Data and text mining

Spathial: an R package for the evolutionary analysis of biological data

Erika Gardini [1,2,*], Federico M. Giorgi [2], Sergio Decherchi [1,*], and Andrea Cavalli [1,2]

[1]Computational and Chemical Biology, Italian Institute of Technology, via Morego 30, 16163 Genoa, Italy;
[2]Department of Pharmacy and Biotechnology, University of Bologna, via Belmeloro 6, 40126 Bologna, Italy.

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Summary: A primary problem in high-throughput genomics experiments is finding the most important genes involved in biological processes (e.g. tumor progression). In this applications note, we introduce *spathial*, an R package for navigating high-dimensional data spaces. *spathial* implements the Principal Path algorithm, which is a topological method for locally navigating on the data manifold. The package, together with the core algorithm, provides several high-level functions for interpreting the results. One of the analyses we propose is the extraction of the genes that are mainly involved in the progress from one state to another. We show a possible application in the context of tumor progression using RNA-Seq and single-cell datasets, and we compare our results with two commonly used algorithms, *edgeR* and *monocle3*, respectively.

Availability and implementation: The R package *spathial* is available on the Comprehensive R Archive Network (https://cran.r-project.org/web/packages/spathial/index.html) and on GitHub (https://github.com/erikagardini/spathial). It is distributed under the GNU General Public Licence (version 3).

Contact: erika.gardini@iit.it, sergio.decherchi@iit.it

Supplementary Information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The recent advent of next-generation sequencing (NGS) techniques has produced a massive amount of high-throughput data for quantitive biology, especially in the field of transcriptomics. The increased depth and sample size of transcript measurement has challenged scientists to create novel algorithms that can use these highly complex datasets to increase our understanding of biological phenomena (Camacho *et al.* (2018)).

Examples of these complex transcriptomic datasets are those generated by the TCGA (Tomczak *et al.* (2015)) and GTEX (Lonsdale *et al.* (2013)) consortia, which collect tens of thousands of human RNA-Seq samples from tumor and physiological tissues, respectively. In the past few years, the development of single-cell sequencing technologies has further increased the sample size of RNA-Seq datasets, albeit at a cost for transcript coverage (Svensson *et al.* (2018)).

RNA-Seq analyses to understand changes in gene expression have built on the previous generation of technological platforms (microarrays), and have focused on characterizing the quantitative differences between two or more groups of samples, a process known as differential gene expression analysis (DGEA).

As the sample sizes increase, so does our ability to detect and study the natural heterogeneity of living systems, whether they are bulk tissues (e.g. interpatient variance in cancer) or single cells (e.g. different cells and cell states in a microenvironment). Moreover, large datasets allow us to measure biological transition processes such as cell differentiation and tumor progression (Pastushenko and Blanpain (2018)). The literature contains some studies on defining cell trajectories (Qiu, X. *et al.* (2017)); however, to our knowledge, there is no general and flexible

algorithm/package for modeling continuous processes and extracting the associated features (i.e. genes or transcripts).

Recently, Ferrarotti *et al.* (2018) designed an algorithm to identify smooth and *energetically meaningful* paths in data space (Ferrarotti *et al.* (2018); Ragusa *et al.* (2019)). This algorithm, the Principal Path algorithm, was inspired by the minimum free-energy path concept in statistical mechanics (Maragliano *et al.* (2006)). It allows the user to navigate and analyze vector spaces morphing from a start point to an end point. The waypoints along the path can be imagined as a chain of springs, with each being a small variation of the previous one. They are therefore particularly interesting from the evolutionary point of view.

Unlike shortest path algorithms (e.g. the Dijkstra shortest path (Dijkstra (1959))), the Principal Path takes into account the concept of smoothness, which can deliver solutions that are much more *cognitively sound* (Ferrarotti *et al.* (2018)). The model can also be considered generative (even if a distribution is not explicitly derived) because the waypoints are interpolated over the data manifold.

