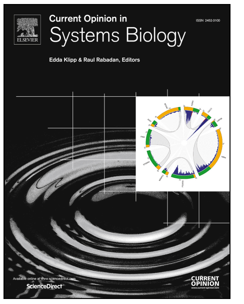
Accepted Manuscript



The Shape of Biomedical Data

Gunnar Carlsson

PII: S2452-3100(16)30032-4

DOI: 10.1016/j.coisb.2016.12.012

Reference: COISB 12

To appear in: *Current Opinion in Systems Biology*

Please cite this article as: Carlsson G, The Shape of Biomedical Data, *Current Opinion in Systems Biology* (2017), doi: 10.1016/j.coisb.2016.12.012.

The Shape of Biomedical Data

Gunnar Carlsson, Department of Mathematics, Stanford University and Ayasdi Inc.

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.

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