

# Cancer Fingerprints by Topological Data Analysis



Ana Carpio

**Abstract** Topological data analysis has arisen as a promising tool to extract information on the structure of a wide variety of datasets. We analyze here its potential in two types of cancer studies. First, we compare time series of images from simulations of metastatic invasion in epithelial tissues. Calculating bottleneck distances of persistent diagrams we can characterize and classify the advancing interfaces of cellular aggregates. Second, we compare mRNA expression values for genes involved in cell cycles extracted from pancreas cancer tissue. We discuss how persistence information from different distances can provide insight on patient/gene clusters.

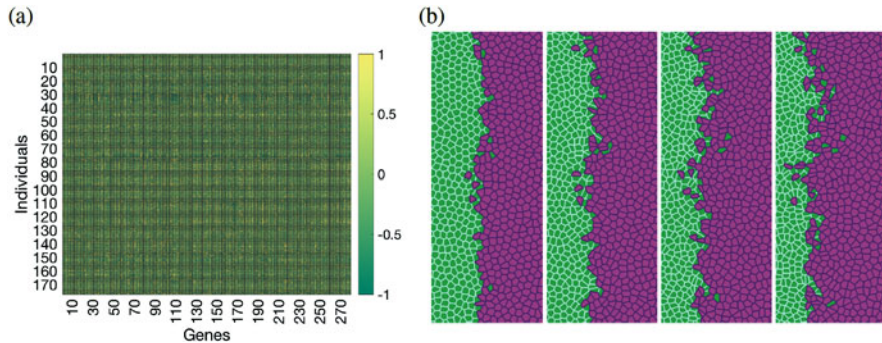
## 1 Introduction

Clinical and experimental studies of illness generate large amounts of data of a different nature. Consider cancer, for instance. Laboratory analyses of gene expression lead to large files containing measurements for different genes [15], see Fig. 1a. Instead, experimental observations of normal and malignant cells [9] yield time series of images, see Fig. 1b. Being able to extract meaningful information from large biomedical datasets, regardless of their nature, is a challenge that requires the development of adequate mathematical and computational tools.

Topological data analysis (TDA) furnishes a framework that provides dimensionality reduction and robustness to noise [2] when studying data clouds, with a certain independence with respect to the metrics selected. Recent studies have pointed out the potential of TDA in biological applications [8, 13, 16]. Biomedical data can often be seen as point clouds in a space of dimension  $D$ . Whereas for images  $D$  is the spatial dimension, for gene expression datasets  $D$  is the number of patients or genes in the study. We will see how to use TDA to extract information in

---

A. Carpio (✉)  
Universidad Complutense de Madrid, Madrid, Spain  
e-mail: [ana\\_carpio@mat.ucm.es](mailto:ana_carpio@mat.ucm.es)



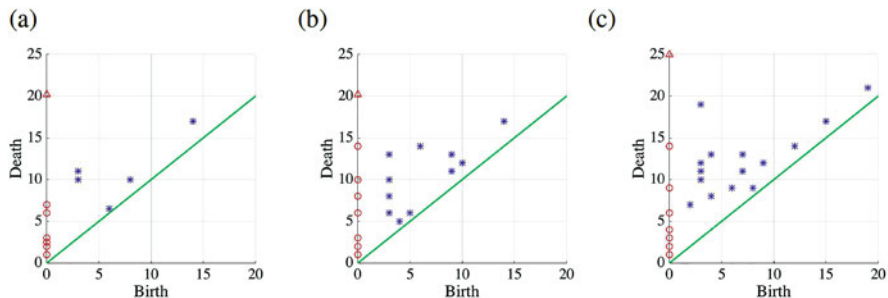
**Fig. 1** (a) Heatmap showing normalized mRNA expressions for a collection of genes within a set of patients, data taken from [6]. (b) Snapshots from a numerical simulation showing the invasion of healthy (green) epithelial tissue by malignant (magenta) cells, reprinted from [1], see [9] for related experimental images

both settings. The paper is organized as follows. Section 2 applies TDA to classify automatically interfaces between healthy and malignant cells in two dimensional images. Section 3 proposes a topology based hierarchical clustering procedure for gene expression data. Finally, Sect. 4 summarizes our conclusions.

