2013); Jones *et al.* (2014)). Since then, a lot of research has been oriented to determine correlations or interactions between miRNAs and specific genes in different types of cancer (Dai *et al.* (2019); Telonis *et al.* (2017); Tan *et al.* (2019); Hu *et al.* (2018)). The also recent boost of machine learning techniques resulted in their successful application for cancer detection and treatment (Cheerla and Gevaert (2017)).

Thus, genomic approaches in cancer research have a strong multi-domain nature, posing interesting problems in the co-analysis of gene and miRNA expression levels, but also requiring complementary data that may include a broad range of clinical information and related metadata of the samples under study. This cross-domain nature of the problem, along with the high dimensionality and volume of data, as well as the nonlinear complexity of the underlying processes make it a challenging representative kind of biomedical data.

In the light of this scenario, the need to broaden the scope of analysis brought by machine learning approaches, resulted in a huge research effort. (Nusrat *et al.* (2019)) provide a deep analysis of the state-of-the-art visualization techniques used for genomic data visualization, as well as a taxonomy of the most widely used. Their study reveals many different options regarding layout, arrangement, encoding, scales, and so forth. Each one of these views has different strong and weak points depending on the intended objective. Other authors propose the use of machine learning-based dimensionality reduction techniques, such as the t-distributed Stochastic Neighbor Embedding (t-SNE), proposed back in 2008 (Van Der Maaten and Hinton (2008)). This technique has been used for single-cell transcriptomics (Kobak and Berens (2019)) and analysis of the gene expression in the mouse brain (Pezzotti *et al.* (2016)), for instance. Also, several computer applications have been recently reported for visualization of genomic data (Ding *et al.* (2018); Huisman *et al.* (2017); Egorov *et al.* (2019)).

However, few works take advantage of the synergies between interaction, data visualization and machine learning (ML), allowing only for simple interactions such as tuning the visualization algorithm or selecting between different views of the data. In a review of this topic, often referred as visual analytics (Endert *et al.* (2017)), the authors suggest a broad niche of opportunities yet to be explored in this field. Algorithmic approaches allow the analysis of massive data, being able to outperform humans in well-defined tasks, but prone to failure under minimal changes in the context or the problem statement. Humans, in turn, benefit from a broad domain knowledge, being able to work on ill-posed problems and perform reasonably well in a vast range of tasks, and are able to find connections and improve answers by means of an iterative and exploratory cognitive process.

This cognitive feedback is depicted in Fig. 1. In the one-way approach (left), common in many conventional visualizations, the user U consumes a fixed outcome of a data visualization or ML software V , requiring a long time (or not allowing) to modify the problem formulation, thereby lending to a narrow analysis and/or a slow improvement in the acquisition of problem-related knowledge. Interactive analysis (right), in turn, introduces feedback in the process, which may dramatically change the outcome of the analysis, resulting in a qualitative improvement of the overall cognitive process. The broader the interaction highway, the higher the bandwidth and the throughput of the knowledge discovery process.

In this sense, many tools featuring interaction and ML have been proposed in recent years for genomic data visualization. For instance, tools like iDREM (Ding *et al.* (2018)), NetworkAnalyst (Xia *et al.* (2015); Zhou *et al.* (2019)) or VIGLA-M (Navas-Delgado *et al.* (2019)) use ML algorithms in different ways using gene expression data and known interactions (e.g. protein-DNA interaction) and produce interactive visualizations of *gene regulatory networks* that allow some kinds of user interaction (zoom, selection, 3D rotation, etc.). Interactive *heatmaps*, enriched with cluster information and interactions like selection, filter, sort, etc. are also another common way of visual analytics in genomics used in tools like GiTools (Perez-Llamas and Lopez-Bigas (2011)) and also VIGLA-M. Other kind of tools provide *context* to the researcher, combining gene expression data and *anatomical visualizations* by showing expression levels of selected genes on the anatomical views using color scales on organs (GEPIA, Tang *et al.* (2017)) or cloropleth-like maps (BrainScope, Huisman *et al.* (2017)). All these tools, in addition, allow the user to carry out different ML computations, like principal component analysis, or differential expression analysis. However, most of them involve a long workflow (ranging from a few seconds to many minutes) for the user from the time the researcher produces a hypothesis until she obtains a result for validation or suggestion of new hypotheses.

Few tools allowing for a truly *fluid exploration* have been proposed, such as SPRING (Weinreb *et al.* (2018)), which includes a steerable force-directed layout built out from gene expression data able to reveal biological relationships on a user-steered, dynamically changing graph topology overlaid with gene expression and other annotations. Despite having a strong visual analytics gist, featuring steerable ML in a dynamic visualization, the method does not consider ways to integrate context information in the layout, such as clinical data, thereby narrowing the possibilities of connecting knowledge from other domains. Also, the need to compute a force layout limits the allowable sample size for the analysis.

In a former work (Diaz-Blanco *et al.* (2012)), the authors presented an interaction technique for data visualization called morphing projections (MP). Despite being a preliminary work, focused on an electrical engineering problem, the underlying idea allowed a comprehensive analysis of multiway data in a highly, multimodal, interactive manner, suggesting it could be well suited to deal with biomedical data, and particularly to explore the highly complex landscape of biological processes in genomic analysis. In this paper we develop this idea in a formal and general way, providing a mathematical framework that connects the approach with relational algebra principles, extending the functionality to allow the combination of an arbitrary number of factors of analysis according to researcher's needs, and propose a methodology for exploratory analysis of complex high dimensional biomedical data. The MP approach is based on reconfigurable data visualizations composed of a user-driven mixture of basis views, each highlighting conceptually

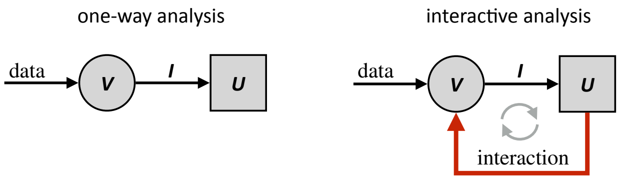


Fig. 1. One-way analysis vs. interactive analysis. In the one-way analysis (left), the user U consumes fixed visual information I produced by visualization software V from the available data. With the interactive analysis (right), the user can modulate V upon received information I by means of interaction, in a continuous improvement cycle.

different traits of the samples under study. We will illustrate the MP technique through some case studies on genomic data analysis, considering basis views that include the biological state obtained by dimensionality reduction projections (e.g. tSNE, deep autoencoders) of gene expression vectors, as well as available relevant related context information (e.g. clinical data, such as cancer type, gender, race, etc). The methodology can be extrapolated to any other biomedical data including vectors of descriptors and/or context data relevant to the target problem, such as those obtained from lab tests, biomedical equipment, etc.

2 Materials and methods

2.1 Materials: data sources and preprocessing

The Cancer Genome Atlas (TCGA) provides gene expression measurements and other transcription data, including more than 20,000 RNA and hundreds of miRNA expression levels, of more than 10,000 tumors from 33 different cancer types. To show the functionality and possibilities of our proposed method we considered gene expression RNAseq data of a selection of 31 cohorts from the TCGA Hub. Data for the cohorts included in our results were downloaded from the Xenabrowser portal (`https://xenabrowser.net/datapages/`). For every cohort we merged a) data containing experimental measurements using the Illumina HiSeq 2000 RNA Sequencing platform, and mean-normalized per gene across all cohorts and b) data with miRNA mature strand expression RNAseq. The resulting table including gene and miRNA expressions was curated by dropping genes with invalid values, and later merged with clinical metadata (downloadable from `https://portal.gdc.cancer.gov`) which included tumor stage, existence of metastasis, race, gender, etc. After processing we got a curated multivariate data table, used in the reported results and demos, with 8580 tumor samples (rows) belonging to 31 cancer types with 19279 attributes (columns) that include 19112 gene expression values, 129 miRNA expression values and 18 clinical variables.

