Separating Topological Noise from Features using Persistent Entropy

Nieves Atienza[1], Rocio Gonzalez-Diaz[1], and Matteo Rucco[2]

[1] Applied Math Department, School of Computer Engineering, University of Seville, Seville, Spain,
{natienza,rogodi}@us.es,
[2] Univ. of Camerino, School of Science and Technology, Computer Science Division, Camerino, IT
matteo.rucco@unicam.it

Abstract. Topology is the branch of mathematics that studies shapes and maps among them. From the algebraic definition of topology a new set of algorithms have been derived. These algorithms are identified with "computational topology" or often pointed out as Topological Data Analysis (TDA) and are used for investigating high-dimensional data in a quantitative manner. Persistent homology appears as a fundamental tool in Topological Data Analysis. It studies the evolution of $k-$dimensional holes along a sequence of simplicial complexes (i.e. a filtration). The set of intervals representing birth and death times of $k-$dimensional holes along such sequence is called the persistence barcode. $k-$dimensional holes with short lifetimes are informally considered to be "topological noise", and those with a long lifetime are considered to be "topological feature" associated to the given data (i.e. the filtration). In this paper, we derive a simple method for separating topological noise from topological features using a novel measure for comparing persistence barcodes called *persistent entropy*.

Keywords: Persistent homology, persistence barcodes, Shannon entropy, topological noise, topological features

1 Introduction

Persistent homology studies the evolution of $k-$dimensional holes along a sequence of simplicial complexes. Persistence barcode is the collection of intervals representing birth and death times of $k-$dimensional holes along such sequence. In persistence barcode, $k-$dimensional holes with short lifetimes are informally considered to be "topological noise", and those with a long lifetime are "topological features" of the given data.

In general, "very" long living intervals (long lifetime) are considered topological features since they are stable to "small" changes in the filtration. Nevertheless, the definition of what a "topological feature" is, depends on the application. This way, the technique presented in this paper should be considered as an option

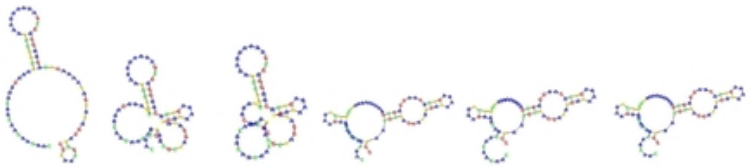

Fig. 1. From left to right: RNA secondary suboptimal structures within different bacteria.

that can be used for discriminating between topological features and topological noise. Moreover, we claim it is very easy (and fast) to compute, and easy to adapt depending on the application.

In [1] a methodology is presented for deriving confidence sets for persistence diagrams to separate topological noise from topological features. The authors focused on simple, synthetic examples as proof of concept. Their methods have a simple visualization: one only needs to add a band around the diagonal of the persistence diagram. Points in the band are consistent with being noise. The first three methods are based on the distance function to the data. They started with a sample from a distribution $\mathbb{P}$ supported on a topological space $\mathfrak{C}$. The bottleneck distance is used as a metric on the space of persistence diagrams. The last method uses density estimation. The advantage of the former is that it is more directly connected to the raw data. The advantage of the latter is that it is less fragile; that is, it is more robust to noise and outliers.

Persistent entropy (which is the Shannon entropy of the persistence barcode) is a tool formally defined in [2] and used to measure similarities between two persistence barcodes. A precursor of this definition was given in [3] to measure how different the intervals of a barcode are in length.

In this paper, we use the difference of persistent entropy to measure similarities between two persistent barcodes. More concretely, we derive a simple method for separating topological noise from topological features of a given persistence barcode obtained from a given filtration (ie., a sequence of simplicial complexes) using the mentioned persistent entropy measurement.