## 2 Classification of Interfaces

Competition between two different media (fluids, for instance) or populations is an ubiquitous phenomenon in many fields. Usually, an interface separating the two components forms. Being able to automatically characterize such interface is important to identify patterns or stages in biological applications. Given several images representing the evolution of fragmented interfaces, our strategy proceeds in the following steps [1]:

1. Extract from each image a point cloud  $X$  defining the interface.
2. Build a Vietoris-Rips filtration  $V(X, r)$  for each point cloud based on the Euclidean distance, that is, a family of simplicial complexes formed joining by edges and triangles the points at a distance smaller than a variable parameter  $r$ , see [17].
3. Calculate the Betti numbers associated to each filtration:  $betti_0(r)$  (number of components) and  $betti_1(r)$  (number of holes) as the filtration parameter  $r$  varies.
4. For each identified component in each filtration, calculate the persistence intervals  $[r_b, r_d]$ , that is, the filtration parameter values at which it appears  $r_b$  (birth) and disappears  $r_d$  (death). They define the  $H_0$  homology.
5. For each identified hole in each filtration, calculate the persistence intervals  $[r_b, r_d]$ . They define the  $H_1$  homology.



**Fig. 2** Persistence diagrams representative of the initial, intermediate and late stages in the invasion process

6. Plot the persistence diagrams formed by the points  $(r_b, r_d)$  defining the persistence intervals for components and holes in each filtration, see Fig. 2.
7. Calculate the Bottleneck distance [11] between the  $H_1$  persistence diagrams.
8. Use k-means or a hierarchical clustering [10] approach to group the interfaces in clusters according to the level of detail required.

For the simulation considered in Fig. 1b, a set of 12 images is classified by K-means in 3 groups: the first three frames correspond to initial stages in which the interface is close to an unfragmented smooth curve, the last two frames correspond to late stages of the invasion period with many fragments and interpenetration, while the remaining frames correspond to an intermediate stage in which fragments may detach and reattach, see Fig. 2.

The study of images involves point clouds in two or three dimensional spaces. Medical records containing the values of several variables monitored over a collection of patients belong to higher dimensional spaces. Their study presents new difficulties.

### 3 Grouping Data

Gene studies in cancer patients have provided large amounts of information which may help to identify genetic features of sickness [15]. We consider here measurements of mRNA gene expression data for pancreas cancer available in [6], taken from the TCGA (the Cancer Genome Atlas) study. In this case, data take the form of numeric matrices  $M = (m_{j,i})$  containing values for a collection of genes  $i = 1, \dots, N$ , from tissue samples corresponding to different patients  $j = 1, \dots, J$ .

The first step consists in normalizing the data. To do so [7], we calculate the means  $\mu_i$  and standard deviations  $\sigma_i$  for each gene over the patients and compute the normalized values  $\tilde{m}_{j,i} = \frac{m_{j,i} - \mu_i}{3\sigma_i}$ . Then, we select a distance and a clustering

strategy to group either patients using information from genes, or genes using information from patients.

### 3.1 Distance Selection

To compare genes or patients, we can use a number of distances [5]:

- The *Euclidean distance* between two columns or rows  $m^1$  and  $m^2$  is their distance as vectors in a  $D$  dimensional space  $d(m^1, m^2) = \|m^1 - m^2\|_2$ .
- The *Earth Mover's distance* (EMD) provides the minimum cost of turning one column (resp. row) into the other [13]

$$emd(m^1, m^2) = \frac{\sum_{k=1}^D \sum_{\ell=1}^D c_{k,\ell} d_{k,\ell}}{\sum_{k=1}^D \sum_{\ell=1}^D d_{k,\ell}},$$

where  $d_{k,\ell} = |m_k^1 - m_\ell^2|$  is the ground distance, and  $c_{k,\ell}$  minimizes the cost  $\sum_{k=1}^D \sum_{\ell=1}^D c_{k,\ell} d_{k,\ell}$  subject to the constraints  $c_{k,\ell} \geq 0$ ,  $1 \leq k, \ell \leq D$ ,  $\sum_{k=1}^D \sum_{\ell=1}^D c_{k,\ell} = D$ ,  $\sum_{k=1}^D c_{k,\ell} \leq 1$ ,  $1 \leq \ell \leq D$ ,  $\sum_{\ell=1}^D c_{k,\ell} \leq 1$ ,  $1 \leq k \leq D$ . The EMD identifies patterns regardless of their location. The distance between two patient profiles that are equal except for a peak about different genes would be small, which is inadequate as different genes may define different illnesses.

