Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in
Systems Biology

The shape of biomedical data

Gunnar Carlsson[1,2]

Abstract

Topological data analysis is a new method for the analysis and modeling of complex data sets, based on the mathematical notion of shape. It is particularly appropriate for problems arising in biology, due to the flexible representation of data sets as network models. In this paper, we show how the method works, give two examples, and discuss the modeling framework that can be constructed based on it.

Addresses

[1] Department of Mathematics, Stanford University, USA
[2] Ayasdi Inc., USA

Corresponding author: Carlsson, Gunnar (carlsson@stanford.edu)

Current Opinion in Systems Biology 2017, **1**:109–113

This review comes from a themed issue on **Future of systems biology**

Edited by **Arnold Levine**

For a complete overview see the Issue and the Editorial

Available online 11 December 2016

http://dx.doi.org/10.1016/j.coisb.2016.12.012

Keywords

Data analysis, Topology, Networks.

Introduction

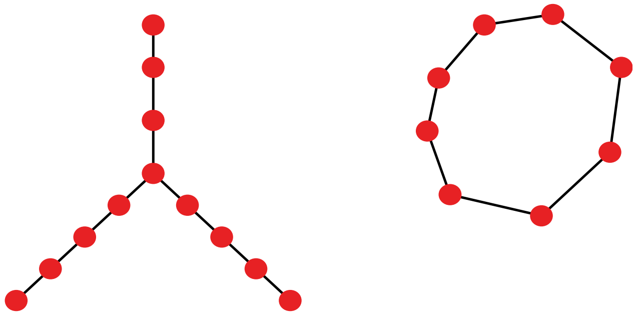
Mathematical modeling has been a very powerful tool for understanding data sets arising from experiments. In physics, for example, simple algebraic models are able to explain naturally arising phenomena, to a very high degree of accuracy. The algebra of matrices and the techniques of differential equations permit us to model physical situations very effectively. There has been a great deal of hope that identical mathematical techniques can be used to good effect to obtain a similar degree of understanding and modeling effectiveness within biology. Although there have been many successes, it is clear that data coming from the life sciences is not modelled nearly as well by the algebraic techniques above as physics is. This is due to several reasons.

- Data arising from the life sciences is inherently much more complex than data from physical systems. Where physical systems are typically driven by a small number of variables which are related by a small number of physical laws, biological systems are instead composed of numerous components which are connected using an extremely complex family of relationships, which are often only understood at a qualitative level.
- Life science data often has a very substantial discrete component, for which continuous algebraic modeling is not useful. Even if the underlying processes are in fact continuous, it is often the case that at a macroscopic level they are best explained by discrete methods. Therefore, it is natural to develop modeling methods that have a discrete nature, or are composed of a combination of continuous and discrete methods.
- The stochastic component to biological systems is much greater than in physical systems, which tends to obscure the structures that are present. It suggests that one needs to develop modeling methodologies that have some degree of robustness to the effect of noise.

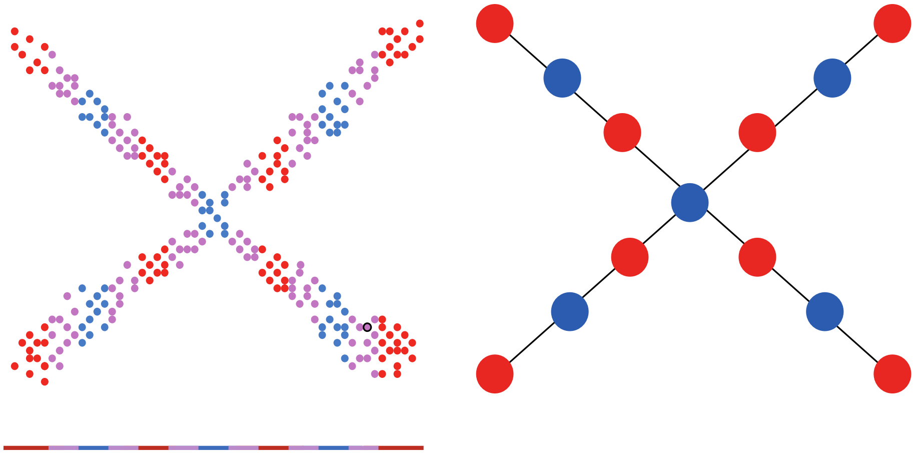
What these observations suggest is that there would be a great deal of value in developing new modeling techniques, that satisfy the above conditions to a greater degree than existing methods. In this paper we will describe a new methodology called *topological data analysis* (*TDA*) that has been demonstrated to be effective for a number of biological problems. TDA is constructed from the mathematical subdiscipline called *topology*, which is the study of shape. The goal is find a methodology intermediate between modeling by algebraic equations, which is continuous but not very flexible, and cluster analysis, which is discrete and therefore misses continuous phenomena.