2.2 Methods: The Morphing framework

We present in this section a framework rooted on relational algebra to formalize the idea of using animated transitions (Heer and Robertson (2007)) between two or more meaningful views by means of *morphing operations*. Morphing operations allow to interactively rearrange a visualization composed of many items (e.g. points of a scatterplot) by means of smooth transitions between several views corresponding to arrangements of data according to different grouping criteria.

2.2.1 Multivariate data table

Gene expression data are often presented as a *gene expression matrix* (GEM) (Jiang *et al.* (2004); Roche *et al.* (2018)), whose rows represent the biological samples and the columns contain the gene expression levels. This information can be extended by adding extra columns with clinical descriptors or other available context information about the samples, resulting in a *multivariate data table*. We shall refer to the rows (samples) as *records* and the columns (gene expressions and clinical data) as *attributes*. Each attribute is the result of discretizing a variable into a finite set of *groups* thereby making it to take a finite set of values (e.g. a set of intervals or bins, a set of classes, or simply the set of unique values of the variable along all records). More formally, a multivariate data table can be defined by the relation D:

$$\begin{array}{c|c} \multicolumn{2}{c}{D} \\ id & a_1, a_2, \cdots a_n \\ \hline \vdots & \vdots \end{array}$$

where each row defines a sample for which *id* is a unique identifier (e.g. primary key, timestamp, etc.), and $a_1, a_2, \cdots, a_n$ are attributes, where attribute a_i takes values from a discrete set $\{g_1^i, g_2^i, \ldots, g_{n_i}^i\}$. In general, the elements g_j^i may define *groupings* of elements in the sample for attribute a_i. For categorical attributes, they will typically represent class labels, whilst for continuous valued attributes the set may contain bins defined by intervals of a variable or all the unique values of this variable present in the sample.

2.2.2 Spatial encodings

A *spatial encoding* can be formally expressed by means of a *relation* E_i:

$$\begin{array}{c|c} \multicolumn{2}{c}{E_i} \\ a_i & P^i \\ \hline g_1^i & \mathbf{p}_1^i \\ \vdots & \vdots \\ g_{n_i}^i & \mathbf{p}_{n_i}^i \end{array}$$

that maps the group value g_j^i of attribute a_i to a position $\mathbf{p}_j^i \in \mathbb{R}^n$. These positions, supposed designed for an interpretable arrangement (e.g. using visual conventions like clock-like, linear or matrix arrays, map coordinates, or distance-preserving dimensionality reduction algorithms), will typically represent the spatial coordinates of an item in a visualization, but more generally can express other visual attributes such as color, size, shape, etc. The encoding E_i can be seen as a *lookup table* to assign a position to every sample (row) of D according to its group value in the attribute a_i. The resulting positions allow to define visualizations that spatially arrange the samples according to attribute a_i in some meaningful way.

2.2.3 Extended dataset

A new table containing the positions of all the samples of D, according to the encoding E_i, can be defined by means of a *natural join* operation:

$$D_{E_i} = D \bowtie E_i \,. \tag{1}$$

The *natural join* operator $\bowtie$ is equivalent to a Cartesian product with the restriction of equality of the common attribute a_i of both relations, resulting in an extended relation D_{E_i} that contains the positions for each sample, as defined in the encoding E_i:

$$\begin{array}{c|c|c} \multicolumn{3}{c}{D_{E_i}} \\ id & a_1, a_2, \cdots a_n & P^i \\ \hline \vdots & \vdots & \vdots \end{array}$$

For each sample, the natural join operation matches the group value attribute a_i to the position assigned to this group by the spatial encoding E_i. Using the conventions of relational algebra, the set of n_i positions for the encoding E_i, as well as the positions for every sample in D, can be respectively obtained in a quite straightforward way as:

$$\Pi_{P^i} D_{E_i} = \Pi_{P^i}(D \bowtie E_i)\,, \tag{2}$$

where $\Pi_x(Y)$ is the *projection operator*, that returns the set of attributes x from relation Y.

2.2.4 Aggregation operations

Similarly, the *groupby* and *aggregation* operations in relational algebra can be used to compute the aggregated values of attribute a_r for the groups g^i_j defined for attribute a_i, to be visualized at positions $\mathbf{p}^i_1, \cdots, \mathbf{p}^i_{n_i}$:

$${}_{a_i}\mathcal{G}_{A(a_r)}(D_{E_i})\,, \tag{3}$$

where $\mathcal{G}$ is the "group by" operator, a_i is the grouping attribute, and $A(\cdot)$ is an aggregate function applied to attribute a_r, that can be typically *sum, count, average, maximum, minimum* or another predefined aggregator.

2.2.5 Morphing operation

Let's denote $\mathcal{C}_N$ the set containing all the sets of N points or positions in $\mathbb{R}^n$. Let's consider p sets of points $P^i = \{\mathbf{p}^i_k\}_{k=1,\cdots,N}$, with $\mathbf{p}^i_k \in \mathbb{R}^n$, and $P^i \in \mathcal{C}_N$ for $i = 1, \ldots, p$. Typically, for $n = 2$, each set of points can be thought to describe a spatial configuration or layout composed of markers in a 2D scatterplot visualization; for larger values of n other visual attributes like color, size, etc., or even more general parameterized graphs can be considered. The *morphing operation* takes the p spatial configurations (i.e. the p sets of N positions) and parametrically returns a new spatial configuration according to a set of user-driven *interaction parameters* $\mathbf{t}$ and a convex *mixing function* λ. More formally, the morphing operation can be defined as $\mu_{\lambda,\mathbf{t}} : \mathcal{C}_N \times \cdots \times \mathcal{C}_N \to \mathcal{C}_N$, being:

$$\mu_{\lambda,\mathbf{t}}(P^1, P^2, \ldots, P^p) \to P^\lambda\,, \tag{4}$$

where $P^\lambda = \{\mathbf{p}^\lambda_k\}_{k=1,\cdots,N}$, and:

$$\mathbf{p}^\lambda_k = \sum_{i=1}^{p} \lambda_i(\mathbf{t})\mathbf{p}^i_k \qquad k = 1, \cdots, N. \tag{5}$$

The mixing function $\lambda : \mathbb{R}^q \to \mathbb{R}^p$ takes a vector of q interaction parameters $\mathbf{t} = t_1, \ldots, t_q$ to produce p *mixing coefficients* $\lambda_1(\mathbf{t}), \ldots, \lambda_p(\mathbf{t})$ such that $\sum_{i=1}^p \lambda_i(\mathbf{t}) = 1$.

Example 1. Linear morphing between two configurations. As a particular case, for $q = 1, p = 2$, the morphing operation $\mu_{\lambda,\mathbf{t}}(A, B)$ having $\lambda_1(\mathbf{t}) = 1 - t$ and $\lambda_2(\mathbf{t}) = t$ for $t \in [0, 1]$, results in a dynamically changing set of points that smoothly *morph* from a configuration A for $t = 0$, to a configuration B for $t = 1$.