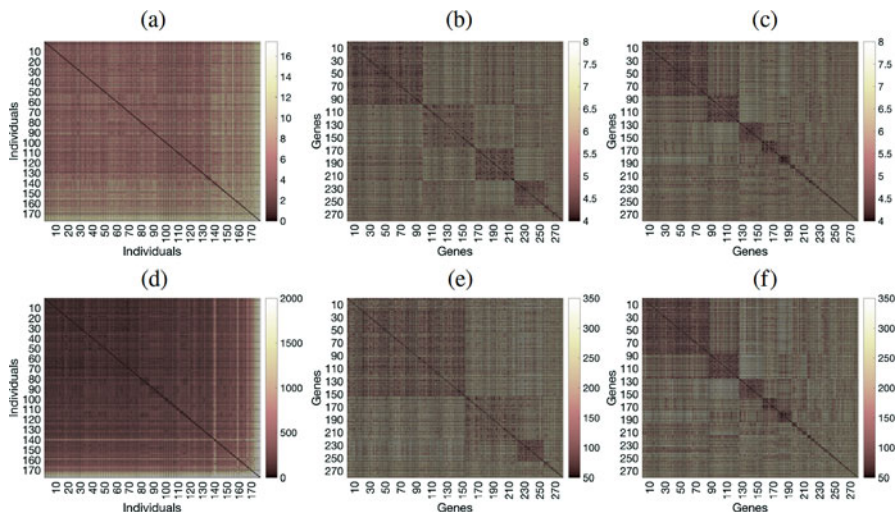
- Considering a set  $S$  of columns (resp. rows)  $m^1, m^2, \dots, m^L$ , the *Fermat  $\alpha$ -distance* between any two of them relative to that set is [14]

$$d_{S,\alpha}(m^1, m^2) = \min \left\{ \sum_{\ell=1}^{k-1} \|y^{\ell+1} - y^\ell\|_2^\alpha \mid (y_1, \dots, y_k) \text{ path from } m^1 \text{ to } m^2 \text{ in } S \right\},$$

for any  $\alpha > 1$ . When  $\alpha = 1$ , we recover the Euclidean distance. The Fermat distance compares items in a set weighting information from all the other items in the same set, which is interesting when we want to compare gene profiles weighting information from cohorts of patients [3].

### 3.2 Distance and Topology Based Clustering

Figure 3 represents gene-gene and patient-patient distances for different gene (resp. patient) orderings. Regardless of the ordering, we can use such distance matrices in hierarchical clustering algorithms [10] and select a natural number of clusters based on inconsistency criteria [12]. Grouping genes (resp. patients) by their clusters we obtain the panels in Fig. 3, which uncover hidden relations in the data.

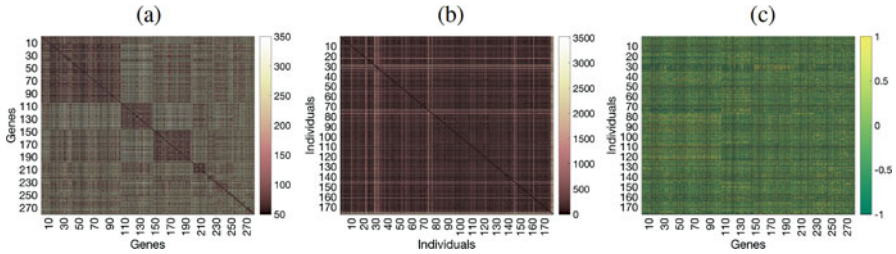


**Fig. 3** Heatmaps representing the distance matrices for the set of data considered in Fig. 1a ordering patients (resp. genes) by cluster groups, as determined by hierarchical clustering with different distances: **(a–c)** Euclidean distances, **(d–f)** Fermat distances with  $\alpha = 3$ . **(a)** and **(d)** compare patients, while the rest compare genes. Panels **(a–b)**, **(d–e)** use the natural number of clusters, as given by inconsistency studies. Instead, **(c)** and **(f)** use 36–37 clusters

Moreover, using any of these distances on the point cloud of patients  $m_{j,\cdot} = (m_{j,1}, \dots, m_{j,N})$ ,  $j = 1, \dots, J$ , or the point cloud of patients  $m_{\cdot,i} = (m_{1,i}, \dots, m_{N,i})$ ,  $i = 1, \dots, N$ , we can implement a similar procedure to that described in Sect. 2, only the distance changes. We construct a filtration, calculate the Betti numbers, as well as the persistence diagrams and intervals. With this information, we can compare datasets from different cancer types or patient studies to identify distinctive features and profiles. Moreover, the  $H_0$  homology provides an additional clustering strategy, different from usual hierarchical clustering. For a fixed filtration parameter value, each component of the simplex constructed for that filtration value defines a cluster. As the filtration parameter varies, we have a topology based hierarchical clustering strategy. Figure 4 displays the same data as Fig. 1a when genes and patients are rearranged following the components of filtrations for a fixed filtration value.

