INTRINSIC DISEASE MAPS USING
PERSISTENT COHOMOLOGY

DANIEL AMIN
CUNY Graduate Center
365 5th Avenue
New York, NY 10016

MIKAEL VEJDEMO-JOHANSSON*
Department of Mathematics
CUNY College of Staten Island
2800 Victory Boulevard
Staten Island, NY 10314

(Communicated by Gunnar Carlsson)

Abstract. We use persistent cohomology and circular coordinates to investi-
gate three datasets related to infectious diseases. We show that all three
datasets exhibit circular coordinates that carry information about the data it-
self. For one of the datasets we are able to recover time post infection from
the circular coordinate itself – for the other datasets, this information was not
available, but in one we were able to relate the circular coordinate to red blood
cell counts and weight changes in the subjects.

1. Introduction. Disease maps were introduced in [5] as a visualization technique
to study progression of infectious diseases from data taken from individuals at dif-
f erent stages of the disease progression. The disease map introduced is, essentially,
a Mapper complex built from physiological data of the population used to map out
the disease. In [5] the authors trace out Mapper complexes that prominently feature
a loop structure, and they are able to connect that loop structure to the phases of
infection.

As observed in [6], Mapper complexes can both obscure and introduce topological
features if the Mapper cover is not approximately a good cover in a topological sense.
This provokes the question: does the loop structure observed in the disease maps
exist in the original data space? Dimensionality reduction techniques carry the same
potential of accidentally introducing topological structures: the same question can
be asked of loop structures that emerge in low-dimensional PCA projections of the
data.

We address this question by examining two datasets drawn from the study of
Malaria in mice and in humans, and one dataset drawn from the study of Hep-
atitis C in humans. One dataset is drawn from [1] and provides weight, glucose
concentration, red blood cell counts and cytometric cell counts from mice infected

2020 Mathematics Subject Classification. 55N31.
Key words and phrases. Disease maps, persistent cohomology, circular coordinates, malaria,
Hepatitis C.
* Corresponding author: Mikael Vejdemo-Johansson.
with malaria parasites. The second dataset is drawn from [5] and contains RNA expression rates for 109 different RNA snippets. The third dataset comes from [4] and contains RNA expression rates for 56,300 different RNA snippets.

On these datasets we then compute circular coordinates using the methodology introduced in [2]. The persistent cohomology computation step provides barcode evidence for the loop structure, and the generated circular coordinate function provides an intrinsic disease phase coordinate that maps out the disease progression in the full data space.

As we look closer into the three datasets under study here, we see a collection of different behaviours appear. All three datasets exhibit circular coordinates, but their interpretations and the extent to which they correlate with temporal aspects vary.

- Circular coordinate comes from a bar approximately 2x the size of the next smaller bar in the persistence barcode. Some gradient behaviours, but largely a clustering of coordinate values corresponding to the two “lobes” in the apparent structure of the data.
- Circular coordinate comes from a bar at least 5x the size of the next smaller bar in the persistence barcode. The coordinate values take on the entire range of the circular coordinates, and we can show a clear connection between the circular coordinate and a time variable.
- Circular coordinate in full dimensionality comes from a bar approximately 2x the size of the next smaller bar, and the coordinate looks to be almost constant. When computing circular coordinates on a projection of the dataset, a very strong circular coordinate appears – but without any connection to the available time variable.

The disease maps in [5] really shows us a best possible case scenario for disease maps; both the Mapper-based methods in [5], and persistent cohomology calculated in the full high-dimensional data space both show a clear circular structure that agrees with available time information.

The data in [4] looks promising based on the approach in [5]: the data is longitudinal of an infectious disease and with a large swathe of RNA expression rates to analyze. However, the circular coordinate structure is all but absent in the full dimension, and while there is a strong circular coordinate in the projection to two dimensions it turns out to be unrelated to the time variable.

Similarly, the data in [1] also looks promising for a disease map approach: it is longitudinal data of the progression of an infectious disease, and with a range of physiological variables to analyze. Here, the circular coordinate structure is relatively weak, and seems to more reflect a soft clustering of the data rather than a circular structure.

Given these varying results, we would recommend a practitioner to combine the disease map approach in [5] with cohomology computations: verifying any apparent cyclical structures with cohomology barcodes computed in the full dimensionality of the raw data.

