Grant ideas for NSF DMS Bio. Math, Data Science or Topology

Math Bio: https://www.nsf.gov/funding/pgm_summ.jsp?pims_id=5690 due November 15

Topology: https://www.nsf.gov/funding/pgm_summ.jsp?pims_id=5551 due Nov. 1

Data Science: http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=504687&org=DMS Due December 9

(Computational and Data-Enabled Science and Engineering in Mathematical and Statistical Sciences (CDS&E-MSS)

1. Goal: Find a polytope for linear programming that has split networks as its vertices.
    a. Could this be dual to the balanced minimal evolution polytope BME(n)?
    b. What is the connection between $\mathfrak{S}$_n and BME(n)? Note that boundary of BME(n) is equal to $\mathfrak{S}$_n for n=4, and that the dimension of $\mathfrak{S}$_n is the dimension of the boundary of BME(n). (See Prop. 5 of https://arxiv.org/abs/1607.06978).
    c. Note that certain facets of BME(n) correspond to certain incompatible splits. These are called "intersecting cherry facets" in https://arxiv.org/abs/1501.05536 , See Fig. 4. Might all split networks correspond to faces of BME(n)?
    d. Find distances (along non-unique geodesics) between split networks in the space CSN_n
    e. Bonus: hopefully the polytope will have a representation whose vertices' dot product with the difference vector allows linear programming—in order to find the best split network. Conjecture: Neighbor-net is a greedy algorithm for this linear program. (See https://arxiv.org/PS_cache/math/pdf/0702/0702515v2.pdf).
    f. Measure the accuracy of linear programming by geodesic distances between the calculated and predicted split network.
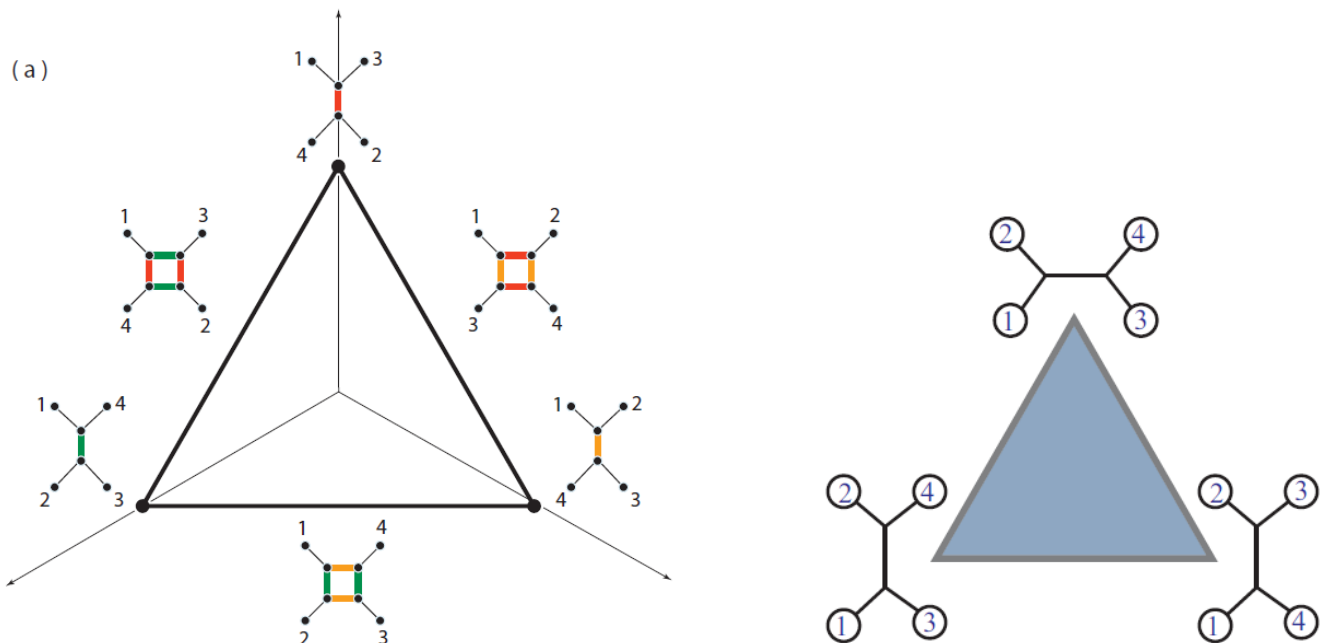
The facets of the triangle on the right do indeed correspond to incompatible splits, as shown on the left! (Note correction to Fig. 9 from https://arxiv.org/abs/1607.06978 on left: leaves 3 and 4 are switched on right edge.)

2. Goal: Prove the conjecture: Agglomerative algorithms for finding split networks, such as neighbor-net (splitsTree), can detect obstructions to trivial persistent homology.

As explained to me by a friend: Persistent Homology is found by inputting a distance matrix on a set of points. For each r > 0, you build a simplicial complex in which a set of points forms a simplicial complex if all their r-balls mutually overlap, and you compute the simplicial homology for every value of r. Of course it changes from being just discrete points (counted by H0) when r is small, to being a single contractible simplex once r is large enough.

