Is that a chromatin loop or not? Scaling Persistent Homology to Large Biological Datasets

Manu Aggarwal[1] and Vipul Periwal[1]

***Abstract*— Biological systems often have functionally critical spatial architecture. Morphological and experimental variability render it difficult to ascertain the significance of observed relative arrangements of constituents. Persistent homology (PH) is a mathematically rigorous method for computing topological invariants from data. The lack of scalability of preexisting PH algorithms prevented computation of the topology of the human genome. We devised an algorithm, Dory, that computed this topology within five minutes using only 6 GB of memory. Dory's computed robust features are validated with known biology in two applications, the role of cohesin subunits in chromatin loop formation and ligand interactions in proteins.**

***Index Terms*— Topological data analysis, algorithm, large data sets, human genome structure, protein structure**

I. Structure and function

Given data on the spatial locations of constituents of a biological system, a common structural pattern of functional importance is a region devoid of the constituents surrounded by a region of high density with constituents close enough to allow interaction, in other words, a hole. For example, holes in the spatial arrangement of human chromosomes in the form of chromatin loops are not random and play crucial roles in gene regulation. As another example, holes in the form of three-dimensional voids in proteins can be functionally important for ligand interaction or forming cages for transport. Hence, finding such patterns in the deluge of biological data sets can reveal functionally important structures.

II. Mathematical and algorithmic background

Identifying holes in a 3D embedding by visual inspection using the human eye is inconsistent and subjective. Persistent homology (PH) is an objective mathematical method for systematically assigning rigorous topological structures, homology groups, to variable experimental observations. Basis elements of the homology groups can be visualized as boundaries around holes when PH is computed for an embedding of data in Euclidean space. However, combinatorial complexity increases the cost of PH computation factorially with the number of points in the data set, making developing scalable algorithms an active area of research. Typical test data sets are limited to a few thousand points, and none of the published algorithms [1] were able to compute PH of the human genome, data sets with millions of points. Further, for scientific use, any computation of significant topological features should come with location information, but this is not provided by extant algorithms.

[1]Laboratory of Biological Modeling, NIDDK, NIH. Email: manu.aggarwal@nih.gov

III. Results and applications

We developed Dory. It is the only algorithm able to compute topology of the human genome at kilobase resolution. It also computes locations of significant topological features with improved precision.

A. Computational benchmarks

Dory computed topology of the human genome within five minutes using around 6 GB of memory on a standard laptop. Its efficiency is not limited to the human genome data set, and our benchmarks show that it uses less memory and run time in almost all of the test data sets generally used in the literature.

B. Cohesin subunits have different roles in chromatin loop formation

Hi-C experiments, based on proximity ligation, can give estimates of pairwise distances between different regions of the human genome in a nucleus at kilobase resolution. We used Dory to compute chromatin loops for the published Hi-C data [2]. Our results corroborate the conclusion that different subunits of the cohesin molecule play distinct specific roles in chromatin loop formation. Additionally, PH computation gives insights into the spatial scales at which topological differences emerge and loops are formed.

C. Topologically different protein homologs

We computed 3D voids in 187k protein structures and found instances where protein homologs have different numbers of significant voids. By computing their location, we showed that the most significant differences in voids are related to changes in ligand interaction. We also found that a protein of the growth arrest and DNA damage family, GADD45-γ, has significantly different topology between human and mice in its dimeric configuration. Experimental data has shown that a related protein, GADD45-α, has differential response to radiation between human and mice.

References

[1] U. Bauer, "Ripser: efficient computation of vietoris–rips persistence barcodes," *Journal of Applied and Computational Topology*, pp. 1–33, 2021.

[2] R. H. van der Weide, T. van den Brand, J. H. Haarhuis, H. Teunissen, B. D. Rowland, and E. de Wit, "Hi-c analyses with genova: a case study with cohesin variants," *NAR genomics and bioinformatics*, vol. 3, no. 2, p. lqab040, 2021.