Example 2. Softmax morphing among several configurations. Another specially useful case considering one tuning parameter *per* encoding ($q = p$) is using the *softmax* function for λ:

$$\lambda_i(\mathbf{t}) = \frac{e^{\alpha t_i}}{\sum_{j=1}^p e^{\alpha t_j}}\,, \tag{6}$$

being α a *sensitivity* parameter that tunes the degree of approximation to the standard *max* function. This function allows the user to smoothly highlight any specific spatial configuration of data by rising its parameter t_i above the others.

Example 3. Trajectory along a path of encodings. Another interesting example is using morphing to make a transition along a trajectory or *path* defined by a sequence of encodings $P^1, \ldots, P^p$ considering a single tuning parameter $t \in [1, p]$:

$$\lambda_i(t) = \frac{\phi_i(t)}{\sum_j \phi_j(t)}, \quad \text{where} \quad \phi_i(t) = e^{-\|t-i\|^2/\sigma^2} \tag{7}$$

for $i = 1, \ldots, p$. This creates a trajectory that smoothly morphs through $P^1, \ldots, P^p$ as t ranges from 1 to p, being σ a smoothness parameter. This kind of morphing is useful to reveal dimensionality reduction mappings of time-evolving data or, more generally, algorithmically generated spatial representations of data for variations of a meaningful parameter.

2.2.6 Spatial configuration of data

Based on the morphing operation, a new encoding relation E_λ containing the lookup table of the *blended* positions for all possible combinations of group values from attributes $a_1, \ldots, a_p$ can be obtained in a general way:

$$\begin{aligned} E_{\lambda,\mathbf{t}} &= \Pi_{a_1,\ldots,a_p,\mu_{\lambda,\mathbf{t}}(P^1,\ldots,P^p)}\,(E_1 \times \cdots \times E_p) \\ &= \Pi_{a_1,\ldots,a_p,P^\lambda}\,(E_1 \times \cdots \times E_p)\,, \end{aligned} \tag{8}$$

where the Cartesian product gives all possible combinations of attributes and positions from $E_1, \ldots, E_p$, and the projection operator takes the actual values for the p classes $a_1, \ldots, a_p$ and the new *blended* positions given by the morphing operation $\mu_{\lambda,\mathbf{t}}$ between the positions P^i:

$E_{\lambda,\mathbf{t}}$

$a_1, \ldots, a_p$	P^λ
$\vdots\ \vdots$	$\vdots$

Proceeding consistently, the resulting extended dataset with the new positions is $D_{E_{\lambda,\mathbf{t}}} = D \bowtie E_{\lambda,\mathbf{t}}$, whose positions for representation are $\Pi_{P^\lambda} E_{\lambda,\mathbf{t}}$ and the aggregated values for an attribute a_r can be computed as ${}_{a_1,\ldots a_p}\mathcal{G}_{A(a_r)}(D \bowtie E_{\lambda,\mathbf{t}})$.

2.2.7 Examples of typical encodings

Circular encoding. A *circular encoding* for attribute a_i locates all elements belonging to groups g^i_j to a discrete set of n_i positions equally distributed in a circle like a *clock*:

$$g^i_j \to \mathbf{p}^i_j = (\cos(2\pi j/n_i), \sin(2\pi j/n_i)), \quad j \in \{1, \ldots, n_i\}\,. \tag{9}$$

Linear and matrix encodings. A *linear encoding* (vertical or horizontal) for attribute a_i locates all elements belonging to groups g^i_j to a discrete set of n_i positions equally distributed in a vertical or horizontal row:

$$g^i_j \to \mathbf{p}^i_j \;=\; (0, j), \quad j \in \{1, \ldots, n_i\} \quad \text{(vertical)}\,, \tag{10}$$

$$g^i_j \to \mathbf{p}^i_j \;=\; (j, 0), \quad j \in \{1, \ldots, n_i\} \quad \text{(horizontal)}\,. \tag{11}$$

A linear morphing operation $\mu_{\lambda,t}(P^h, P^v)$ between an horizontal and a vertical encoding for attributes a_h and a_v yields a *matrix encoding* consisting of a regular grid with all combinations of possible values for both encodings.

Dimensionality reduction encodings. Dimensionality reduction encodings for a given attribute a_i map all its elements according to a *spatialization principle* ("similar ≈ close"), assigning close positions $\mathbf{p}^i_j$ and $\mathbf{p}^i_k$ to elements from groups g^i_j and g^i_k as long as they are similar in some sense. For this type of encoding the most typical case assumes one group per sample in the dataset and that a high dimensional feature vector $\mathbf{x}_j$ (e.g. composed of expressions of a selected group of genes or miRNA) is available for every sample j. Computation of positions $\mathbf{p}_j^i$ can be done using nonlinear dimensionality reduction algorithms, such as tSNE ($t-$*stochastic neighbor embedding* Van Der Maaten and Hinton (2008)). The tSNE considers the conditional probability of neighborhood in the high dimensional space:

$$p_{j|i} = \frac{\exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/2\sigma_i^2)}{\sum_{k\neq i}\exp(-\|\mathbf{x}_i - \mathbf{x}_k\|^2/2\sigma_i^2)} \, , \tag{12}$$

and a Student $t-$distribution of probability of the projected points $\mathbf{y}_i$ in the visualization space:

$$q_{j|i} = \frac{(1+\|\mathbf{y}_i - \mathbf{y}_j\|^2)^{-1}}{\sum_{k\neq i}(1+\|\mathbf{y}_i - \mathbf{y}_k\|^2)^{-1}} \, . \tag{13}$$

The algorithm minimizes the Kullback-Leibler distance $KL(p_{j|i}||q_{j|i})$ between both distributions.

2.3 Morphing projections in gene expression analysis

2.3.1 Morphing between tSNE views

Consider the *expression vectors* $\mathbf{x}_j^1$ and $\mathbf{x}_j^2$, each composed of the expression levels of different groups of genes for sample j. Both vectors can be thought to describe the genetic state of sample j according to different biological processes. Let's consider sets of points $P_j^1 = \{\mathbf{p}_j^1\}$ and $P_j^2 = \{\mathbf{p}_j^2\}$, containing the tSNE projections of $\mathbf{x}_j^1$ and $\mathbf{x}_j^2$, respectively, for all samples $j = 1,\ldots,N$ on a low dimensional space (e.g. 2D meant for visualization):

$$\mathbf{x}_j^1 \to \text{tSNE} \to \mathbf{p}_j^1 \qquad \mathbf{x}_j^2 \to \text{tSNE} \to \mathbf{p}_j^2 \, .$$

Point sets P^1 and P^2 reveal the similitudes between the samples according to the gene activities included in $\mathbf{x}^1$ and $\mathbf{x}^2$. Similar gene expression profiles will lead to close positions in a scatterplot representation of the positions, where clusters contain groups of samples with similar biological states that, depending on the chosen set of gene expressions, may reveal cell differentiation (tissues), cancer-related biological processes, etc. Combining P_j^1 and P_j^2 by means of the mixing function λ as $\lambda_1(t) = 1-t$ and $\lambda_2(t) = t$ a new set of points P^λ for visualization can be obtained as:

$$\mathbf{p}_j^\lambda = (1-t)\mathbf{p}_j^1 + t\mathbf{p}_j^2 \, . \tag{14}$$

The resulting point set P^λ smoothly transitions between representations P^1 and P^2. The representation of P^λ for user-driven values of t gives insight in the differences between both sets of genes for classifying the biological activity of the samples. Note that within the context of a visual display, available context information for each point, such as the cancer type, severity, metastasis condition, etc. can be encoded using color, size, tooltips or different markers, so the user will have immediate visual feedback on subtle differences regarding the ability of both groups of genes in providing differentiation between target biological conditions.

