Topology of the biomolecular world

organized by
Gunnar Carlsson and Guowei Wei

Workshop Summary

1. Overall Structure of the Conference

The conference proceeded using the standard AIM methodology, with lectures in the mornings, and group based work in the afternoons. The talks concerned many different situations in which topology can be used to study the biomolecular structure and function relationship, which is central in the understanding of rules of life. Topics included a general introduction to persistent homology and topological data analysis (TDA) more generally, the application of feature generation via persistent homology for deep learning and other methods, the geometry and topology of DNA and RNA, general discussions of various complex constructors and their applications, as well as dynamical systems. The discussion during and following these talks was quite lively.

2. Specific Foci

In this section, we will give the content of what was done in the groups. There was a wide variety of different topics covered.

(1) **DNA origami:** DNA origami were introduced in 2006 by Paul Rothemund and because the desired nano structures are easier obtainable than using smaller DNA based building blocks, this method for molecular assemblies has become very popular in many laboratories around the world. To date, there is no mathematical study of the types of DNA strands that assemble in a given structure, or the strand properties that impose certain geometries or topologies. The group addressed the question: given a topological surface (or a class of surfaces) what are the families of strands that produce those surfaces. We started with analysis of toy examples and obtained a small observation about the properties of the strands that provide a planar nano structure. We observed that more realistic situations need further sophisticated analysis. Persistent homology offers a potential tool for the systematical characterization of DNA origami.

(2) **Chromosome architecture:** We focused on the challenge of generating random "biological" structures so that the significance of computational/theoretical results can be quantified. This was motivated by the well-publicized 'fractal globule' model for chromosome architecture based on experimental Hi-C data that Mariel Vasquez addressed in her talk. We discussed the potential of Markov chain Monte Carlo (MCMC) sampling methods (such as in the RNA secondary structure combinatorial model) and recent results on random self-avoiding walks under confinement from Jie Liang's group. We debated what can be "known" experimentally, that is what

these chromosomal capture methods actually are able to measure, and sketched some possible research directions. The possibility of following up through a SQuaREs proposal is appealing.

(3) **Time varying persistence and clustering:** The group met for two afternoons to discuss the closely related problems of time-varying persistence and time-varying clustering. Discussions largely centered around the identification of a suitable formal model for multiscale, time-varying cluster structure. In the time-independent setting, dendrograms are the standard multiscale clustering model. There is an emerging consensus in some TDA circles that the natural extension of a dendrogram to the time varying-setting is a dendrogram-valued cosheaf over the real line. Under mild finiteness assumptions, this object is equivalent to a zig-zag of dendrograms. Such zig-zags may be rather complex combinatorially, and the problem of how to exploit them in practical applications them remains open. The group discussion focused on understanding this cosheaf perspective and explored the question of how one might be able to make use of this perspective in practice. The group came away with a better understanding of this fundamental problem. It became clear that, given the complexity of the objects involved, the appropriate way to study them likely depends on the nature of the data and the needs of the application. It will likely be useful in future discussions to focus on particular choices of data type and application. We also explored time varying persistence from the cosheaf perspective. Many of the same ideas carry over to the setting of functors taking values in a category of vector spaces, rather than in the category of sets.

(4) **Conformation spaces for cycloalkanes** The goal in this discussion group was to understand the topological structure of conformation spaces for cycloalkanes, which are molecules of the form C_nH_{2n} - a ring of n carbon atoms, each of which bonds to 2 hydrogen atoms. Currently the only well understood example is the conformation space for cyclooctane, C_8H_{16}, which has the structure of a 2-sphere with a Klein bottle attached along 2 circles. Two approaches were pursued: a) trying to understand a simpler example, C_6H_{12} - cyclohexane, by estimating the number of degrees of freedom, using known conformations (e.g., chair, boat, etc), their energy levels, and the chemically feasible switches between them. b) the second approach was computational; first constructing a dense sample of the conformation space (as a quotient of a subset of $\mathbb{R}^{9n}$), and then doing topological inference via persistent homology. It turned out that sampling the conformation space, even for small n, was harder than we had anticipated; after some literature review we arrived to a strategy but did not have time to implement it.

(5) **Fiberwise persistent (co)homology** The goal was to adapt the Leray spectral sequence of a fibration to the setting of discrete data, in order to compute persistent cohomology in degrees greater than 2. The main idea is that one can obtain maps from data to interesting spaces - e.g. to Grassmannians or projective spaces - and given an open cover for the target space, one can compute the persistent cohomology of the preimage of each open set. This can be treated as the input for a Leray-type chomology spectral sequence of $F[t]$-modules, whose infinity page – under suitable conditions – should be related to some form of persistent cohomology of the data. Progress: We clarified how to compute the successive quotients in the spectral

sequence in terms of the initial $F[t]$-generators, and got started on a small computational example (Klein bottle of natural image patches, fibered over the primary circle). This particular approach permits the iterative understanding of complex data sets. For example, in initial computations (which may involve absolute persistent homology) one may arrive at the understanding that the data has a natural map to a well understood space, such as a circle. The idea is then to use the fiberwise analysis over the circle to get to a quicker and more precise understanding of the space than one would obtain by applying homology directly.

(6) **Deep Learning and TDA:** Deep learning is a very popular methodology, which has been used very successfully for numerous of classification problems, especially for images. For example Guowei Wei and his group have used the method, applied to images constructed from persistence bar codes of biomolecular complexes to understand databases of targets and drugs. One problem with the methodology, though, is that there is a certain lack of transparency around what the calculation is doing, and how it is doing what it does. It is additionally vulnerable to so-called adversarial attacks. The group discussed various aspects of this problem. In particular, we discussed how to use features generated from persistence barcodes as input to a learning system, how to use the topology of data to understand the "guts" of what is going on throughout the calculations, and how to use homology of topological models such as mapper as features which can permit a deep learning system to select such models. The bulk of the time was used to discuss the possibility of building some very small models where one can visualize and understand what is happening. After a great deal of discussion, we settled on one such small model and did initial calculations with it. The results were interesting, and we agreed that we would continue our work on it as a joint project.