## 4 Conclusions

We have discussed the potential of persistence studies based on different distances combined with clustering strategies to extract information from point clouds of data of medical interest. Applied to time series of images of cellular arrangements, it provides a tool to automatically classify specific image features. Applied to gene



**Fig. 4** Fernet distance reordered following  $H_0$  clusters (a) for genes and (b) for patients. Panel (c) shows the data rearranged following the  $H_0$  clusters

expression data, it opens new perspectives to gain a better understanding of hidden relations. Similar techniques could be exploited to study clinical data from other illnesses, immune disorders for instance [4].

**Acknowledgments** Research supported by Spanish MICINN grants MTM 2017-84446-C2-1-R and PID2020-112796RB-C21.

## References

1. L.L. Bonilla, A. Carpio, C. Trenado, Tracking collective cell motion by topological data analysis, *PLoS Comput Biol* 16 (2020) e1008407.
2. G. Carlsson, Topology and data, *Bull. Amer. Math. Soc.* 46 (2009) 255–308.
3. A. Carpio, L.L. Bonilla, J.C. Mathews, A.R. Tannenbaum, Fingerprints of cancer by persistent homology, *bioRxiv* 777169, 2019
4. A. Carpio, A. Simón, L.F. Villa, Clustering methods and Bayesian inference for the analysis of the time evolution of immune disorders, *arXiv:2009.11531* 2020
5. A. Carpio, A. Simón, A. Torres, L.F. Villa, Pattern recognition in data as a diagnosis tool, *Journal of Mathematics in Industry* 12 (2022) 3.
6. E. Cerami, J. Gao, U. Dogrusoz et al, The cBio cancer genomics portal: An open platform for exploring multidimensional cancer genomics data, *Cancer Discov* 2 (2012) 401–404.
7. Y. Chen, F. D. Cruz, R. Sandhu, A. L. Kung, P. Mundi, et al, Poediatric sarcoma data forms a unique cluster measured via the Earth Mover’s Distance, *Sci. Rep.* 7 (2017) 7035.
8. M.R. McGuirl, A. Volkening, B. Sandstede, Topological data analysis of zebrafish patterns. *Proc. Nat. Acad. Sci.* 117 (2020) 5113–5124.
9. S. Moitrier, C. Blanch, S. Garcia, K. Sliogeryte et al., Collective stresses drive competition between monolayers of normal and Ras-transformed cells, *Soft Matter* 15 (2019) 537–545.
10. L. Kaufman, P.J. Rousseeuw, *Finding groups in data: An introduction to cluster analysis*, Hoboken: Wiley-Interscience, 1990.
11. M. Kerber, D. Morozov, A. Nigmatov, Geometry helps to compare persistence diagrams, *ACM J. Exp. Algorithmics*, 22 (2017) 1.4.
12. T. Kovacheva, A hierarchical clustering approach to find groups of objects, *Proceedings of the IV Congress of Mathematicians, Macedonia; 2008.* pp 359–373.
13. A.H. Rizvi, P.G. Camara, E.K. Kandror, T.J. Roberts et al., Single-cell topological RNA-seq analysis reveals insights into cellular differentiation and development, *Nat. Biotech.* 35 (2017) 551–560.

14. F. Sapienza, P. Groisman, M. Jonckheere, Weighted Geodesic Distance Following Fermat's Principle. Proc 6th International Conference on Learning Representations (ICLR), 2018.
15. The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium, Pan-cancer analysis of whole genomes, *Nature* 578 (2020) 82–93.
16. C. Topaz, L. Ziegelmeier, T. Halverson, Topological data analysis of biological aggregation models, *PLoS ONE* 10 (2015) e0126383.
17. A. Zomorodian, G. Carlsson, Computing persistent homology. *Discrete and Computational Geometry*, 33 (2002) 249–274.