This paper is structured as follows: in Section 2, we describe the persistent cohomology and circular coordinates approach that we are using, as well as the visualization choices we make for our later visualizations of the data. In Section 3, we describe the datasets we study and the processing we have performed. In Section 4, we show the results of applying our methods to the chosen data and discuss the resulting visualizations and their content.
2. Method. The circular coordinates approach goes through the following steps: First construct a Vietoris-Rips complex from the data. Compute persistent cohomology of the Vietoris-Rips complex. The bars in the barcode correspond to more or less significant circle-valued coordinate functions. Pick a bar and truncate the Vietoris-Rips complex to end during the span of that bar. Using the boundary map \( \delta^0 : C^0 VR_n(X) \to C^1 VR_n(X) \), and a representative cocycle \( \zeta \) of the coclass corresponding to the chosen bar, minimize \( \| \zeta - \delta^1 \alpha \|_2 \). A map from \( VR_n(X) \to [0,1]/\{0 \sim 1\} \) can be constructed as \( \theta = \alpha \pmod{1} \).

We will describe each step in this process in more detail next.

2.1. Cohomology. Cohomology is dual to the more well known homology functor. The construction starts with the chain complex of a simplicial complex: \( C_n X \) is a graded vector space with homogenous components \( C_n X \) spanned by the \( n \)-simplices in \( X \), equipped with a boundary map \( \partial : C_n X \to C_{n-1} X \) that obeys \( \partial^2 = 0 \). The boundary map for a simplicial complex is usually taken to be \( \partial[v_0, \ldots, v_d] \mapsto \sum_j (-1)^j [v_0, \ldots, \hat{v}_j, \ldots, v_d] \), where \( \hat{v}_j \) means leaving \( v_j \) out of the vertex list for that simplex. For a simplicial complex \( X \), its cochain complex \( C^* X \) with coefficients in a field \( \mathbb{k} \) is a graded vector space with homogenous components \( C^n X = \text{hom}(C_n X, \mathbb{k}) \). The cochain complex is equipped with a coboundary operator \( \delta : f \mapsto f \circ \partial \).

Since circular coordinates as described in [2] are generated by cohomology in degree 1, we can focus our attention to three cochain spaces: \( C^0 \) represents functions defined on vertices, \( C^1 \) represents functions defined on edges and \( C^2 \) represents functions defined on triangles of the simplicial complex. The corresponding cochain complex of a simplicial complex \( X \) is

\[
C^0 X \xrightarrow{\delta^0} C^1 X \xrightarrow{\delta^1} C^2 X
\]

We define degree 1 cohomology \( H^1 X = \ker \delta^1 / \text{img} \delta^0 \). Here, an edge function \( f \) is in \( \text{img} \delta^0 \) if \( f(u, v) = g \circ \partial[u, v] = g(v - u) = g(v) - g(u) \) for some vertex function \( g \). In other words, the coboundaries are precisely edge functions that measure some sort of potential differences between vertices. An edge function \( f \) is in \( \ker \delta^1 \) if \( \delta^1 f(a, b, c) = 0 \). Now, \( \delta^1 f(a, b, c) = f \circ \partial[a, b, c] = f(bc - ac + ab) = f(b, c) - f(a, c) + f(a, b) \), vanishes precisely when \( f(a, b) + f(b, c) = f(a, c) \). Hence the cocycles are precisely edge functions that are path independent around triangles in the simplicial complex.

2.2. Circular coordinates. Given a cocycle \( \zeta \) representing some cohomology class \( [\zeta] \in H^1(X; \mathbb{Z}) \), [2] recalled a classical topological construction of a function \( \theta : X \to S^1 \) that maps all vertices to 0 and wraps edges around the circle with winding number \( \zeta(u, v) \) assigned to each edge \( [u, v] \). [2] proved that this map can be smoothed to a map that maps each edge \( [u, v] \) linearly to an interval of length \( \bar{\zeta} = \zeta + \delta^0 f \), where \( f \in C^0(X; \mathbb{R}) \). The resulting \( \theta \) can be calculated as \( \theta = f \pmod{1} \).

2.3. Vietoris-Rips. In order to perform Cohomology and Circular Coordinates calculations, we convert our input dataset in the form of point cloud, into simplicial complexes. Vietoris-Rips complex is one of the constructions we can use among
many others, however, in this work, it’s the only method we focus on. The Vietoris-Rips complex \( V_{\epsilon}X \) has vertices given by the dataset \( X \) and includes a \( k \)-simplex whenever \( k + 1 \) vertices lie pairwise within distance \( \epsilon \) from each other.

