Geometric and topological methods in data analysis, with applications in biology and medicine: projects

Jean-Daniel.Boissonnat@inria.fr, Frederic.Cazals@inria.fr, Mathieu.Carriere@inria.fr

Academic year: 2023-24

Contents

- Each student chooses one project and returns a report to the teacher (see Contact at the end of each project before Sunday Dec. the 24th.
- Constraint: no two students can pick the same project.

1 Algorithms for loop closure: a comparison

In computational structural biology, *loop closure* is the problem concerned with the reconstruction of a protein backbone loop, given fixed endpoints. This problem is crucial since protein loops adopt motions constrained by fixed domains. The goal of this project is twofold. First, to compare two recent algorithms for this task: the first one is database dependent and uses reinforcement learning; the second one is *ab initio* and uses a global kinematic model to represent the protein backbone. Second, to ponder on extensions to go beyond the backbone modeling only.

Question 1. Summarize the algorithm from [1] (2 pages maximum). What is the role of reinforcement learning?

Question 2. Summarize the algorithm from [2] (2 pages maximum).

Question 3. The previous two algorithms need to cope with a difficult problem: the high dimensionality of the search space explored. Explain how this difficulty is dealt with.

Question 4. Backbone sampling algorithms omit side-chains, which are critical in protein structure. Explain what is a side-chain *rotamer*, and what is a *backbone dependent rotameric* library.

Question 4. Sketch a modification of the algorithms presented in [1] and [2] to incorporate conformations of side-chains.

Contact: Frederic.Cazals@inria.fr

2 Animating molecular machines

Question 1. Consider a non directed graph G, with n vertices and m edges [1]:

- Assume G is a tree: how many cycles does it have?
- Assume G is a connected graph with n vertices and $m > n - 1$ edges: how many *independent cyles* does it have? See also Fig. 1.
- Answer the previous question if the graph G has p connected components.

Question 2. Cycles in a graph have a vector space like structure: even if the number of cycles is exponential in the number of vertices, it is possible to compute a cycle basis. Define a *minimum cycle basis* (MCB) and summarize the algorithm from [3] to compute it (maximum 2 pages).

Question 3. Many molecular machines behave as *robots*, with rigid parts moving with respect to one another. One may also think about these machines as *molecular craddles* [4]. When two rigid regions linked by a *flexible linker* (a flexible loop as studied in class), sampling conformations of the loop makes it possible to explore the possible relative positions of the two rigid regions.

The situation is more complex when there are multiple rigid regions, and many linkers connecting them. Consider a models of such a machine as a graph: its n nodes are the rigid regions; its m edges are the linkers connected the nodes. Also assume that one has at its disposal a loop sampling algorithm such as [2].

Using the notion of MCB and a loop sampler, sketch h a method that could sample the relative positions of the rigid regions.

Contact: Frederic.Cazals@inria.fr

[1] See also `https://en.wikipedia.org/wiki/Cycle_basis` and the references therein.

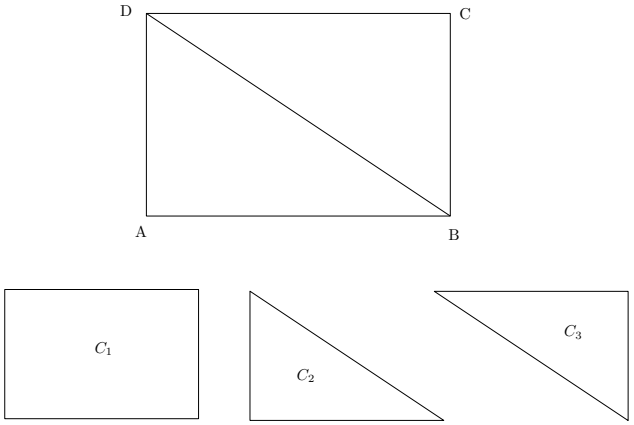

Figure 1: **A graph with four vertices and 5 edges.** This graph has 3 cycles but only has 2 independent cycles: adding any two cycles and counting edges modulo 2 yields the third cycle.

3 Topological Data Analysis and Computational Biology

Pick one of the following articles, which are applications of topological data analysis to biology.

- Cámara, Levine, Rabadán, Inference of ancestral recombination graphs through topological data analysis, PLOS Computational Biology, 2016.

  `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4988722/pdf/pcbi.1005071.pdf`

- Kuchroo et al., Single-cell analysis reveals inflammatory interactions driving macular degeneration, Nature Communications, 2023.

  `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10162998/pdf/41467_2023_Article_37025.pdf`

- Hoekzema et al., Multiscale methods for signal selection in single-cell data, Entropy, 2022.

  `https://www.mdpi.com/1099-4300/24/8/1116/pdf?version=1661340602`

Then, follow these steps:

1. Read the article carefully. Don't hesitate to look at the bibliography or any other resources if there is a definition or method that you find difficult.
2. Summarize the article. Explain what problem is being solved by the authors, what new theoretical and/or experimental contributions they bring to the state-of-the-art, and what biological discovery has been achieved.
3. Discuss the strong and weak points of the methods proposed in the article (such as efficiency, complexity, running time, etc). If applicable, implement the methods and provide some examples to illustrate your point.
4. Suggest potential improvements: what would you do to solve the weaknesses identified in the previous question?

Contact: Mathieu.Carriere@inria.fr

4 Computing a molecule

- Let $B_1, ..., B_n$ be n disks in the plane. The boundary of the union B of the B_i consists of arcs of circles whose endpoints are called vertices.

  Show that, although the circles bounding the B_i may intersect in $\Omega(n^2)$ points (i.e. a quadratic number of points), the number of vertices of the union of the B_i is $O(n)$, i.e. only linear in the number of disks

- Assuming that you have at your disposal an optimal algorithm to compute the convex hull of n points in R^3, suggest an algorithm to compute B and evaluate its complexity.

- . Same questions for a molecule, i.e. a set of balls (atoms) in R^3. For Question 2, you will assume to have at your disposal an algorithm to compute the convex hull of points in R^4.

  Hint : You can consult any text book on Voronoi diagrams and weighted Voronoi diagrams, e.g. Boissonnat, Chazal, Yvinec : Geometric and Topological Inference, chapter 4. https://geometrica.saclay.inria.fr/team/Fred.C

Contact: Jean-Daniel.Boissonnat@inria.fr

References

[1] A. Barozet, K. Molloy, M. Vaisset, T. Simeon, and J. Cortés. A reinforcement-learning-based approach to enhance exhaustive protein loop sampling. *Bioinformatics*, 36(4):1099–1106, 2019.

[2] T. O'Donnell and F. Cazals. Enhanced conformational exploration of protein loops using a global parameterization of the backbone geometry. *J. Comp. Chem.*, 2023.

[3] Kurt Mehlhorn and Dimitrios Michail. Minimum cycle bases: Faster and simpler. *ACM Transactions on Algorithms (TALG)*, 6(1):1–13, 2009.

[4] M. Simsir, I. Broutin, I. Mus-Veteau, and F. Cazals. Studying dynamics without explicit dynamics: a structure-based study of the export mechanism by AcrB. *Proteins: structure, function, and bioinformatics*, 89:259–275, 2021.