Fig. 2 (a) shows five steps (frames) of the morphing representation between two tSNE views of samples with prostate cancer; the left view ($t = 0$), shows the tSNE projection based on all available genes in the dataset, 19112, and the right view ($t = 1$), shows the tSNE projection based on expressions of 129 miRNA.

The resulting animated transition between both projections (intermediate values of t between 0 and 1) reveals "on the fly" different groupings of the samples of the same cancer using both representations. Apart from a detectable group $\mathbf{A}$ in the gene view revealing samples with a similar global gene behavior, a particularly noticeable, and potentially more significant cluster, $\mathbf{B}$, emerges within seconds (a user gesture in the interface) when giving more weight to the miRNA view, which might have clinical relevance with implications in diagnosis, prognosis or analysis of response to treatment, and may also suggest possibilities of biomarkers with potentially useful clinical applications.

Note also that during navigation between views the user can visually track a selection the samples along the change; this allows the user to keep in mind their belonging to the emerging groups, being a powerful aid in the identification of the common traits of the discovered groups, specially if combined with tooltip and selection mechanisms.

2.3.2 Morphing between tSNE and clinical views

Clinical data provide essential information to complement gene expression data, since in most cases the ultimate target is to discover connections between the gene activity and its clinical manifestation. Typical clinical variables include cancer type, gender, ethnicity, type of tumor (metastatic), stage of the tumor, etc. These variables often contain a reduced number of groups (e.g. two groups male/female for sex) so that all samples are mapped to positions $P^{\text{clinical}} = \{\mathbf{p}_j^{\text{clinical}}\}$. Fig. 2(b) shows four morphing steps between an all-gene tSNE projection for three cancer types (prostate, ovarian, melanoma) and a dominant gender encoding with data split around two different locations $\{\mathbf{p}^{\text{male}}, \mathbf{p}^{\text{female}}\}$. The morphing sequence shows how for gender-specific cancers (prostate and ovarian, dark and light blue, respectively) points are *not* split towards the two male/female positions in the gender view on the right, while for melanoma (green), which affects both genders, the cluster is split into two groups of points that move towards the two end positions in the gender view. Note that due to the linear nature of the morphing operation, the relative positions among samples of a same gender are preserved in the right view, allowing for an independent (and consistently comparable) analysis on both groups.

3 Results

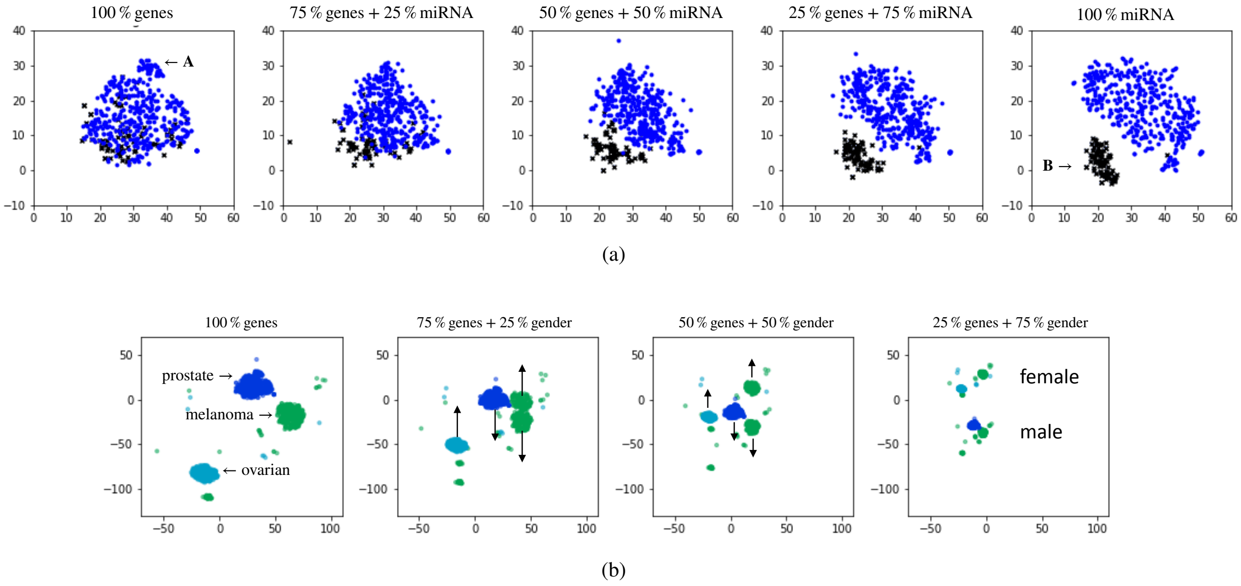
We present in this section three case studies using MP to explore data including gene expressions and related clinical metadata. It should be pointed out that, despite the descriptions presented here are rather detailed, in all cases the whole discovery process may take only a few seconds.

Case 1 specifically focuses in the workflow and in revealing the strengths of the method in presenting gene expression and disease information in a rich number of ways, with agile procedures to change among qualitatively different perspectives of a problem. The video, available in the supplementary material, includes first a free use of the tool (without a predefined goal analysis), demonstrating how its exploratory richness favors discovery. Then, as an example, it shows how salient displacement of certain samples between gene and miRNA expression tSNEs reveals a misclassification in the original TCGA dataset and explains the reason for that. This second part is detailed below.

Cases 2 and 3, also with accompanying videos in the supplemental material, describe the discovery process using the MP approach to validate known results in the literature. It must be pointed out that the focus of the examples are not the results themselves, but to show how interactivity and user steered adaptive views speed up the discovery process and pose a highly competitive alternate way of analysis over other methods in the exploration tasks.

3.1 Case study 1: Morphing Projections as an exploratory tool

This case study starts with the tSNE map of the gene expression for all the samples in the TCGA set. This map offers a *big picture* where samples appear well clustered by tissues due to the strong influence of gene activity for cell differentiation over other factors. An alternative tSNE map



(a)

(b)

Fig. 2. Morphing between two views showing several transition frames: a) tSNE-genes vs. tSNE-miRNA in prostate cancer. MP reveals two different groupings **A** and **B** of the samples, according to gene (left) or miRNA (right) expression levels. Cluster **B** emerges while the user changes (in a quick gesture, moving a slider) the view weights from genes to miRNA; b) tSNE-genes vs. gender. Melanoma samples (green) are split in two clusters (male/female). Gender specific tumors (prostate and ovarian, dark and light blue), in turn, are not split.