Description of TDA

The output of TDA, rather than a system of algebraic or differential equations, is instead a *topological network*, by which we mean a collection of nodes and a collection of edges connecting some of the nodes. Such networks are defined purely combinatorially, as lists of nodes and edges with information about which nodes are connected by each edge, but they can be viewed as a shape by using a standard network algorithm. Here are some examples of some topological networks.



The actual distances and layouts of the nodes and edges is only an artifact of the layout algorithm used. The connections are the key piece of information, and that is what is stored. When one constructs a topological network model of a particular data set, each of the nodes corresponds to a collection of data points, and these collections may overlap. We draw an edge between two nodes if the corresponding collections share at least one data point. The construction is based on a similarity measure between data points, measured quantitatively by a *distance function* or a *metric*. Such distance functions assign numerical values to pairs of data points, and are abstractions of our familiar notions of distance in two or three dimensional space, but can be defined in many different ways. For example, if one if is considering genetic sequences, a useful measure of distance between two sequences is a count of the number of entries in which the sequences differ. This is called *Hamming distance*, and variants of it are extremely useful in the study of databases of sequences in an alphabet. Other examples are correlation distance, angle distance, and generalized Euclidean distance in high dimensions. The point of these measures is that if the distance is small, it should reflect that the points are similar, and if it is large, that they are dissimilar. The relationship between the distance function and the network is that the points that belong to an individual node are assigned to that node based on their similarity, so that the distances between points within a given node are relatively small. Also, if two points belong to nodes that are connected to each other, they will also be expected to be close to each other (and therefore similar). On the other hand, if two points belong to nodes for which one must traverse a large number of edges to move from one to the other, then one expects that their distance will be large and that they are therefore quite dissimilar. The network now serves as a map for the data set, giving a high level view of the structure of the data set. At a conceptual level, the construction of the network is relatively easy to describe. The starting point is a data set equipped with a metric and one or more projections to the usual line. Each projection is just the assignment of a number to each data point. The situation is illustrated by the image on the left below.



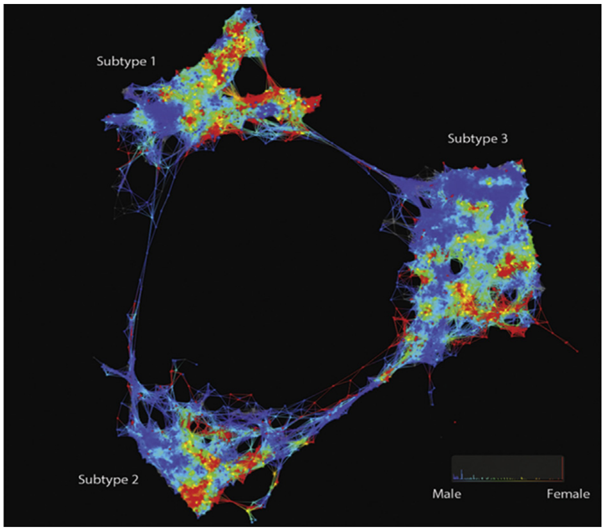
The data set is represented by the points on the left, and the axis below them represents the projection. The number assigned to each point is in this case simply its downward projection to the number line. The number line is covered by red and blue intervals, with the overlap of two such intervals colored in purple. Each red or blue interval determines a "bin" in the data set, by considering the set of points whose projection lies in the interval in question. A clustering step in each bin separately (using a predetermined clustering algorithm), and a node is created for each cluster that appears. Since the bins overlap, it is quite possible for two nodes to correspond to overlapping clusters within the data set, and when this occurs we draw an edge between the corresponding nodes. This construction creates the network on the right. It is clearly a simplified representation of the structure of the data set on the left. Of course, networks occurring in the analysis of real data sets are typically more complex than the example given above, but nevertheless yield substantial simplification.

Once the network has been constructed, one can apply standard graph layout algorithms to place it on the screen. Once this is done, it is possible to provide tools analogous to those used in graphics editors to allow a user to interact with the model by selecting regions and performing analyses on the subgroups in various ways, which will be described below. In this way, the methodology provides a very powerful method for analyzing data, rather than simply visualizing it.

For a more detailed description of the method, see [3,4].