2 Related work

Persistent homology based techniques are nowadays widely used for analyzing high dimensional dataset and they are good tool for shaping these dataset and for understanding the meaning of the shapes. There are several techinques for building a topological space from the data. The main approach is to complete the data to a collection of combinatorial objects, i.e. simplices. A nested collection of simplices forms a simplicial complex. Simplicial complexes can be obtained from graphs and point cloud data (PCD) [4,5]. For example, PCD can be completed to simplicial complexes by using the Vietoris-Rips approach. Vietoris-Rips filtration is a versatile tool in topological data analysis. It is a sequence of simplicial

complexes built on a metric space to add topological structure to an otherwise disconnected set of points. It is widely used because it encodes useful information about the topology of the underlying metric space. The mathematical details of Vietoris-Rips filtration are given in Section 3.

Let's take a look at Fig. 1, it represents a collection of RNA secondary suboptimal structures within different bacteria. All the shapes are characterized by several circular substructures, each of them is obtained by linking different nucleotides. Each substructure encodes functional properties of the bacteria. Persistent homology properly identifies these substructures. For the love of preciseness, Mamuye et al. [6], used Vietoris-Rips complexes and persistent homology for certifying that there are different species but characterized with the same RNA suboptimal secondary structure, thus these species are functionally equivalent.

In [7], the authors proposed a new methodology based on information theory and persistent homology for classifying real length noisy signals produced by small DC motors. They introduced an innovative approach based on "auto mutual information" and the "CAO's method" for providing the time delay embedding of signals. The time delay embedding transforms the signal into a point cloud data in $\mathbb{R}^d$, where d is the dimension of the new space. Vietoris-Rips complex is then computed and analyzed by persistent homology. The authors classified the signal in two classes, respectively "properly working" and "broken".

However, Vietoris-Rips based analysis suffers of the selection of the parameter ϵ. Generally speaking, for different ϵ, different topological features can be observed. In [7], ϵ was selected as the euclidean distance among the points in the new space. We remark that the parameter ϵ does not have a unique physical meaning and it depends on the problem under analysis. For example, in [8], several applications of Vietoris-Rips based analysis to biological problems have been reported and examples of different ϵ with different meaning were found. In order to select the best ϵ, some statistics have been provided what it is known as " persistence landscape" [9]. Landscape is a powerful tool for statistically assessing the global shape of the data over different ϵ. Technically speaking, a landscape is a piecewise linear function that basically maps a point within a persistent diagram (or barcode) to a point in which the $x-$coordinate is the average parameter value over which the feature exists, and the $y-$coordinate is the half-life of the feature. Landscape analysis allows to identify topological features and which are not. In Section 5 we propose an alternative approach to landscape. The main difference between landscape and our method is that the former uses the average of ϵ, while the latter works directly on a fixed ϵ.

3 Background

This section provides a short recapitulation of the basic concepts needed as a basis for the presented method for separating topological noise from features.

Informally, a topological space is a set of points each of them equipped with the notion of neighboring. A simplicial complex is a kind of topological space

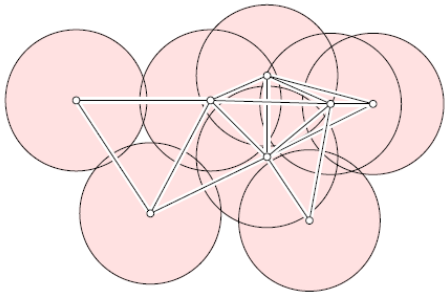

Fig. 2. [12, p. 72] Nine points with pairwise intersections among the disks indicated by straight edges connecting their centers, for a fixed time ϵ. The Čech complex $\check{C}_\epsilon(V)$ fills nine of the ten possible triangles as well as the two tetrahedra. The Vietoris-Rips complex $VR_\epsilon(V)$ fills the ten triangles and the two tetrahedra.

constructed by the union of n-dimensional simple pieces in such a way that the common intersection of two pieces are lower-dimensional pieces of the same kind. More concretely, an abstract *simplicial complex* K is composed by a set K_0 of $0-$simplices (also called vertices V, that can be thought as points in $\mathbb{R}^n$); and, for each $k \geq 1$, a set K_k of $k-$simplices $\sigma = \{v_0, v_1, \ldots, v_k\}$, where $v_i \in V$ for all $i \in \{0, \ldots, k\}$, satisfying that:

- each $k-$simplex has $k+1$ faces obtained removing one of its vertices;
- if a simplex σ is in K, then all faces of σ must be in K.