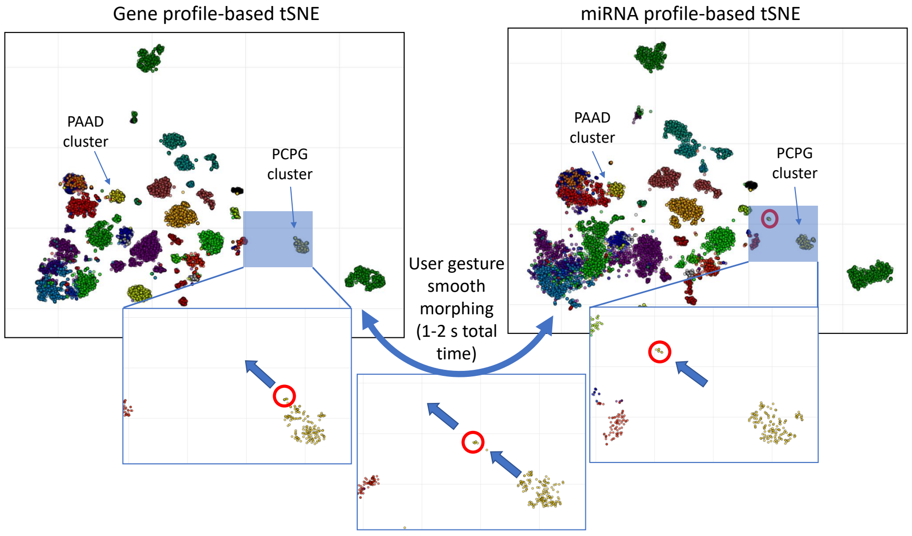


Fig. 3. Morphing between two tSNE maps: based on gene expression (left) and on miRNA expression (right). Two clusters have been highlighted: PCPG, containing pheochromocytomas and paragangliomas, and PAAD, which are pancreatic adenocarcinomas. The left frame shows, inside a red circle, 6 samples located next to the PCPG cluster which are classified in the TCGA database as PAAD tumors (although the used palette mapping has assigned them similar hues, making them almost indistinguishable). Increasing the encoding for the miRNA tSNE, reveals that these samples move away together (depicted in the three insets). These samples are, in fact, pancreatic neuroendocrine tumors (panNETs). The morphing operation reveals that, though they share common gene profiles, they can be differentiated by their miRNA profile. As seen in the accompanying video included in the supplementary material, the whole exploratory process seen in the figure just takes the time of a drag gesture as the user moves a slider (a few seconds). The picture is updated all along in real time so the user never loses continuity.

representation can be made using the miRNA expression, where samples cluster by similar miRNA profiles. A morphing operation between these two maps allows the user to track samples that fly away their original cluster, or group together.

This is indeed the case in the example shown in Fig. 3. On the left, we can see the tSNE map of the gene expression, where two discrete clusters have been highlighted: PCPG, which are neuroendocrine tumors of the adrenal medulla or paraganglia (pheochromocytoma and paraganglioma), and PAAD, which are pancreatic adenocarcinomas, derived from the exocrine pancreas. Interestingly, 6 PAAD did not cluster with PAAD, but grouped with PCPGs. Exploration of the National Cancer Institute GDC data public portal reveals that those 6 PAADs are not actually pancreatic adenocarcinomas but pancreatic neuroendocrine tumors (PanNETs). PAAD and PanNETs are two distinct entities that, although emerge from the same organ, differ in their biological and clinical properties. Whereas PAAD derive from the exocrine cells of the pancreas and have poor prognosis, PanNETs derive from the neuroendocrine pancreatic cells and tend to have better prognosis. Thus, PanNETs should not have been included in the PAAD study and, actually, this misclassification has been claimed by Peran *et al.* (2018) in a revised version of the TCGA data.

This example also illustrates that MP allows boost the cognitive process by the interactive and remarkably expeditious reconfiguration of the scatterplots combining gene and miRNA expression profiles. As shown in Fig. 3, increasing the encoding from the miRNA profile-based tSNE reveals that the 6 PanNET samples contained in the PCPG cluster, move away together from the PCPG cluster. Thus, by introducing the miRNA encoding in a continuous and fluid manner, a clear pattern emerges ("pop-up" effect) allowing the discovery. This is due to one of the Gestalt principles, the *principle of common fate*, where stimulus elements are likely to be perceived as a unit if they move together. In this case morphing analysis revealed that PanNETs and PCPGs share a similar gene-expression profile but can be differentiated by their miRNA profile. Collectively, these data alert that precautions should be taken when using publicly available databases, and deliver a fast and efficient tool to depurate big data and provide scientifically sounded hypothesis.

3.2 Selection of a cancer type for analysis

The next two case studies are depicted in Fig. 4. In both cases the procedure starts with the tSNE map of gene expression for all available cancers. The cancer type attribute is available (along with many other clinical variables) in the data matrix for which a circular encoding can be defined. By progressively adding weight to this circular encoding (increasing $t^{\mathrm{cir}}_{\mathrm{cancer}}$ the mixing weight $\lambda^{\mathrm{cir}}_{\mathrm{cancer}}$ is increased according to a *softmax* function Eq. (6)) the original tSNE layout of the samples is smoothly reorganized towards a circular view where different cancer types are clustered on regularly spaced positions of the circle. In this view, with $\lambda_{\mathrm{tSNE}} \ll \lambda^{\mathrm{cir}}_{\mathrm{cancer}}$, the user can focus on a particular cancer type. Since linear morphing has been used, this view preserves the mutual distances among samples in the tSNE, being like having multiple single-cancer tSNEs.

3.3 Case study 2: Overexpression in Serum Amyloid A1 as marker for fatal outcome in clear cell renal cancer

In this example we use the MP approach to analyze the effect of a protein Serum Amyloid A1 (SAA1) over clear cell renal cancer (CCRC – KIRC in TCGA database). According to the literature (Paret *et al.* (2010)), this protein is overexpressed in those tumors with a worst prognosis generally resulting in metastasis and fatal outcome for patients.

As shown in Fig. 4, the first step is to focus on the clear cell renal cancer samples, simply zooming over its position in the circular encoding showing the different cancer types. Later, we can add an horizontal encoding for the SAA1 gene expression. By smoothly increasing its weight λ_{SAA1}, an horizontally ordered layout of tumor items by SAA1 expression emerges (tumors with low expression levels of SAA1 to the left, and with high expression levels to the right). Adding up a new vertical encoding for tumor stage λ_{stage}, tumors of the 4 stages get progressively arranged into 4 horizontal layers, being the upper ones for late-stage tumors. Having the samples separated by cancer and tumor stage, and being the horizontal coordinate related to SAA1 expression, the user immediately observes that samples concentrate on the left side for stages I and II, however, for stage IV tumors, they concentrate on the right side, meaning higher levels of expression.

Taking advance to this "tailored" layout, the user can also select two groups of tumors, one for stage I and another for stage IV, for statistical confirmation of the previous visual observation. Results show that SAA1 expression levels for stage I tumors yield $\mu = -1.76$, $\sigma = \pm 3.48$, while stage IV tumors get $\mu = 1.12$, $\sigma = \pm 3.66$. A one-way ANOVA test ($N = 192$) yields an F-value$= 26.75$ and a significance level p-value < 0.001, which confirms that the previous difference is highly relevant.

3.4 Case study 3: Correlation of miR-210 and CA9 in tumors with hypoxia

This case study shows how MP allow a visual confirmation of the correlation between miR-210 (hsa-miR-210-3p) and CA9 expressions in tumors with hypoxia (McCormick *et al.* (2013)). First a visual separation of all cancer types can be obtained using a circular encoding by increasing $\lambda^{\mathrm{cir}}_{\mathrm{cancer}}$. Then, adding vertical and horizontal encodings for CA9 and miR-210 with smaller (user-tuned) mixture coefficients $\lambda^{\mathrm{hor}}_{\mathrm{mir210}}, \lambda^{\mathrm{ver}}_{\mathrm{CA9}}$, naturally results in local xy scatterplots of miR-210 vs CA9 for each cancer.