Here, we present a readily usable R package, dubbed *spathial*, which implements the Principal Path algorithm to analyze progressions in large-scale transcriptomic datasets, such as those arising from bulk and single-cell RNA-Seq.

2 The package

Here, we introduce a novel R package, *spathial*, which implements the Principal Path algorithm for the analysis of multidimensional biological datasets.

The algorithm is based on the following minimization problem:

$$\min_{W,u} \sum_{i=1}^{N} \sum_{j=1}^{N_c} \|\phi(x_i) - \omega_j\|^2 \delta(u_i, j) + s \sum_{i=0}^{N_c} \|\omega_{i+1} - \omega_i\|^2$$

where N is the number of samples, Nc is the number of waypoints, $\phi(\cdot) : \mathbb{R}^{\mathrm{d}} \rightarrow \mathbb{R}^{\mathrm{d}'}$ is the (possibly nonlinear) transformation mapping of the d-dimensional input space, x_i is a sample of the $N \times d$ matrix X arranged in a row-wise fashion, ωj is a waypoint of the $N_c \times d'$ matrix W arranged in row-wise fashion, and $\delta(ui, j)$ is the Kronecker delta where u_i are cluster memberships.

This functional is an extension of k-means clustering, where the first and last clusters are fixed, while the other clusters are evolved according to the functional, which induces a curve topology due to the regularization term. All the clusters are waypoints for the path and are topologically connected by a chain of springs (Ferrarotti *et al.* (2018)). The s hyperparameter regulates the trade-off between data-fitting and smoothness of the inferred path. The selection of s is critical. The supplementary materials provide details on how s is selected in this package together with the discussion regarding the computational complexity of the algorithm.

Conceptually, the algorithm allows one to infer a relevant transition or evolutionary path that can highlight the features involved in a specific process. It can thus be useful in all the scenarios where the temporal (or pseudo-temporal) evolution is the main problem (e.g. tumor progression, cell cycle analysis). The input of the algorithm (together with the full data matrix) comprises two points, which represent the boundary conditions of the algorithm: the start point and the end point. Given the boundaries, the algorithm learns a smooth transition path connecting them. Along the path, there are new intermediate data samples which gradually morph from the start point to the end point. In this way, it is possible to move from two known states and analyze which features are involved in the transition between the two states.