The underlying topological space of K is the union of the geometric realization of its simplices: points for 0-simplices, line segments for 1-simplices, filled triangles for 2-simplices, filled tetrahedra for 3-simplices and their n-dimensional counterparts for n-simplices. We only consider finite (abstract) simplicial complexes with finite dimension, i.e., there exists an integer n (called the dimension of K) such that for $k > n$, $K_k = \emptyset$ and for $0 \leq k \leq n$, K_k is a finite set. See [10] and [11] for an introduction to algebraic topology.

Two classical examples of abstract simplicial complexes are Čech complexes and Vietoris-Rips complexes (see [12, Chapter III]). Let V be a finite set of points in $\mathbb{R}^n$. The *Čech complex* of V and r denoted by $\check{C}_r(V)$ is the abstract simplicial complex whose simplices are formed as follows. For each subset S of points in V, form a closed ball of radius $r/2$ around each point in S, and include S as a simplex of $\check{C}_r(V)$ if there is a common point contained in all of the balls in S. This structure satisfies the definition of abstract simplicial complex. The *Vietoris-Rips complex* denoted as $VR_r(V)$ is essentially the same as the Čech complex. Instead of checking if there is a common point contained in the intersection of the $(r/2)-$ball around v for all v in S, we may just check pairs adding S as a simplex of $\check{C}_r(V)$ if all the balls have pairwise intersections. We have $\check{C}_r(V) \subseteq VR_r(V) \subseteq \check{C}_{\sqrt{2}r}(V)$. See Fig.2.

Homology is an algebraic machinery used for describing topological spaces. The $k-$Betti number β_k represents the rank of the $k-$dimensional homology

group of a given simplicial complex K. Informally, β_0 is the number of connected components, β_1 counts the number of loops in $\mathbb{R}^2$ or tunnels in $\mathbb{R}^3$, β_2 can be thought as the number of voids and, in general, β_k can be thought as the number of k-dimensional holes.

Persistent homology is a method for computing $k-$dimensional holes of K at different spatial resolutions. The key idea is as follows: First, the space must be represented as a simplicial complex and a distance function must be defined on the space. Second, a filtration of the simplicial complex, that is a nested sequence of increasing subsets (referred above as different spatial resolutions), is computed. More concretely, a filtration of a simplicial complex K is a collection of simplicial complexes $\{K(t)|t \in \mathbb{R}\}$ of K such that $K(t) \subset K(s)$ for $t < s$ and there exists $t_{\max} \in \mathbb{R}$ such that $K_{t_{\max}} = K$. The filtration time (or filter value) of a simplex $\sigma \in K$ is the smallest t such that $\sigma \in K(t)$.

Then, persistent homology describes how the homology of a given simplicial complex K changes along filtration. If the same topological feature (i.e., $k-$dimensional hole) is detected along a large number of subsets in the filtration, then it is likely to represent a true feature of the underlying space, rather than artifacts of sampling, noise, or particular choice of parameters. More concretely, a $k-$dimensional Betti interval, with endpoints $[t_{start}, t_{end})$, corresponds to a $k-$dimensional hole that appears at filtration time t_{start} and remains until filtration time t_{end}. The set of intervals representing birth and death times of homology classes is called the *persistence barcode* associated to the corresponding filtration. For more details and a more formal description we refer to [12].