The resulting layout can be seen in Fig 4 (c). Immediately, the user can spot emerging visual correlations on a per-cancer basis. Particularly, it can be observed that there is a negligible correlation for stomach cancer, while a visually appreciable correlation can be seen for CCRC as well as for papillary renal cancer. The layout easily allows the user to make individual selections for each cancer type. Numerical computation of correlations on the three selections confirm the observations being $r_{\mathrm{stomach}} = 0.13\,(p = 0.011)$, $r_{\mathrm{CCRC}} = 0.8\,(p < 0.001)$ and $r_{\mathrm{papillaryRC}} = 0.63\,(p < 0.001)$, the last two ones showing a weak but relevant correlation, consistent with McCormick *et al.* (2013).

4 Discussion

A key aspect of the MP approach relies on combining it with interaction mechanisms. In (Yi *et al.* (2007)) the authors describe seven categories of interaction (*select, explore, reconfigure, encode, abstract/elaborate, filter, connect*) that give a comprehensive view of interaction mechanisms used in data visualization. Most of these categories are embraced by the MP approach. For a fixed set of values of the interaction parameters t_i, the resulting scatterplot allows the user to perform a broad spectrum of these interaction operations. Thus, the user can *select* a subset of the points (for instance, a group of samples sharing a clinical condition, like having metastasis, belonging to a given population group, or having a certain range of expression values for a given gene or miRNA); these *interesting* points can be highlighted over the non-selected points, and this selection may be kept on upcoming rearrangements of the scatterplot, revealing the roles of the selected samples in other views that may be based on completely different contexts. Also, using scatterplots as the basis of representation, the approach admits interaction mechanisms falling in the *abstract/elaborate* category of Yi, such as *zoom* operations and *tool-tips* showing detailed information, as well as *explore* mechanisms like *panning*.

However, one of its most relevant and distinctive features over other approaches relies on allowing highly *reconfigurable* scatterplots. By manipulating the interaction parameters $t_1, t_2, \ldots, t_q$, from which a set of mixing coefficients $\lambda_1, \ldots, \lambda_p$ are computed using the mixing function $\lambda(\mathbf{t})$, the user modifies in a smooth manner the current scatterplot composed of a weighted combination of meaningful spatial encodings (e.g. tSNE-based 2D maps of samples, circular or linear encodings describing particular gene expressions, cancer types, gender or races, etc., that act as *basis views* or layouts), producing animated transitions that provide an immediate feedback. User actions on the t_i during this operation become, in this way, closely coupled to what the user sees, thereby producing a powerful *virtuous* cycle in the analytics process that fully engages visuomotor mechanisms of the cognitive process.

Inspired in the generic model on visualization proposed by Van Wijk in (Van Wijk (2005)), Fig. 5 describes the workflow of the MP approach described above. The amount of knowledge K gained by the user depends on the visual information (the current scatterplot) fed to the user through

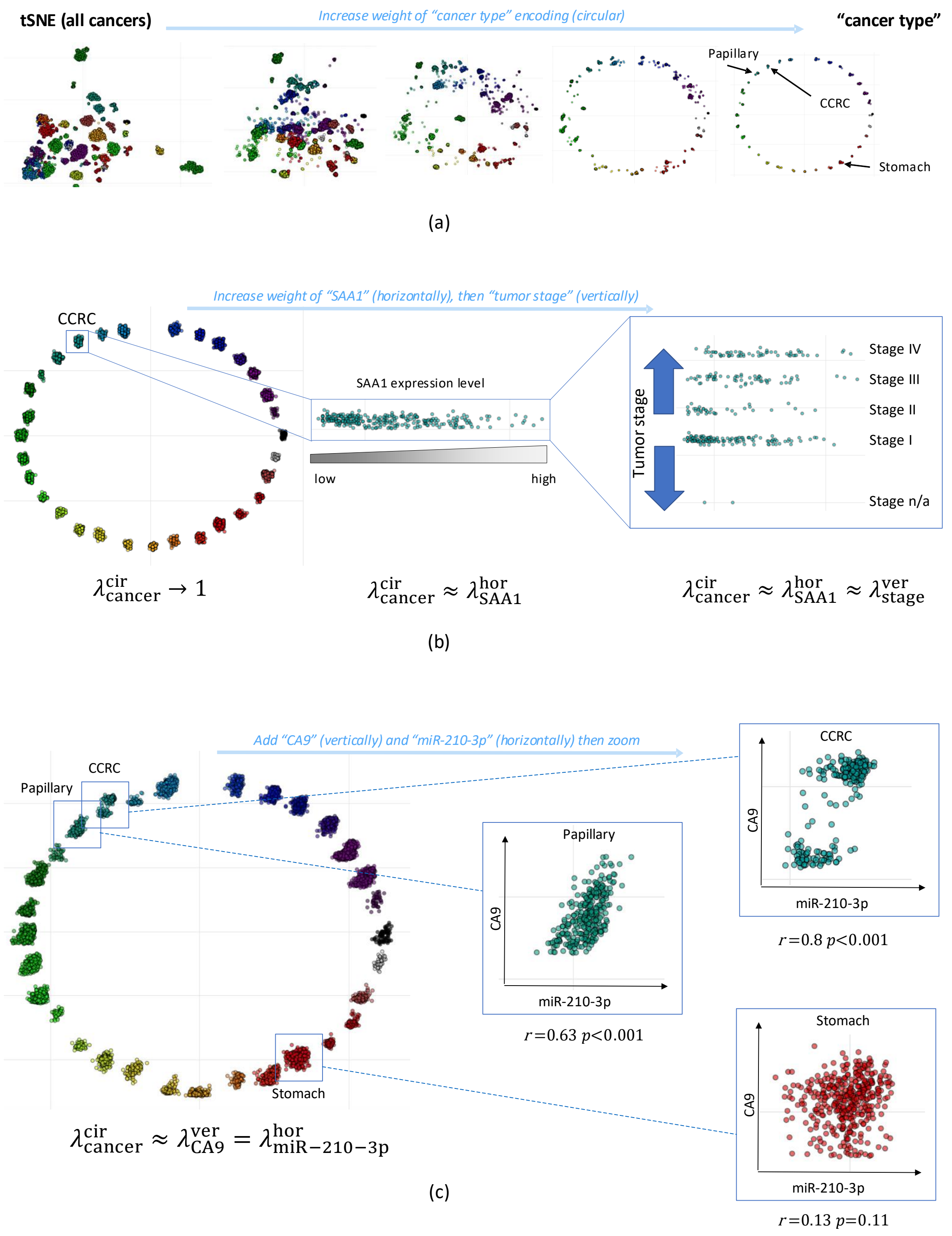


Fig. 4. Case studies using morphing approach for discovery of relevant facts. In a) the user selects the target cancer by increasing the weight of cancer type; case 1 is described in b), showing the discovery of overexpressed serum amyloid A1 gene (SAA1) in late stages of clear cell renal cancer (CCRC / KIRC), first selecting the target cancer and then increasing the weights of vertical tumor stage encoding; case 2 is shown in c) showing the visual discovery of relevant correlations between hypoxia related expressions, by increasing the weights of vertical and horizontal encodings of miR-210 and CA9 levels.

the perception system P, and the current user's knowledge. Also, based on current knowledge K, the user modifies the view through interactive exploration, E, that include the classical interaction mechanisms zoom, selection, pan, etc., and reconfiguration of the current view by means of the interaction parameters t_i according to the morphing operation.