Homology classes that exist over a large range of r values are viewed as "persistent" and likely to be a significant feature rather than an accident of sampling. For example, draw yourself a typical annulus in the plane and scatter some points around it representing samples. You can see that there will then be a range of r-values for which the balls link up and 1-dimensional homology will recognize the underlying circular structure, before r gets too big and some ball covers everything. There might be some other structures that arise briefly and then vanish again, but their "lifespan" will be dwarfed by that of the big ring.

People print out "bar codes" with one axis corresponding to r and bars parallel to the r axis showing where homology elements appear and disappear. If you see a significant-looking bar then you can delve in and try to interpret what its meaning is.

    a. Given a discrete space (points and pairwise distances), what is the relationship between:
    (1) the persistent homology, (often pictured as a "triangle-free network construction, which approximates the computed homology" as on the 3rd page of this short paper https://arxiv.org/abs/1511.01429) , and
    (2) the output of the neighbor-net algorithm? (The latter gives similar pictures that are related to the associahedron, as seen on page 7 of this paper: https://arxiv.org/abs/math/0702515).

    b. Googling for "persistent homology" and "neighbor net" turned up a very nice thesis as the first hit, but even there the question was not addressed. (The thesis downloads automatically from here: http://academiccommons.columbia.edu/download/fedora_content/download/ac:198974/CONTENT/Emmett_columbia_0054D_13361.pdf) See picture below.

    c. An example where this type of comparison is actually made is this paper on viruses: http://www.pnas.org/content/110/46/18566.full.pdf (especially see the appendix: http://www.pnas.org/content/suppl/2013/10/29/1313480110.DCSupplemental/sapp.pdf)
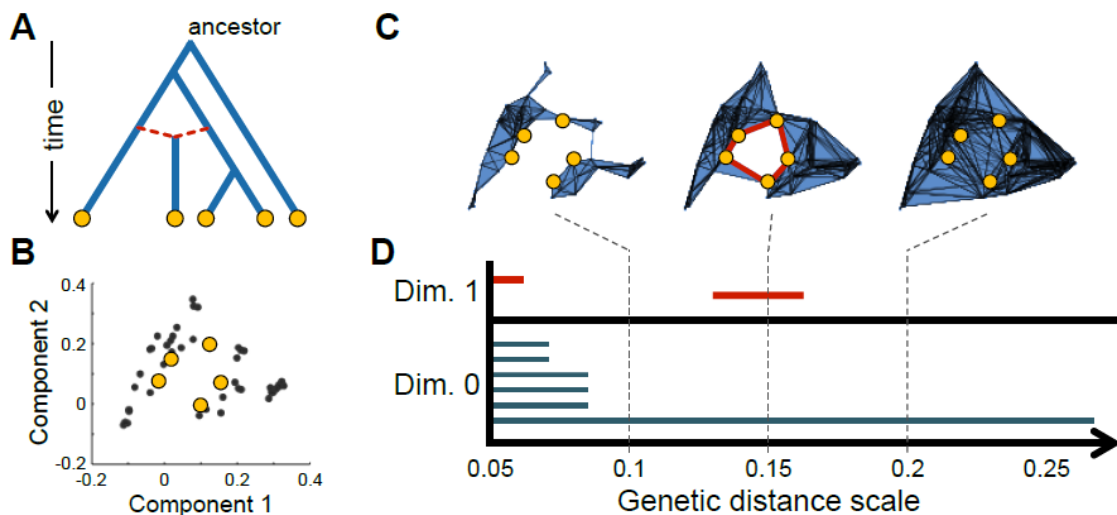

Figure 2.24: Applying persistent homology to genomic data. (A) An evolutionary genealogy including reticulation. (B) Data projected into 2-dimensions. (C) Construction of a filtered simplicial complex. (D) The resulting multiscale barcode diagram.