2.4. **Persistent cohomology.** Persistent cohomology is a method of finding persistent cocycles that are stable across a range of \( \epsilon \)-values. The result is a barcode diagram portraying intervals where specific cocycles satisfy the cocycle condition without being coboundaries themselves. From this diagram we pick the longest interval, typically, which indicates a stable cocycle corresponding to a stable circle-valued coordinate, in the hope of capturing the shape of data instead of noise.

2.5. **Lifting to integer coefficients.** Most persistent cohomology software packages work with coefficients in a finite field \( \mathbb{F}_p = \mathbb{Z}/p\mathbb{Z} \). The circular coordinates construction however requires the use of a cocycle with coefficients in the integers \( \mathbb{Z} \). The method suggested in [2] is to pick coefficients in the range of \( \{-\frac{p-1}{2}, \ldots, -1, 0, 1, \ldots, \frac{p-1}{2}\} \) – that way the most common coefficients of \( \pm 1 \) can be represented by \( \pm 1 \). Computationally, this can be done by replacing a coefficient value \( \alpha \) with \( (\alpha + \frac{p-1}{2} \mod p) - \frac{p-1}{2} \).

2.6. **Harmonic smoothing.** As discussed above in Section 2.2, after obtaining integer cocycle \( \zeta \), we want to convert it into a real cocycle to smooth out the mapping. [2] suggests using the \( L_2 \)-norm to measure smoothness. This turns the smoothing of an integer cocycle into a least squares optimization problem. By a proposition from [2], it turns out that if \( f \) minimizes the least squares problem

\[
\min_f \| \zeta - \delta^0 f \|^2
\]

then \( \theta = f \pmod{\mathbb{Z}} \) forms a smooth circle-valued coordinate function.

2.7. **Visualization.** To visualize the data, we use PCA projection onto two coordinates. To visualize specific circular coordinate functions, we color scatterplots of the PCA coordinates by the coordinate function values. In order to show the periodic nature of the coordinates, we use the HSV (Hue/Saturation/Value) colormap to assign colors to the span \([0, 1]\) in such a way that the colors assigned to values near 0 agree with the colors assigned to values near 1. We also include scatterplots of ground truth time data against circular coordinates to illustrate how the circular coordinate captures the disease progression. These scatterplots have to be read as wrapping around vertically.

3. **Data.** Two malaria datasets we use are found from two separate papers; one discusses host energy source and its importance for disease tolerance [1], and the other maps disease space to track resilience to infections [5].

The first paper studies how energy level matters for each mouse with respect to disease tolerance, by treating mice with different amount of glucose and comparing the effects of the disease as well as survival rate. The data consists of mouse features and study parameters such as weight of mouse, amount of glucose given, Accuri count and RBC (red blood cells count) taken daily. RBC is measured by taking samples of \( 4 \mu L \) of blood from mice and diluting them with Hanks’ Balanced Salt Solution. Accuri is a flow cytometry measurement tool, which detects and counts the number of cells in samples. In this dataset, Accuri represents the whole absolute count of all cells within the blood samples including both red blood cells and parasites contained. These total of four features (weight, glucose, RBC, Accuri)
are then normalized using the Standard Scaler from scikit-learn [3] and used for the subsequent topological analysis.

The second paper focuses on disease progression and how different each malaria case (mice and humans) goes through the disease. The data consists of relatively similar features as the first data with weight, RBC taken daily, however, it also includes RNA expressions by extraction from the blood samples. In this data, we decide to only use RNA expressions in disease mapping. Results did not differ noticeably from the omission of the physiological variables. The detailed processes of extracting the RNA expressions data is explained in the paper [5]. Their processing includes median centering and quartile normalization. There are 109 total RNA expressions used for analysis.

In addition to the two Malaria datasets, we also use data from [4]. This is a longitudinal transcriptomic study of Hepatitis C infections in human hosts, with spontaneous viral clearance: patients with Hepatitis C infections were sampled at four time intervals: a pre-infection baseline, an early acute (2-9 weeks after infection), a late acute (15-20 weeks after infection) and a late follow up (25-71 weeks after infection). From each sampled time and each of 14 subjects, a blood sample was taken, and RNA extracted and sequenced, then aligned to a the human genome reference (hg19). Gene read counts from 56 300 genes were deposited in the Gene Expression Omnibus database under the GEO Series accession number GSE119117. These gene read counts form the data set we analyze.