4.1 Computational limitations and suitability for big data

The workload of the numerical computation, based on simple math operations, is negligible with respect to graphical burden. The computational efficiency of our approach is mainly limited by the number of items being displayed in the graphical display and the need to comply with a reasonable latency to support fluid animated transitions. In our experiments a web-based interface using canvas to display elements (developed under Python/Bokeh, with javascript callbacks) is able to display orders of 10^4 items (e.g. tumour samples, etc.) in a fluid manner. However, using desktop interfaces and advanced GPU-based graphical libraries could improve this substantially.

It must be pointed out, however, that the morphing projections framework admits aggregation operations, as described in Eq. (3), making it suitable for big data applications. Applications could be developed using aggregations that compact many items into a single group (binned aggregation), and Abstract/Elaborate (aggregate/disaggregate) interactions to switch between detail or big picture, in this way the number of displayed elements may be several orders of magnitude smaller than the overall analyzed items, allowing to tackle big data problems.

5 Conclusions

As biomedical data grows, researchers in the field are in need of more and more powerful techniques to manipulate this overwhelming amount of data. While there has been an enormous boost in machine learning and visualization techniques, these usually offer static views of this data, are sometimes complex to use, and do not take full advantage of the expert's knowledge in the analysis process.

Adding a smooth real-time interaction that allows the expert to re-arrange the visualized data according to different criteria is a powerful approach to both improve the analysis process, and the expert's knowledge, and it is a very efficient way to lay down relevant questions and establish new research hypotheses.

In this work we have presented the use of a technique called *morphing projections*, which relies on a set of basis views (typically 2D scatterplots) consisting of spatial encodings that lay out the samples according to meaningful attributes, or by sample similarities, that are blended by the user along the analysis process into a live animated view that is a combination of the basis views in different user-defined proportions. Morphing projections combines high doses of interaction with a highly visual approach, allowing the user to reconfigure the views "on the fly" to focus on demand into different sets of attributes of interest. Its capabilities were illustrated with two case studies in the field of genomics that reproduced some findings in the literature.

The main strength of this technique relies in that it combines in a seamless way one or more state-of-the-art techniques for data visualization of gene expression like the t-SNE (or any other dimensionality reduction techniques) and other 2D scatterplot visualizations, making it able for the user to compose in seconds, for instance, views with local tSNE plots arranged by cancer type, race, gender, disease stage or any other criteria, to do complex selections at any moment on the current view for posterior analysis, or to track the trajectories of the selected samples for user-driven changes in the layout according to different criteria.

Indeed, for instance, in case study 1, we learnt that PanNETs differ from PCPGs and PAADs in their miRNA expression profile. This is a novel, not previously reported finding. Any future investigations aimed at uncovering disease-related mechanisms, clinically useful biomarkers, or therapeutic targets, should be aware of these common and specific molecular traits of PAADs, PanNETS and PCPGs.

It must be pointed out, however, that the main role of MP is to facilitate the exploration of large biomedical datasets (providing a fast, friendly and interactive visual analysis of the data), not to generate medical results or evidences. The use of the tool allows to raise hypotheses that could lead to the generation of new knowledge, but this will always depend on the quality of the data and the good judgment of the user along the process. In this regard, we have shown, in the section on results, the exploratory process (use cases) that would be followed by a user whose decision making were based on expert medical criteria. Also, as part of a fair methodological use of a visual approach, the conclusions obtained must be accompanied by the assumptions and hypotheses used to generate them to avoid common pitfalls (such as cherry picking or data dredging); it should also be kept in mind that MP is an early stage in a longer process, where the observations should not be taken as conclusive and must be subsequently validated by other means.

We have also presented a rigorous formulation of the morphing projections idea in two ways. First, including a mathematical framework of the morphing operations with a close connection to relational algebra, that lays the basis for future generalizations of the approach, and, in addition, allows to pose the method operations in terms of primitives available in numerous software libraries and tools of the field, facilitating the development of new exporatory tools. And second, presenting the connections of our approach to the factors that improve the user's cognitive process for data analysis, highlighting the user-interaction and data visualization features of the approach.

Supplementary material

The following supplementary material to this paper has been provided: a) videos for the case studies described in the paper; b) a small demo web app with the ability to reproduce the reported cases; c) the details and code of the whole procedure described in section 2.1.

References

Akcakaya, P., Ekelund, S., Kolosenko, I., Caramuta, S., Özata, D. M., Xie, H., Lindforss, U., Olivecrona, H., and Lui, W.-O. (2011). mir-185 and mir-133b deregulation is associated with overall survival and metastasis in colorectal cancer. *International journal of oncology*, **39**(2), 311–318.

Cheerla, N. and Gevaert, O. (2017). Microrna based pan-cancer diagnosis and treatment recommendation. *BMC bioinformatics*, **18**(1), 32.

Dai, X., Ding, L., Liu, H., Xu, Z., Jiang, H., Handelman, S. K., and Bai, Y. (2019). Identifying interaction clusters for mirna and mrna pairs in tcga network. *Genes*, **10**(9), 702.

Di Leva, G. and Croce, C. M. (2013). mirna profiling of cancer. *Current opinion in genetics & development*, **23**(1), 3–11.

Diaz-Blanco, I., Dominguez-Gonzalez, M., Cuadrado-Vega, A., Diez-Gonzalez, A., and Fuertes-Martinez, J. (2012). Morphingprojections: Interactive visualization of electric power demand time series.

Ding, J., Hagood, J. S., Ambalavanan, N., Kaminski, N., and Bar-Joseph, Z. (2018). idrem: Interactive visualization of dynamic regulatory networks. *PLoS computational biology*, **14**(3), e1006019.

Egorov, A. A., Sakharova, E. A., Anisimova, A. S., Dmitriev, S. E., Gladyshev, V. N., and Kulakovskiy, I. V. (2019). svist4get: a simple visualization tool for genomic tracks from sequencing experiments. *BMC bioinformatics*, **20**(1), 113.

Endert, A., Ribarsky, W., Turkay, C., Wong, B. W., Nabney, I., Blanco, I. D., and Rossi, F. (2017). The state of the art in integrating machine learning into visual analytics. In *Computer Graphics Forum*, volume 36, pages 458–486. Wiley Online Library.

Eyking, A., Reis, H., Frank, M., Gerken, G., Schmid, K. W., and Cario, E. (2016). Mir-205 and mir-373 are associated with aggressive human mucinous colorectal cancer. *PloS one*, **11**(6), e0156871.

Heer, J. and Robertson, G. (2007). Animated transitions in statistical data graphics. *Visualization and Computer Graphics, IEEE Transactions on*, **13**(6), 1240–1247.

Hu, Y., Dingerdissen, H., Gupta, S., Kahsay, R., Shanker, V., Wan, Q., Yan, C., and Mazumder, R. (2018). Identification of key differentially expressed micrornas in cancer patients through pan-cancer analysis. *Computers in biology and medicine*, **103**, 183–197.