3. Goal: Improve our Branch and Bound (BnB) method of finding the Balanced minimal Evolution tree.
   a. Describe the Splitohedron. Do certain sets of incompatible splits correspond to faces?
   b. Prove convergence and consistency: we conjecture that the BnB method on the splitohedron for any input distance matrix will eventually converge to a binary tree; and that if the distance matrix approaches a true tree metric d_T (additive, obeys 4-point condition) then the output of the BnB method will approach the tree T.
   c. Prove safety radius: We conjecture that the safety radius of our BnB method on the splitohedron is ½.
   d. Study statistical accuracy of approximations to our method: time-limited BnB, forced component values, etc.
4. Goal: Use graph associahedra to tile a space of split networks.
5. Connections: The Permutohedron, the linear ordering polytope, the Birkhoff polytope are all related to the caterpillar tree facets of the BME polytope: same number of vertices, n!.
6. Connections: Acyclic orientations of a graph G make the vertices of the Graphical Zonotope for G. This polytope ZG is a cellular quotient of the graph associahedron KG; that is, the fan of KG subdivides the fan of ZG. An acyclic orientation of G, with its sinks and sources, can be seen as a time-directed phylogenetic network. Graph tubings, which are certain topological bases on the space of nodes, can be seen as level curves where the gradient follows the edge orientations. Graphical zonotopes tile Euclidean space.
7. Connections: A collection of nested or non-intersecting subsets of set X corresponds to a tree on the set of leaves X. A binary tree has smallest subsets of size 2. Thus a graph tubing yields a phylogenetic tree (unrooted) with leaves the nodes of G. For a path graph, this tree is planar using the ordering of the nodes. Thus a graph tubing is simultaneously a phylogenetic tree whose leaves are the nodes of G, and an acyclic orientation of the graph G. Note that two different tubings can give the same phylogenetic tree, and two different tubings can give the same acyclic orientation.
   a. Conjecture: Given an acyclic orientation on G and a phylogenetic tree (unrooted) with leaves the nodes of G, we can recover the graph tubing on G.
8. Connections: The permutoassociahedron projects to the BME polytope. Do permuto-graph-associahedra project to split network polytopes?
9. The vertices of the BME polytope are powers of 2. The vertices of the graph associahedra are differences of powers of 3.
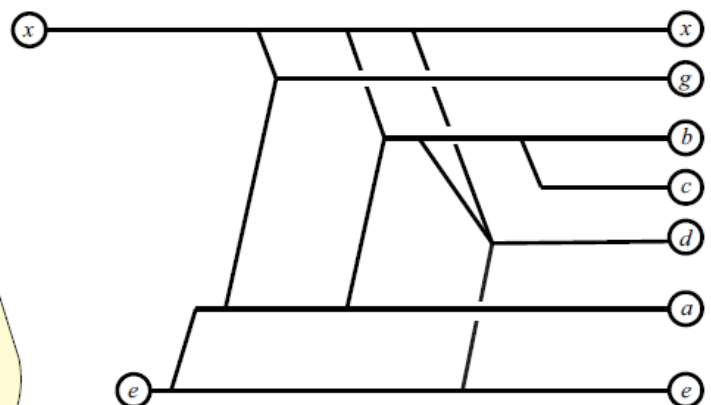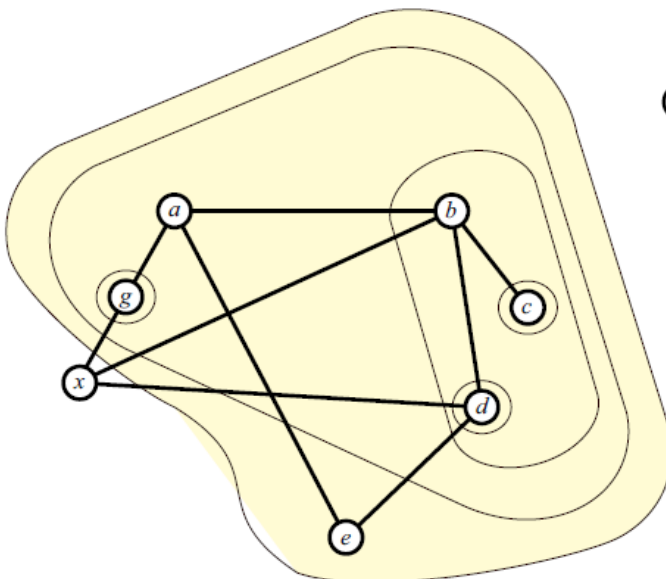
10. Goal: show that multitriangulations project nicely onto split networks, and that the (conjectural) polytope of multitriangulations projects onto the BME(n) polytope and/or the S_n complex. Check here: http://metodosestadisticos.unizar.es/~egc09/index_archivos/Trabajos/vincent.pdf and check here: https://arxiv.org/pdf/1009.5344.pdf see ref DKK here: https://www.mat.univie.ac.at/~slc/ .
    a. It sure seems that S_5 has a close relationship to BME(5). For instance, compare Figure 13 on page 13 of https://arxiv.org/abs/1607.06978 with Figure 7 on page 14 of https://arxiv.org/abs/1501.05536 . However, is this relationship direct (embedding?) or dual (embedding of a polar)?
11. Goal: use quick linear programming on small sets of data streams to show evolving phylogenetic trees as data flows into a sampler.
    a. It now occurs to me that the content could be used to infer "follower" relationships. It would basically be like textual analysis: two streams (tweets, posts) that shared text or sentiments could be assumed likely to both be copying from a single original. Of course the "rooting" must be done via extra information, but the tree would reveal which branching nodes are one step away from the root.
    b. Idea: An interesting approach would be to get the five streams by starting with a single stream and mutating it to get the others. You could also mutate the mutations, and so on. Then the tree should show this relationship!
12. Possible outline:
    a. What we know about the space of circular split networks.
    b. What we know about the BME polytope and splitohedron.
    c. What we know about multitriangulations.
    d. What we conjecture about their relationship.
    e. What we know about neighbor-net.
    f. What we know about persistent homology.
    g. What we conjecture about the relationship.
    h. What we know about graph-associahedra.
    i. What we know about zonotopes.
    j. What we conjecture about their relationship.
    k. Applications to data streams.