4 Persistent Entropy

In order to measure how much the construction of a filtered simplicial complex is ordered, a new entropy measure, the so-called *persistent entropy*, were defined in [2]. A precursor of this definition was given in [3] to measure how different the intervals of a barcode are in length. In [13], persistent entropy is used for addressing the comparison between discrete piece-wise linear functions.

Given a Čech or Vietoris-Rips filtration $F = \{K(t)|t \leq T\}$ (in practice one will never construct the filtration up to the end and will stop at a certain time T), and the corresponding persistence barcode $B = \{[a_i, b_i) : 1 \leq i \leq n\}$, let $L = \{\ell_i = b_i - a_i : 1 \leq i \leq n\}$. The *persistent entropy* H of the filtration F is:

$$H_L = -\sum_{i=1}^{n} \frac{\ell_i}{S_L} log \frac{\ell_i}{S_L}, \qquad \text{being } S_L = \sum_{i \in I} \ell_i.$$

Note that the maximum persistent entropy would correspond to the situation in which all the intervals in the barcode are of equal length. Conversely, the value of the persistent entropy decreases as more intervals of different lengths are present. More concretely, if B has n intervals, the possible values of the persistent entropy H_L associated with the barcode B lie in the interval $[0, \log(n)]$.

The following result supports the idea that persistent entropy can differentiate long from short intervals as we will see in the next section.

Theorem 1. *For a fixed integer i, $1 \leq i \leq n$, let $L_i = \{\ell_{i+1}, \dots \ell_n\}$, $S_i = \sum_{j=i+1}^{n} \ell_j$ and let H_i be the persistent entropy associated to L_i. Let*

$$L'(i) = \{\ell'_1, \dots, \ell'_i, \ell_{i+1}, \dots, \ell_n\}, \quad \text{where } \ell'_j = S_i/e^{H_i}, \quad \text{for } 1 \leq j \leq i.$$

Then $H_L \leq H_{L'(i)}$.

Proof. Let us prove that $H_{L'}(i)$ is the maximum of all the possible persistent entropies associated to lists of intervals with n elements, such that the last $n-i$ elements of any of such lists is $\{\ell_{i+1}, \dots, \ell_n\}$. Let $M = \{x_1, \dots, x_i, \ell_{i+1}, \dots, \ell_n\}$ (where $x_j > 0$ for $1 \leq j \leq i$) be any of such lists. Let $S_x = \sum_{j=1}^{i} x_j$. Then, the persistent entropy associated to M is:

$$H_M = \sum_{j=1}^{i} \frac{x_j}{S_x + S_i} \log\left(\frac{x_j}{S_x + S_i}\right) + \sum_{j=i+1}^{n} \frac{\ell_j}{S_x + S_i} \log\left(\frac{\ell_j}{S_x + S_i}\right).$$

In order to find out the maximum of H_M with respect to the unknown variables x_k, $1 \leq k \leq i$, we compute the partial derivative of H_M with respect to those variables:

$$\frac{\partial H_M}{\partial x_k} = \frac{1}{(S_x + S_i)^2} \left(-S_i H_i + S_i \log\left(\frac{S_i}{x_k}\right) + \sum_{j \neq k} x_j \log\left(\frac{x_j}{x_k}\right)\right).$$

Finally, $\{x_k = \frac{S_i}{e^{H_i}} : 1 \leq k \leq i\}$ is the solution of the system $\{\frac{\partial H_M}{\partial x_k} = 0 : 1 \leq k \leq i\}$. □

5 Separating topological features from topological noise

Let us start with a sample V from a distribution $\mathbb{P}$ supported on a topological space $\mathfrak{C}$. Suppose the Vietoris-Rips filtration F is computed from V, and the persistence barcodes B is computed from F. The following are the steps of our proposed method, based on persistent entropy, to separate topological noise from topological features in the persistence barcode B, estimating, in this way, the topology of $\mathfrak{C}$.