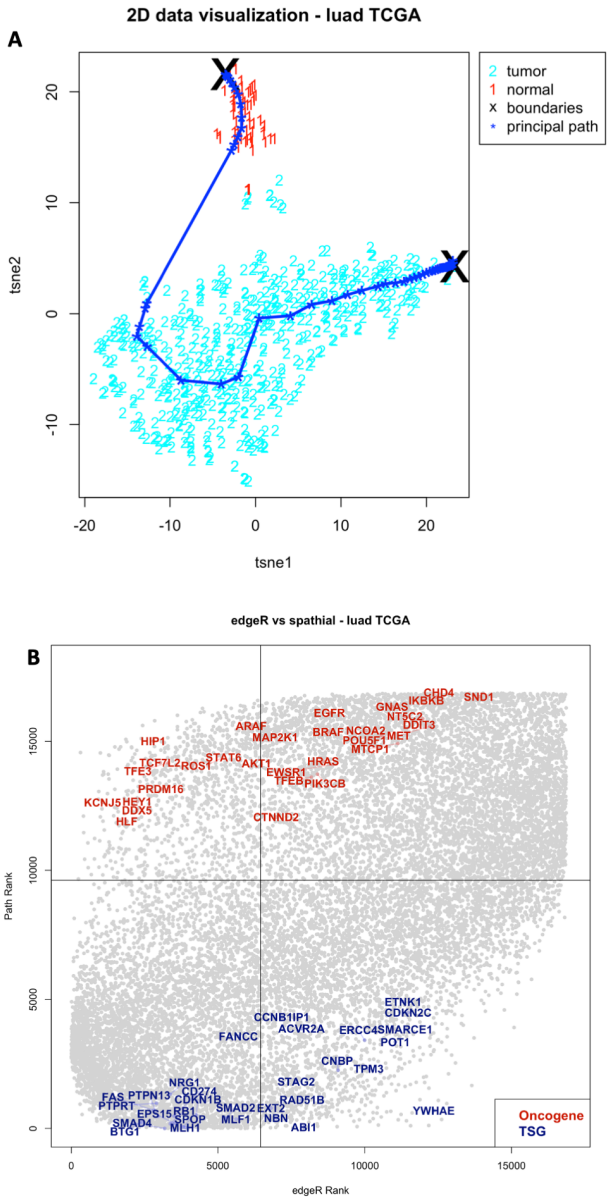

Fig. 1. Results on the TCGA Lung Adenocarcinoma dataset. (A) 2D visualization of the Principal Path together with the data points. The x and the y coordinates are the output of the dimensionality reduction performed with tSNE (Van der Maaten, L. et al. (2008)). The start point and the end point of the Principal Path are the most distant points from the centroid of the tumor samples and the centroid of the normal samples, respectively. The Principal Path is composed of 50 intermediate points (waypoints) plus the boundaries (B) Comparison of *spathial* and *edgeR*. The x and y coordinates are the position in the rank for *edgeR* and *spathial*, respectively. The ranks are the statistical significance $(log10(pvalue)sign(FoldChange))$ for *edgeR* and the Pearson's correlation score for *spathial*. Colored genes are the 60 oncogenes and tumor suppressor genes (from the Cancer Gene Census list (v86 - Sondka, Z. et al. (2018)) for which *spathial* disagreed the most strongly with *edgeR* and for which the *spathial* rank was better than *edgeR*.

The package *spathial* offers the option of running the Principal Path algorithm using very high-level functions. It subdivides the workflow for constructing the path into a few steps:

- selection of the boundaries (start and end points). *spathial* provides three different options: a visual selection by the user, classes centroids, or selection of the samples using their row-name. However, users can choose their own strategy, extract the row-names of the boundaries and set them using the third mode;
- prefiltering (optional): allows one to obtain a local solution, which does not involve the entire dataset. This procedure removes some data points and forces the Principal Path algorithm to go through a restricted number of samples. This can create smoother paths but at the same time can prune some available data;
- execution of the Principal Path algorithm with the boundaries selected during the first step and with the input data (filtered or not filtered).

After the Principal Path algorithm is run, users obtain the coordinate of the waypoints (new interpolating samples). *spathial* provides some utility functions for the analysis of the output. In particular, users can compute the labels of the waypoints (assigned as the label of the nearest point). Additionally, they can plot the 2D visual representation of the datapoints together with the path waypoints. This utility function takes as input the data points and the waypoints of the path. If those are in 2D, the function directly plots them. If not, it performs a dimensionality reduction using tSNE (Van der Maaten, L. *et al.* (2008)) and then plots the points. However, users can adopt their preferred dimensionality reduction strategy and give 2D coordinates to the function. Finally, *spathial* allows the user to compute some statistical information about the waypoints. In particular, it allows one to obtain the Pearson's correlation of the waypoint features with the path progression. Path progression is here defined as the ordered sequence of waypoint indices from 0 (the start point) to N_c+1 (the end point). In this way, users can obtain the features that are correlated with the progression (features involved in the transition between the start point and the end point), and they can perform a feature selection according to the Pearson's correlation scores. The function also provides the associated p-value and q-value.

While the package can be applied to any data, its current focus is the analysis of transcriptomic data, which include some of the largest (in terms of number of features and samples) and most commonly generated datasets.

3 Results

We performed several experiments to compare *spathial* with existing tools and to demonstrate its flexibility.

