Persistent Homology Analysis of Biomolecular Data

By Guo-Wei Wei

Technological advances in the past few decades have fueled the exponential growth of "omic" data in biology. Understanding the rules of life from existing omic data sets, which offer unprecedented opportunities for mathematicians, remains an important mission of the field. Biomolecular structure-function relationships is a major rule of life, and recognizing this relationship is the holy grail of biophysics and a central issue in experimental biology.

Geometric modeling is vital to the comprehension of biomolecular structure-function relationships. It also bridges the gap between biological data and theoretical models, such as quantum mechanics, molecular mechanics, statistical mechanics, thermodynamics, and multiscale models.

However, geometry-based models are frequently inundated with too much structural detail and thus often computationally intractable. Topology provides the ultimate abstraction of geometric complexity by concerning only the connectivity of different components in a space and characterizing independent entities, rings, and higher-dimensional faces of the space in terms of topological invariants or Betti numbers.

To study topological invariants in a discrete data set—like atoms in a biomolecule—algebraic topology utilizes simplicial complexes under various settings, such as the Vietoris-Rips complex, Čech complex, or alpha complex. Specifically, a 0-simplex is a vertex, a 1-simplex an edge, a 2-simplex a triangle, and a 3-simplex a tetrahedron, as illustrated in Figure 1. Algebraic groups built on these simplicial complexes are used in simplicial homology to systematically compute Betti numbers for a given data set [7].

Nevertheless, traditional topology and homology are truly free of metrics or coordinates and thus keep too little geometric information to be practically useful for biomolecules. Persistent homology, a new branch of algebraic topology, embeds multiscale geometric information into topological invariants to achieve an interplay between geometry and topology [14]. It creates a variety of topological spaces of a given object by varying a filtration parameter, such as the radius of balls or the radius of balls of the level set of a real-valued function. As such, persistent homology neglects chemical and biological information during topological simplification and is thus not as competitive as geometry or physics-based representation in quantitative predictions. Element-specific persistent homology, or multicomponent persistent homology built on colored biomolecular networks, has been introduced to simultaneously retain chemical and biological information during topological abstraction [2]. This approach enciphers biological properties—such as hydrogen bonds, van der Waals interactions, hydrophilicity, and hydrophobicity—into topological invariants, rendering a potentially revolutionary representation for biomolecules [1, 3].

Rational drug design is an imperative life science problem that ultimately tests our understanding of biological systems. Designing efficient drugs to cure diseases is one of the most challenging tasks in the biological sciences. Multicomponent persistent homology plays a crucial role in hot-spot prediction, drug-binding pose analysis, binding affinity prediction, structure optimization, toxicity analysis, and pharmacokinetic simulation. For example, the integration of machine learning with feature vectors of colored graphs and multicomponent persistent homology provided the best free energy ranking for Set 1 (Stage 2) in D3R Grand Challenge 2, a worldwide competition in computer-aided drug design.1

References


Guo-Wei Wei is a professor of mathematicst at Michigan State University.