1. Order the intervals in B by decreasing length. Then $L = \{\ell_i = b_i - a_i : 1 \leq i \leq n\}$ satisfies that $\ell_i \leq \ell_j$ for $i < j$;
2. Compute the persistent entropy H_L of B. Denote $H_{L'(0)} := H_L$.
3. From $i = 1$ to $i = n$,
   a. Compute the persistent entropy $H_{L'(i)}$ for $L'(i) = \{\ell'_1, \dots, \ell'_i, \ell_{i+1}, \dots, \ell_n\}$, being $\ell'_k = \frac{S_i}{e^{H_i}}$ for $1 \leq k \leq i$ as in Th. 1.
   b. Compute $H_{rel(i)} = (H_{L'(i)} - H_{L'(i-1)})/(\log(n) - H_L)$.
   c. If $H_{rel(i)} > \frac{i}{n}$, then the associated interval $[a_i, b_i)$ represents a topological feature. Otherwise, the interval $[a_i, b_i)$ represents noise.

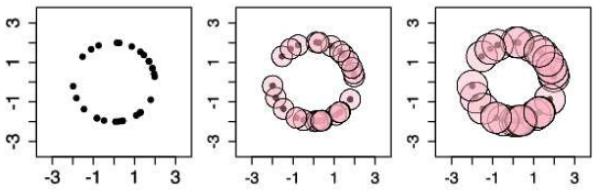

Fig. 3. Left: 30 data points sampled from a circle of radius 2. Middle: Balls of radius 0.5 centered at the sample points. Right: Balls of radius 0.8 centered at the sample points.

Steps 1, 2 and 3.a can be considered as a general method for any kind of application. For $1 \leq i \leq n$, $H_{L'(i)}$ is the entropy of the barcode obtained by replacing the intervals $\ell_1, \ldots, \ell_i$ by i intervals that maximize the entropy. Observe that $H_{L'(0)} = H_L$, $H_{L'(i)} < H_{L'(j)}$ for $0 \leq i < j \leq n$ and $H_{L'(n)} = \log(n)$ by Th. 1.

Step 3.b and 3.c are used to test a possible dissimilarity measure to differentiate topological features from noise. These two steps could be modified later depending on the application. In this paper, we use $H_{L'(i)} - H_{L'(i-1)}$ to measure the influence of the current interval ℓ_i in the initial persistent entropy H_L. It is in order to appreciate this influence, why we divide $H_{L'(i)} - H_{L'(i-1)}$ by the difference of the possible maximal entropy (which is $\log(n)$) and H_L. Then, we compare the resulting $H_{rel}(i)$ with $\frac{i}{n}$ since $H_{rel(i)}$ is affected by the total number of intervals and the number of intervals we are replacing.

We have applied our methodology to two different scenarios. First, we take 30 data points sampled from a circle of radius 2 (see Fig. 3.Left). This example has been taken from paper [1]. Vietoris-Rips complex for $t = 0.5$ can be deduced from the picture shown in Fig. 3.Middle which consists of two connected components and zero loops. Looking at Vietoris-Rips complex for $t = 0.8$ (see Fig. 3.Right), we assist at the birth and death of topological features: at $t = 0.8$, one of the connected components has died (was merged with the other one), and a loop appears; this loop will die at $t = 2$, when the union of the pink balls representing the distance function becomes simply connected.

In our method, an interval is considered to be a feature if $H_{rel(i)} > \frac{i}{n}$. In Table 5.Left, we have applied our method to the intervals that make up the barcode (without differentiating dimension). This way, only the intervals with length 2 (that corresponds to the connected component that survives until the end) and 1.2 (that correspond to the loop that appears at $t = 0.8$ and disappears at $t = 2$) are considered features. Later, in Table 5.Righ,t we have applied our method to the intervals that make up the 0-barcode (i.e., the lifetime of the connected components along the filtration). This way, the intervals with length 2 and 0.7 (that corresponds to the connected components that dies just before the loop is created) are considered features. This example highlight that we the

results may be different depending on if we apply our method to the whole set of intervals of the barcode or if we do it dimension by dimension.