4. Results. In order to visualize the topological analysis, a dimension reduction in the form of Principal Component Analysis is done across features at the end of the process to acquire 2 dimensions for mapping. In Figures 1b and 3b we show these dimension reduced data points, colored by the corresponding circular coordinate function. The result of topological analysis for the data from the first paper [1] is shown in Figure 1. As seen, we pick the longest bar that represents the longest, most stable cocycle across \( \epsilon \). The barcode is not as resoundingly compatible with the cohomological signature of a circle as the barcode in Figure 3a, and while the most dominant cocycle still gives rise to a circular coordinate, it is more reminiscent of a soft clustering of the data than a clear circular structure; in Figure 1b we see a somewhat green region shifting into red or vice versa, with intermediary stages.

From the first dataset, looking at Figure 2a, we can see that the circular coordinate \( \theta \) creates a separation in the RBC (red blood cells count). The authors of the first paper [1] discuss the importance of energy levels with respect to survival and disease tolerance. It is fascinating to see that our mapping corroborates that idea; as red blood cells count increases, the disease gets into a different stage (recovery) while the lower counts corresponds to a different stage (sick) and some intermediary transition in between. Similarly, \( \theta \) also gives some separation to percent weight changes, although not as distinct. As discussed before, weight changes correspond to energy levels of animals infected. The mapping shows two groups here as well, one with small negative or positive weight changes, the other with big negative weight changes. As energy levels of each animal decreases, they get into the sick disease stage, while they get into recovery in the other direction.

The second dataset from [5], in which we only focus on RNA expressions, performs really well. As seen in Figure 3, the barcode graph shows that there is a very stable cocycle we can use. This, in return, results in a very clear mapping of the
Figure 1. Topological analysis on the energy levels and malaria data from [1]. The barcode graph to the left has one bar noticeably longer than the other bars. The corresponding circle-valued coordinate is plotted with the hsv colormap to the right, we see that most of the data splits into two relatively constant regions, with some points in a transition between them.

Figure 2. Further exploration of the dataset from [1]: the circular coordinate value discriminates quite clearly between low/high red blood cell counts (left) and between weight loss/weight gain (right).

disease progression. Exploring further, we can also see how the circular coordinate \( \theta \) works.

In Figure 4a, we can see how the earliest and latest stages of the disease progression look identical to the circular coordinate – these are the times immediately after infection and the times when recovery has progressed relatively far. In between these two identical seeming stable regions lies a region – ranging from day 5 to day 15 – where the circular coordinate increases monotonically with time. We
Figure 3. Topological analysis on RNA Expressions. This barcode graph shows a very clear long bar indicating a highly persistent cocycle encoding a circular coordinate inherent to the data. The corresponding circular coordinate has been plotted with the hsv colormap to the right. We can see the coordinate map constant near the bulk of the points, and then sweeping through a range of values as we go around the cycle. This indicates that the cycle we see in the PCA plot and in the Mapper graph in [5] is present in the full 109-dimensional data space.

also show the PCA plot of the data colored by body temperature in Figure 4b. Comparing Figures 4b and 3b, we can see that the region of points where the circular coordinate is changing corresponds precisely to the region of points where the body temperature is elevated.

The third dataset, with transcriptomics published as GSE119117 in the GEO database, and described in [4] looks quite promising when exploring the data in low-dimensional projections. As can be seen in Figure 6, the 2-dimensional projection contains a circular structure witnessed by a single long bar in the degree 1 persistent cohomology barcode and an exemplary circular coordinate. However, when looking at the same data in its full dimensionality, in Figure 5, the picture is more muddled: the longest bar in the barcode is still long relative to the other bars – but nowhere near as dramatically much longer, and while the corresponding circular coordinate function shows variation along the visibly apparent cycle in the plot in Figure 5b, the variation in the coordinate is much more limited and not as strong an indication of a relevant structure.

In the context of disease maps, it is immediately interesting to find out whether the long bar and strong circular coordinate that can be seen in Figure 6 actually corresponds to disease progression in some meaningful way. However, as can be seen in Figure 7, this does not seem to be the case: the circular structure may be an artifact of the PCA projection, or may be representing some other phenomenon, but it is not representing time-post-infection.
Figure 4. Further exploration: To the left we are plotting the circular coordinate from Figure 3b against the variable Days Post Infection from the original dataset. We can see clearly how the circular coordinate increases with time post infection and marks out a region stretching from approximately day 5 to day 15. To the right, we plot body temperature on the PCA plot. We can see that the points where the coordinate function varies in Figure 3b are precisely the points with deviations in body temperature.