First, we experimented on the TCGA lung adenocarcinoma RNA-Seq dataset (The Cancer Genome Atlas Research Network1 *et al.* (2013)), comprising 562 gene expression profiles RPM-normalized (19637 genes each). Each sample is labelled as "tumor" or "normal" according to the TCGA barcode. The aim of the experiment is to navigate the space from the normal samples to the tumor samples. In this case, the start point was the most distant normal sample from the tumor centroid and the end point was the most distant tumor sample from the normal centroid. We selected these start and end points because we were searching for the extremes, conceptually the *most normal* sample and the *most diseased* sample. We considered as ground truth the oncogenes and tumor suppressor genes (tsg) listed in the Cancer Gene Census (Sondka, Z. *et al.* (2018)). The prefiltering was not executed since the search is for a global solution. Finally, the Principal Path algorithm was run with 50 intermediate points (waypoints) plus the boundaries. Fig. 1A shows the samples (colored according to the labels) and the path.

We compared the first 1000 best q-value ranked genes for *spathial* with the relevant genes extracted by a commonly used tool for DGEA, *edgeR* (Robinson, M.D. *et al.* (2010)) (again the first 1000 best q-values genes). Tables in supplementary materials (ST 1 and ST 2) show the details for this comparison. Some genes that *spathial* identified as being involved in the progression were not identified as such by *edgeR* (and vice versa).

To further highlight the genes found by *spathial* and missed by *edgeR* (see Fig. 1B), we selected the most correlated genes for *spathial* using the quantiles and setting two thresholds such that 70% of values fell below the first threshold and 30% fell above the second threshold, representing the most positively and negatively correlated genes. From among the positively correlated genes for *spathial*, we selected the oncogenes and compared them with the *edgeR* results. In particular, we analyzed how these oncogenes are placed by the *edgeR* and *spathial* ranking respectively according to their statistical significance (computed as $-log10(pvalue) * sign(FoldChange)$) and their Pearson's correlation scores. Finally, we selected the first 30 genes for which *spathial* disagreed the most strongly with *edgeR* and for which the *spathial* rank was better than *edgeR*. The same comparison can be performed by selecting the tumor suppressor genes (tsg) from among the most negatively correlated genes for *spathial*. Fig. 1B shows the subset of 60 genes selected as described above. The x and y coordinates are the position in the rank for *edgeR* and *spathial*, respectively. Oncogenes and tumor suppressor genes (tsg) should be placed at the end and the beginning of the rank, respectively (the rank is with sign and ascending), because they should be highly positively and highly negatively correlated in the transition from normal to tumor. Therefore, the red genes on the left (oncogenes) and the blue genes (tsg) on the right are those that *spathial* (but not *edgeR*) identified as involved in the transition from normal to tumor.

Other experiments on the TCGA liver hepatocellular carcinoma and the breast invasive carcinoma RNA-Seq datasets (The Cancer Genome Atlas Research Network1 *et al.* (2013)) are shown in the supplementary materials. Figures SF 3 and SF 4 and tables ST 3, ST 4, ST 5, ST 6 show the resulting path and the comparison with *edgeR*.

We performed a second experiment with *spathial* on a single-cell RNA-Seq dataset. In this case, we selected the dataset used in the experiments in Karlsson *et al.* (2017). This dataset comprises 96 human myxoid liposarcoma cells, each described with a gene expression profile (23928 genes each). Cells are labelled as "G1", "S", "G2/M" according to their experimentally determined cell cycle phase. The aim of the experiment is to navigate the space from the "G1" samples to the "G2/M" samples. The start point was the "G1" centroid and the end point was the "G2/M" centroid. There was no prefiltering because the search was for a global solution. Finally, the Principal Path algorithm was run with 50 intermediate points (waypoints) and the boundaries. Figure SF 5 shows the samples (colored according to the labels) and the path.

We computed the q-value for each gene, then selected the genes with high statistical significance (the first 1000 best q-value ranked genes). Then, we compared them with the statistical information extracted with *monocle3*, a package for computing single-cell trajectory analysis (Qiu, X. *et al.* (2017)). In particular, one can use *monocle3* to learn the graph and find genes that are differentially expressed across a single-cell trajectory computing the MoranâĽ™s I test. Here too, we selected the first 1000 best q-value ranked genes (this q-value is computed on the Moran's I scores and adjusted according to the Benjamini-Hochberg method). We detected a significant overlap between *monocle3* and *spathial* gene predictions (see supplementary material, tables ST 7 and ST 8). However, some genes identified by *spathial* as being involved in the progression were not identified as such by *monocle3* (and vice versa). Some of those genes belong to "Group2" and "Group3" of the Karlsson *et al.* (2017) experiment and are known to respectively decrease and increase in expression from G1 toward mitosis; the decrease/increase information is used as ground truth.