Table 1. Results of our method applied to the intervals that make up: Left) the barcode (i.e., without differentiating dimension); and Right) the 0-barcode; both associated to the Vietoris-Rips filtration obtained from 30 data points sampled from a circle of radius 2 (see Fig. 3.Left).

ℓ_i	$\frac{H_i}{\log(n)}$	$H_{rel(i)}$	Feature
2.	0.967011	0.542391	yes
1.2	0.985761	0.260088	yes
0.7	0.991422	0.07853	no
0.45	0.992506	0.0150434	no
0.45	0.993746	0.0171948	no
...	...	...	...

l_i	$\frac{H_i}{\log(n)}$	$H_{rel(i)}$	Feature
2.	0.985109	0.77248	yes
0.7	0.991032	0.0905039	yes
0.45	0.992167	0.0173301	no
0.45	0.993463	0.0198057	no
0.4	0.994199	0.0112466	no
...	...	...	...

Consider now a set V of 400 points sampled from a 3D torus. The barcodes (separated by dimension) computed from the Vietoris-Rips filtration associated to V are showed in Fig. 4. We have applied our method to the 0-barcode (lifetime of connected components along the V-R filtration) and the 1-barcode (lifetime of loops along the V-R filtration). See Table 5.Left and 5.Right, respectively. The interval of length 1.9 in the table on the left corresponds to the connected component that survives until the end. The intervals of length 1.531 in the table on the right corresponds to the two tunnels of the 3D torus. In Table 5 we show the results of our method applied to all the intervals of the barcode without separating by dimensions. We can see in this case that we obtain the same features as before plus the interval representing the void.

Table 2. Results of our method applied to the 0-barcode (table on the left) and the 1-barcode (table on the right) associated to the Vietoris-Rips filtration obtained from 400 points sampled from a 3D torus.

ℓ_i	$\frac{H_i}{\log(n)}$	$H_{rel(i)}$	Feature
1.9	0.996295	0.442767	yes
0.396	0.996325	0.00449624	no
0.387	0.996351	0.00386916	no
0.387	0.996376	0.00389884	no
0.387	0.996403	0.00392887	no
...	...	...	...

ℓ_i	$\frac{H_i}{\log(n)}$	$H_{rel(i)}$	Feature
1.531	0.918238	0.238936	yes
1.531	0.950752	0.302654	yes
0.27	0.952044	0.012028	no
0.261	0.953275	0.011451	no
0.234	0.954209	0.00869544	no
...	...	...	...

Table 3. Results of our method applied to the barcode (without differentiating dimension) associated to the Vietoris-Rips filtration obtained from 400 points sampled from a 3D torus.

ℓ_i	$\frac{\ell_i}{L}$	ℓ'_i	$\frac{\ell'_i}{L'(i)}$	$\frac{H_{L'(i)}}{\log(n)}$	$H_{rel(i)}$	Feature
1.9	0.0145219	0.268369	0.00207708	0.971259	0.0799069	yes
1.531	0.0117016	0.262812	0.00205432	0.972992	0.0554616	yes
1.531	0.0117016	0.257239	0.00203115	0.974775	0.0570812	yes
1.234	0.00943158	0.253276	0.00201566	0.975978	0.0385369	yes
0.396	0.00302667	0.252916	0.00201511	0.976021	0.00137745	no
...	...	...	...	...	...	...