Sample applications

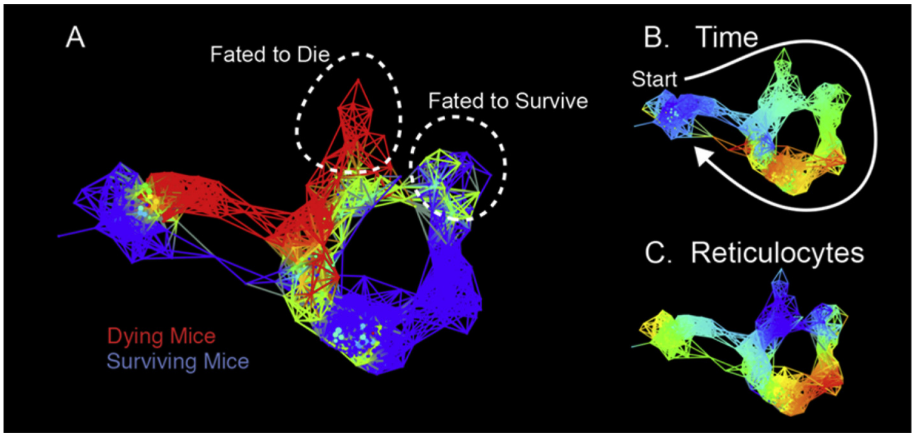
One application of this method has been to the study of type 2 diabetes, by a group at the Icahn School of Medicine at Mount Sinai [1]. They have studied a data set that includes electronic medical records information, as well as genotyping data, for a cohort of 11,210 patients.



The resulting network is displayed above. One sees that it consists of three large structures, connected by “thin wires”. The authors interpret the corresponding groups as being distinct, with groups 1,2, and 3 being enriched for diabetic nephropathy and diabetic retinopathy, cancer malignancy and cardiovascular diseases, and cardiovascular diseases, neurological diseases, allergies and HIV infections, respectively. Although the network construction used both EMR clinical data as well as genotyping, the authors also found interpretations of the groups in the genotypic information. This work is an example of how this methodology can support precision medicine, since one strongly expects that there will be different treatments based on different group memberships.

A second application is to the study of progression of infectious disease [2]. This work was conducted by the D. Schneider laboratory at Stanford University. They studied microarray transcriptomic data collected from mice infected with malaria, and constructed a topological network model based on it. The microarray data was collected daily for each individual.

The resulting network is shown below, with various colorings. One is by the time stamp, and it demonstrates that the passage of time corresponds to counterclockwise movement around the loop in the network. In addition, one can color each node by the fraction of member data points that ultimately died, and this is shown in the left hand network. The existence of such a model solves two problems related to the understanding of the progression of infectious disease, by creating a model that is based on the actual state of the individual rather than using time directly as a measure of the degree of progression. This is important because different individuals traverse the states in different amounts of time, and because for a particular individual, one might not know the time of



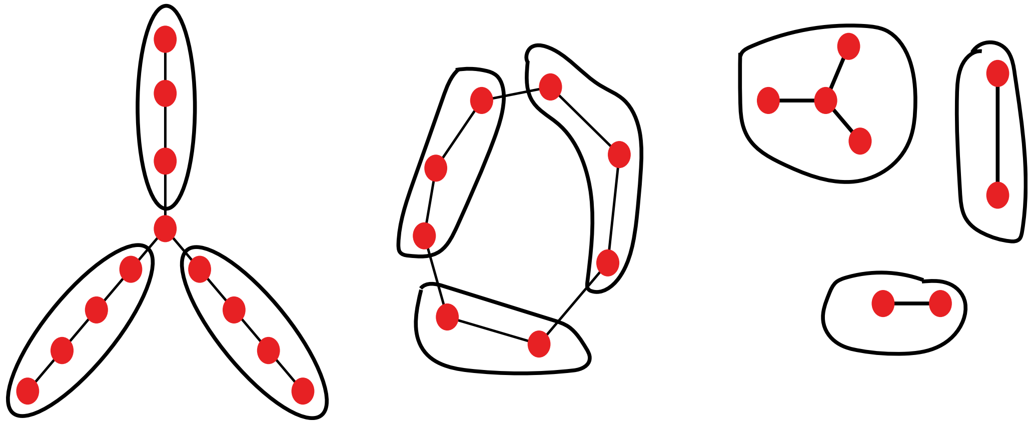
infection. Placement of the state of a given patient in the topological model will permit the assessment of where the individual is in the progression of the disease, without requiring information about the time of infection. Additional information in the model is likelihood to survive or die when one is in a particular state. The left hand network shows a substantial red region where the individuals whose state belongs to that region are very likely to die from the disease.

Other interesting applications have appeared in [5] and [6].

TDA as a modeling framework

One can interpret topological models as simply a new form of visualization. However, it is more than that; it is in fact a new modeling mechanism, with capabilities beyond the simple display of the network representation of the data on the screen. The capabilities include the following.