The supplementary material contains all the detailed results, tables, and

figures, all the datasets, and the scripts for reproducing the experiments and figures.

4 Conclusions

We have developed *spathial*, an R implementation of the Principal Path algorithm. *spathial* can be readily used to identify progression paths, either temporal or pseudo-temporal, in data space. We applied this algorithm to transcriptomic and single-cell RNA-Seq datasets because these applications demonstrate its flexibility in coping with different problems. However, the package can be applied to any omics. Results show that the tool is able to retrieve information missed from other packages and vice versa.

The package is available on the Comprehensive R Archive Network (https://cran.r-project.org/web/packages/spathial/index.html) and on GitHub (https://github.com/erikagardini/spathial).

Acknowledgements

We thank Grace Fox for copyediting and proofreading.

References

Camacho, D. M., Collins, K. M., Powers, R. K., Costello, J. C., and Collins, J. J. (2018). Next-generation machine learning for biological networks. *Cell*, **173**(7), 1581–1592.

Dijkstra, E. W. (1959). A note on two problems in connexion with graphs. *Numerische mathematik*, **1**(1), 269–271.

Ferrarotti, M. J., Rocchia, W., and Decherchi, S. (2018). Finding principal paths in data space. *IEEE transactions on neural networks and learning systems*.

Karlsson, J., Kroneis, T., Jonasson, E., Lekholm, E., and StÃ¥hlberg, A. (2017). Transcriptomic characterization of the human cell cycle in individual unsynchronized cells. *Journal of Molecular Biology*, **429**.

Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., Young, N., *et al.* (2013). The genotype-tissue expression (gtex) project. *Nature genetics*, **45**(6), 580.

Maragliano, L., Fischer, A., Vanden-Eijnden, E., and Ciccotti, G. (2006). String method in collective variables: Minimum free energy paths and isocommittor surfaces. *The Journal of chemical physics*, **125**(2), 024106.

Pastushenko, I. and Blanpain, C. (2018). Emt transition states during tumor progression and metastasis. *Trends in cell biology*.

Qiu, X., Mao, Q., Tang, Y., Wang, L.,Chawla, R., Pliner, H. A., and Trapnell, C. (2017). Reversed graph embedding resolves complex single-cell trajectories. *Nature Methods* **14**, 979âŁ“982. (2017)

Ragusa, E., Gastaldo, P., Zunino, R., Ferrarotti, M. J., Rocchia, W., and Decherchi, S. (2019). Cognitive insights into sentic spaces using Principal Paths. *Cognitive Computation*, pages 1–20.

Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* , **26**(1), 139-140.

Sondka, Z., Bamford, S.,Cole, C. G., Ward, S. A., Dunham, I., and Forbes, S. A. (2018). The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers *Nature Reviews Cancer*, **18**, 696âŁ“705.

Svensson, V., Vento-Tormo, R., and Teichmann, S. A. (2018). Exponential scaling of single-cell rna-seq in the past decade. *Nature protocols*, **13**(4), 599.

The Cancer Genome Atlas Research Network1, Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R., Ozenberger, B. A., Ellrott, K., Shmulevich, I., Sander, C., Stuart, J.M. (2013). The Cancer Genome Atlas Pan-Cancer analysis project. *Nature Genetics*, **45**, 1113âŁ“1120.

Tomczak, K., Czerwińska, P., and Wiznerowicz, M. (2015). The cancer genome atlas (tcga): an immeasurable source of knowledge. *Contemporary oncology*, **19**(1A), A68.

Van der Maaten, L., and Hinton, G.E. (2008). Visualizing data using t-sne. *Journal of Machine Learning Research* , **9**, 2579-2605.