Figure 5. Topological analysis on Hepatitis C RNA Expressions, in full 56,300 dimensions. There is a barcode with one bar significantly longer than all the other bars, indicating a possible circular structure. It produces a circular coordinate function that we depict in the right figure.

5. Conclusion. We apply persistent cohomology and circular coordinates to study two different datasets related to Malaria disease progressions and one related to Hepatitis C progression. In one dataset we find a weak connection between the circular coordinate values and disease indicators such as red blood cell counts and
Figure 6. Topological analysis on Hepatitis C RNA Expressions, after dimensionality reduction to 2 dimensions. In the 2-dimensional projection there is a loud and clear topological signal as can be seen with the long bar in the left figure. It produces a circular coordinate function that we depict in the right figure.

Figure 7. The Hepatitis C RNA Expressions data was collected in four time intervals for each experimental subject: pre-infection, early acute, late acute and late follow up. We plot these four time stamps against the circular coordinate in the left-hand figure, and use the four time stamps for coloring the projected data in the right-hand figure. Even though the projected data has a very clear circular coordinate function, it does not seem to capture the time variable.
weight loss/gain. In the second dataset we find a resoundingly dominant persistence bar corresponding to a circular coordinate that correlates very closely with time-post-infection and that picks out a time region ranging between day 5 and day 15 after initial infection in which the circular coordinate changes are concentrated. This time region we show to correspond precisely to the time region where body temperature is elevated and thus to the time region where the body response to the infection is at its highest. The third dataset shows examples of both of these behaviours – in full 56,300 dimensional space, there is a dominant persistent cohomology bar, but only at about twice the length of the other bars in the barcode, and the corresponding circular coordinate is mostly (but not completely) constant across the dataset. In projection to 2 dimensions using PCA, the dataset instead shows a very clear circular structure with a very dominant persistent cohomology bar and a textbook example of a circular coordinate.

Dimensionality reduction techniques – both linear techniques such as PCA, and topological ones such as Mapper – tend to come with theoretical guarantees that short distances in the full space translate to short distances in the reduced space, but no guarantee that long distances remain long after reduction. This is exactly what causes behaviours such as in the third dataset: the dimensionality reduced version has a much much stronger topological signal than the full dimensionality dataset does.

All in all, our tale is rather a cautionary one. Disease mapping can reflect cyclical structures present in full high-dimensional data space – but both Mapper paradigms and PCA projections can be prone to accidentally introducing cyclical structures in dimensionality reduced data. Cohomology barcodes and circular coordinate maps can help identify such false positives.

REFERENCES

[1] K. Cumnock, A. S. Gupta, M. Lissner, V. Chevee, N. M. Davis and D. S. Schneider, Host energy source is important for disease tolerance to malaria, *Current Biology*, 28 (2018), 1635–1642.
[2] V. de Silva, D. Morozov and M. Vejdemo-Johansson, Persistent cohomology and circular coordinates, *Discrete Comput. Geom.*, 45 (2011), 737–759.
[3] F. Pedregosa, G. Varoquaux and A. Gramfort, et al., Scikit-learn: Machine learning in Python, *J. Mach. Learn. Res.*, 12 (2011), 2825–2830.
[4] B. R. Rosenberg, M. Depla, C. A. Freije, D. Gaucher and S. Mazouz, et al., Longitudinal transcriptomic characterization of the immune response to acute hepatitis c virus infection in patients with spontaneous viral clearance, *PLoS Pathogens*, 14 (2018).
[5] B. Y. Torres, J. H. M. Oliveira, A. T. Tate, P. Rath, K. Cumnock and D. S. Schneider, Tracking resilience to infections by mapping disease space, *PLoS biology*, 14 (2016).
[6] M. Vejdemo-Johansson and A. Leshchenko, Certified mapper: Repeated testing for acyclicity and obstructions to the nerve lemma, in *Topological Data Analysis*, Abel Symposia, 15, Springer, Cham, 2020, 491–515.

Received November 2020; 1st revision January 2021; 2nd revision February 2021.

E-mail address: damin@gradcenter.cuny.edu
E-mail address: mvj@math.csi.cuny.edu