- **Selection of groups within the data set based on the geometric structure of the network**. Once the data set is displayed on the screen, one can select groups within the network (and therefore within the data set) using lasso or related similar, as is done in graphics editing software such as Adobe Illustrator or Photoshop. Such selection is often done based on coarse geometric features, such as flares, clusters, or loops.



- **Identify explanatory variables for a group**. Once a group is selected, and if the data set is defined by numerical variables, one can ask for the most explanatory variables for the group, in an appropriate sense. One such appropriate sense is to find the variables whose distribution on the subgroup is maximally different from its distribution on the entire data set, under some measure of similarity of distributions. One useful such measure is the Kolmogorov–Smirnov test. This kind of identification of correlated variables is often performed on an entire data set, but what this method permits is the application of it to subsets of the data set, which tends to give stronger identification of the relevant variables.
- **Represent functions on the data using coloring of the network**. Each node consists of a collection of data points, and if we have numerical values attached to each data point, we may color the node to represent the average value of the function over these data points. In this way, it is very easy to investigate the behavior of functions on the data set, and to understand maxima and minima. Coloring takes the place of graphing of functions for this kind of modeling.
- **Represent the distribution of categorical values using coloring of the network**. If we have a binary variable, we can color each node by a quantity representing the fraction of the data points that take the value 1. This generates "hot spots" within the network for concentration of points of this value.
- **Represent group membership using coloring of the network**. Given a data set and a group within that data set, one can color each node by the fraction of points in the collection corresponding to the node that belong to the group. For example, given a genomic data set, one might color by the membership in a particular ethnic group, and potentially observe that the distributions of different ethnic groups are distinct.
- **Hot spot analysis**. Using all three coloring methods described above, one can locate "hot spots", by which we mean regions in the network where, for example, survivors in a study will be much overrepresented.
- **Feature selection using networks of columns**. When a data set is given as a data matrix, where the rows are the data points and the columns are the "features", one can construct a network model on the columns rather than on the rows. In this case, each node consists of a collection of features, and each node consists of features which are similar. Understanding of the geometry of such a data set of features is extremely

useful. For example, if one colors by a density measure, one can identify that some features have many other features that represent information that is very highly correlated with it. This can mean that these features have a disproportionate effect on the analysis, and suggests how one might choose finite sets of features.

- **Split highlighting using multiple network models**. There are a number of choices involved in modeling a data set with TDA. One is a set of columns to consider, when the data set is defined as a data matrix. Another is the choice of similarity measure, and yet another is a set of functions (called "lenses") that are used to define the network model. Once obtains different networks this way, and it is very instructive to study both networks simultaneously. A useful way to study the network representations simultaneously is to select a group in one network and then to color other network using the selected group from the first.
- **Refined linear modeling using network segmentations**. Linear regression is a systematic procedure for predicting a numerical outcome variable based on a collection of other variables, called the independent variables. It is solved as an optimization problem, and there is typically a unique solution to the problem. One can use the error in the prediction as a quantity to color the network, and hot spots under this value will then indicate systematic errors in prediction. There are systematic methods for adding features to the linear problem that will improve the prediction. These features are constructed as numerical quantities attached to the nodes of the network. Overfitting is certainly a possibility here, and needs to be assessed.

Conclusion

Topological data analysis is a novel framework for modeling complex data sets, certainly including the ones that arise in the biomedical sciences. The framework includes a number of extremely useful capabilities, and produces non-conventional models that reflect heterogeneity within a data set as well as temporal behavior. We are currently at the beginning of the study of the applicability of these techniques. The applications produced so far are very encouraging, and suggest that many biological and medical problems can be usefully approached this way.

References

1. Li Li, *et al.*: **Identification of type 2 diabetes subgroups through topological analysis of patient similarity**. *Sci Transl Med* 28 Oct 2015, **7**:311ra174, http://dx.doi.org/10.1126/scitranslmed.aaa9364.
2. Torres B, *et al.*: **Tracking resilience to infections by mapping disease space**. *PLoS Biol* 2016, **14**:e1002436.
3. Carlsson G: **Topology and data**. *Bull Amer Math Soc* 2009, **46**: 255–308.
4. Lum PY, *et al.*: **Extracting insights from the shape of complex data using topology**. *Scientific Rep* 2013, **3**, 1236.
5. Nielson JL, *et al.*: **Topological data analysis for discovery in preclinical spinal cord injury and traumatic brain injury**. *Nat Commun* 2015, **6**, 8581.
6. Hinks TSC, *et al.*: **Innate and adaptive T cells in asthmatic patients: relationship to severity and disease mechanisms**. *J Allergy Clin Immunol* 2015, **136**:323–333.