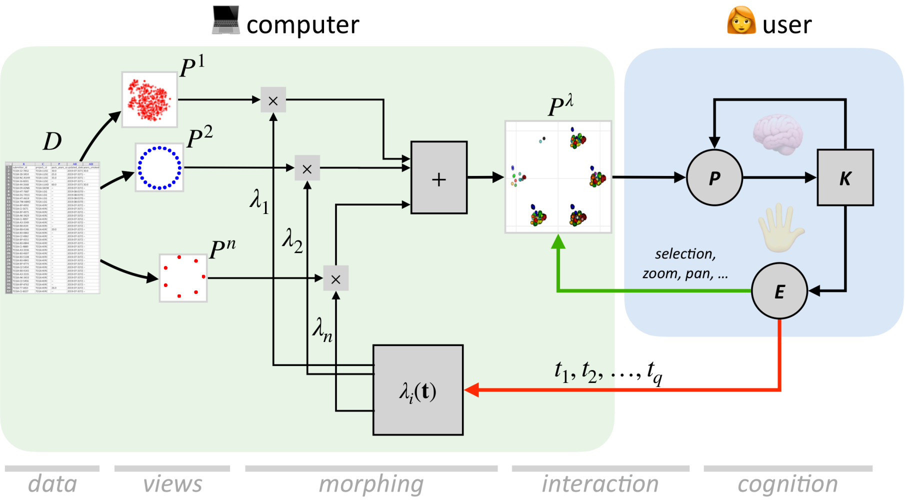


Fig. 5. Knowledge generation model of the morphing projections approach. A set of basis views P^i, generated from data D, are blended according to weights λ_i dynamically tuned by the user using interaction parameters, t_i. The resulting view, P^λ, enters the perception system P, increasing knowledge K, which changes the way the user perceives, and suggests new exploration actions, E, through interaction mechanisms.

Huisman, S. M., van Lew, B., Mahfouz, A., Pezzotti, N., Höllt, T., Michielsen, L., Vilanova, A., Reinders, M. J., and Lelieveldt, B. P. (2017). Brainscope: interactive visual exploration of the spatial and temporal human brain transcriptome. *Nucleic acids research*, **45**(10), e83–e83.

Jiang, D., Tang, C., and Zhang, A. (2004). Cluster analysis for gene expression data: a survey. *IEEE Transactions on knowledge and data engineering*, **16**(11), 1370–1386.

Jones, K., Nourse, J. P., Keane, C., Bhatnagar, A., and Gandhi, M. K. (2014). Plasma microrna are disease response biomarkers in classical hodgkin lymphoma. *Clinical cancer research*, **20**(1), 253–264.

Kamal, N., Wiebe, S., Engbers, J., and Hill, M. (2014). Big data and visual analytics in health and medicine: From pipe dream to reality. *J Health Med Informat*, **5**, e125.

Kobak, D. and Berens, P. (2019). The art of using t-sne for single-cell transcriptomics. *Nature communications*, **10**(1), 1–14.

McCormick, R., Blick, C., Ragoussis, J., Schoedel, J., Mole, D., Young, A., Selby, P., Banks, R., and Harris, A. (2013). mir-210 is a target of hypoxia-inducible factors 1 and 2 in renal cancer, regulates iscu and correlates with good prognosis. *British journal of cancer*, **108**(5), 1133–1142.

Navas-Delgado, I., García-Nieto, J., López-Camacho, E., Rybinski, M., Lavado, R., Berciano Guerrero, M. Á., and Aldana-Montes, J. F. (2019). VIGLA-M: Visual gene expression data analytics. *BMC Bioinformatics*, **20**.

Nusrat, S., Harbig, T., and Gehlenborg, N. (2019). Tasks, techniques, and tools for genomic data visualization. In *Computer Graphics Forum*, volume 38, pages 781–805. Wiley Online Library.

O'Donoghue, S. I., Baldi, B. F., Clark, S. J., Darling, A. E., Hogan, J. M., Kaur, S., Maier-Hein, L., McCarthy, D. J., Moore, W. J., Stenau, E., *et al.* (2018). Visualization of biomedical data. *Annual Review of Biomedical Data Science*, **1**, 275–304.

Paret, C., Schön, Z., Szponar, A., and Kovacs, G. (2010). Inflammatory protein serum amyloid a1 marks a subset of conventional renal cell carcinomas with fatal outcome. *European urology*, **57**(5), 859–866.

Peran, I., Madhavan, S., Byers, S. W., and McCoy, M. D. (2018). Curation of the pancreatic ductal adenocarcinoma subset of the cancer genome atlas is essential for accurate conclusions about survival-related molecular mechanisms. *Clinical Cancer Research*, **24**(16), 3813–3819.

Perez-Llamas, C. and Lopez-Bigas, N. (2011). Gitools: Analysis and visualisation of genomic data using interactive heat-maps. *PLoS ONE*, **6**(5).

Pezzotti, N., Lelieveldt, B. P., van der Maaten, L., Höllt, T., Eisemann, E., and Vilanova, A. (2016). Approximated and user steerable tsne for progressive visual analytics. *IEEE transactions on visualization and computer graphics*, **23**(7), 1739–1752.

Roche, K. E., Weinstein, M., Dunwoodie, L. J., Poehlman, W. L., and Feltus, F. A. (2018). Sorting five human tumor types reveals specific biomarkers and background classification genes. *Scientific reports*, **8**(1), 1–12.

Tan, H., Huang, S., Zhang, Z., Qian, X., Sun, P., and Zhou, X. (2019). Pan-cancer analysis on microrna-associated gene activation. *EBioMedicine*, **43**, 82–97.

Tang, Z., Li, C., Kang, B., Gao, G., Li, C., and Zhang, Z. (2017). GEPIA: A web server for cancer and normal gene expression profiling and interactive analyses. *Nucleic Acids Research*, **45**(W1), W98–W102.

Telonis, A. G., Magee, R., Loher, P., Chervoneva, I., Londin, E., and Rigoutsos, I. (2017). Knowledge about the presence or absence of mirna isoforms (isomirs) can successfully discriminate amongst 32 tcga cancer types. *Nucleic acids research*, **45**(6), 2973–2985.

Trevino, V., Falciani, F., and Barrera-Saldaña, H. A. (2007). Dna microarrays: a powerful genomic tool for biomedical and clinical research. *Molecular Medicine*, **13**(9), 527–541.

Võsa, U., Vooder, T., Kolde, R., Fischer, K., Välk, K., Tõnisson, N., Roosipuu, R., Vilo, J., Metspalu, A., and Annilo, T. (2011). Identification of mir-374a as a prognostic marker for survival in patients with early-stage nonsmall cell lung cancer. *Genes, Chromosomes and Cancer*, **50**(10), 812–822.

Van Der Maaten, L. and Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, **9**(November 2008), 2579–2625.

Van Wijk, J. (2005). The value of visualization. In *16th IEEE Visualization 2005 (VIS 2005)*. IEEE Computer Society.

Weinreb, C., Wolock, S., and Klein, A. M. (2018). SPRING: A kinetic interface for visualizing high dimensional single-cell expression data. *Bioinformatics*, **34**(7), 1246–1248.

Xia, J., Gill, E. E., and Hancock, R. E. (2015). NetworkAnalyst for statistical, visual and network-based meta-analysis of gene expression data. *Nature Protocols*, **10**(6), 823–844.

Xu, G., Zhang, Y., Wei, J., Jia, W., Ge, Z., Zhang, Z., and Liu, X. (2013). Microrna-21 promotes hepatocellular carcinoma hepg2 cell proliferation through repression of mitogen-activated protein kinase-kinase 3. *BMC cancer*, **13**(1), 469.

Yi, J. S., ah Kang, Y., and Stasko, J. (2007). Toward a deeper understanding of the role of interaction in information visualization. *IEEE transactions on visualization and computer graphics*, **13**(6), 1224–1231.

Zhou, G., Soufan, O., Ewald, J., Hancock, R. E., Basu, N., and Xia, J. (2019). NetworkAnalyst 3.0: A visual analytics platform for comprehensive gene expression profiling and meta-analysis. *Nucleic Acids Research*, **47**(W1), W234–W241.