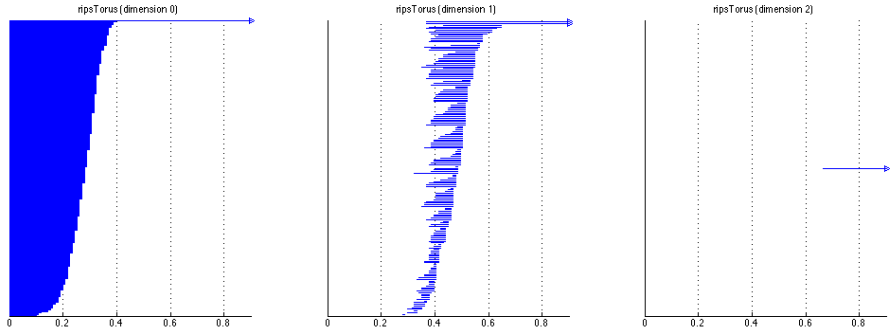

Fig. 4. Barcodes (separated by dimension) computed from the Vietoris-Rips filtration associated to a point cloud lying on a 3D torus. Left: lifetimes of connected components. Middle: lifetimes of tunnels. Right: lifetimes of voids.

6 Conclusions and future work

In this paper, we have derived a method for separating topological noise from topological features using the Shannon entropy of persistence barcode. We have proven that the method is consistent by proving that in step i of the method we replace i intervals by the same number of intervals but with the length that maximizes the entropy. This way we "neutralize" the effect of such i intervals and, by computing the difference of the entropies obtained in step $i-1$ and step i, we can deduce if the interval at position i is a topological feature or not.

We intend to adapt our method to study RNA data from healthy and unhealthy cells. We argue the method will let to highlight the topological features that are formed by the most relevant genes associated to pathologies.

Acknowledgments. Authors are partially supported by IMUS, University of Seville under grant VPPI-US and Spanish Government under grant MTM2015-67072-P (MINECO/FEDER, UE). We also thank the reviewers for their valuable and constructive comments.

References

1. B. T. Fasy, F. Lecci, A. Rinaldo, L. Wasserman, S. Balakrishnan, A. Singh, Confidence sets for persistence diagrams, The Annals of Statistics 6 (2014) 2301–2339.
2. M. Rucco, F. Castiglione, E. Merelli, M. Pettini, Characterisation of the idiotypic immune network through persistent entropy, in: Proc. Complex, 2015.
3. H. Chintakunta, T. Gentimis, R. Gonzalez-Diaz, M. J. Jimenez, H. Krim, An entropy-based persistence barcode, Pattern Recognition 48 (2) (2015) 391–401.
4. J. Binchi, E. Merelli, M. Rucco, G. Petri, F. Vaccarino, jholes: A tool for understanding biological complex networks via clique weight rank persistent homology, Electronic Notes in Theoretical Computer Science 306 (2014) 5–18.
5. H. Adams, A. Tausz, Javaplex tutorial (2011).
6. A. Mamuye, E. Merelli, M. Rucco, Persistent homology analysis of the rna folding space, in: Proc. 9th EAI Conference on Bio-inspired Information and Communications Technologies (BICT 2015), 2015.
7. M. Rucco, E. Concettoni, C. Cristalli, A. Ferrante, E. Merelli, Topological classification of small dc motors, in: 1st Int. Forum on Research and Technologies for Society and Industry (RTSI), IEEE, 2015, pp. 192–197.
8. N. Jonoska, M. Saito, Discrete and Topological Models in Molecular Biology, Springer, 2013.
9. P. Bubenik, Statistical topological data analysis using persistence landscapes, The Journal of Machine Learning Research 16 (1) (2015) 77–102.
10. A. Hatcher, Algebraic topology cambridge university press, Cambridge, UK.
11. J. R. Munkres, Elements of algebraic topology, Vol. 2, Addison-Wesley, 1984.
12. H. Edelsbrunner, J. Harer, Computational topology: an introduction, American Mathematical Soc., 2010.
13. M. Rucco, R. Gonzalez-Diaz, M. J. Jimenez, N. Atienza, C. Cristalli, E. Concettoni, A. Ferrante, E. Merelli, A new topological entropy-based approach for measuring similarities among piecewise linear functions, CoRR abs/1512.07613.