

Research Statement - Megan Owen

I am interested in combinatorics and computational geometry problems in computational biology, bioinformatics, computer science, and communication theory. My current research focus is on developing efficient computational methods for studying the evolutionary histories of organisms. Specifically, several distance measures have been proposed to quantitatively compare alternative phylogenetic trees for the same set of species. My dissertation research focuses on using combinatorics and computational geometry to design an algorithm to efficiently compute one such distance. I plan to extend this distance to phylogenetic networks.

A phylogenetic tree depicts how a set of organisms, such as primates, evolved from a single common ancestor. The leaves of a phylogenetic tree represent existing species, while the internal nodes represent their ancestors. Some applications of phylogenetic trees are in studying the evolution of viruses, such as HIV, which helps vaccine development; in studying biodiversity; in prediction of the function of genes; and in studying the evolution of manuscripts, language and culture [2]. There are many different algorithms to construct phylogenetic trees from biological data, but their accuracy can be affected by such factors as the underlying tree shape or the rate of mutation in the DNA sequences used. To compare these methods through simulation, or to find the likelihood that a certain tree is generated from the data, researchers need to be able to compute a biologically meaningful distance between trees [3].

In 2001, Billera et al. [4] introduced a novel distance measure for phylogenetic trees. They represented a phylogenetic tree with n leaves as a point in the space formed from a set of Euclidean regions, called orthants, one for each topologically different tree. Two orthants are connected if their corresponding trees are considered to be neighbors. Because the tree space is a CAT(0) space and thus has non-positive curvature, there is a unique locally shortest path, or geodesic, between any two trees. The length of this path defines the distance between the two trees.

In my thesis, I give a practical algorithm for computing the distance between any two trees within this tree space. My algorithm is derived from two main ideas. First, the properties of the geodesic imply that it is restricted to certain orthants of the tree space. We can model this section of tree space as a partially ordered set, where each element of the partially ordered set corresponds to one of the orthants in the tree space. This partially ordered set provides a means to enumerate all orthant sequences that could contain the geodesic, as each such orthant sequence corresponds to a maximal chain, or path, in the partially ordered set. We can efficiently construct this partially ordered set from the two trees, since each of its elements also corresponds to a closed set. The closure operator is defined using the idea of incompatible edges between the two trees, where two edges are called incompatible if they cannot exist in the same phylogenetic tree.

Second, given a particular orthant sequence, or particular maximal chain in the partially ordered set, we can find the length of the shortest path through that space by translating the problem into one of finding the shortest path through a subspace of a lower dimensional Euclidean space. By viewing this problem as a touring problem, in which regions must be visited in a certain order, and adapting a proof of Dror et al. [5], I showed that this new shortest path problem can be solved in linear time. Note that the general problem of finding a shortest path through a Euclidean space of 3 or more dimensions with obstacles is NP-hard [6], and the

sub-problems for which a polynomial algorithm exists have not been characterized. Thus, my solution is also an interesting result within computational geometry.

I then showed that any part of a shortest path through the space corresponding to the partially ordered set is also the shortest path in the space corresponding to the interval it induces in the partially ordered set. This implies that we can use a dynamic programming-based algorithm on the Hasse diagram of the partially ordered set to find the geodesic. This algorithm is polynomial in the number of elements in the partially ordered set. However, for certain pairs of trees, the size of the partially ordered set is exponential in the number of leaves in the tree.

My algorithm for computing the distance between two phylogenetic trees has proven to be practical when tested on a set of trees with 43 leaves each. I am currently working to analyze and improve the algorithm to provably run in polynomial time with respect to the number of leaves. I will release an implementation of this algorithm for use by computational biologists.

While phylogenetic trees have traditionally been used to represent the evolutionary history of organisms, not all organisms evolved from a single parent organism. For example, bacteria can transfer genes to other bacteria which are not their descendants through a process called lateral gene transfer (LGT). LGT is pervasive enough among early prokaryotes that Doolittle [1] argued that their evolution cannot be adequately represented by a phylogenetic tree, and suggests employing a phylogenetic network instead. Other lineage combination events such as hybrid speciation are also impossible to express using a tree structure. Generally, a phylogenetic network is a graph representing the evolutionary history of a set of organisms. Several mathematical models of phylogenetic networks, including directed acyclic graphs, have been proposed [7].

Many of the questions asked about phylogenetic trees can also be asked about phylogenetic networks. However, because phylogenetic networks are such a recent idea, work on their analysis is only in its infancy, making this research topic extremely exciting. In the next year, I plan to focus on the problem of meaningfully comparing phylogenetic networks. It is essential to have a valid comparison metric, because algorithms for generating phylogenetic networks are often tested by simulating genetic data from a given phylogenetic network, and using this data to try to reconstruct the original network. Researchers must then be able to quickly compare the reconstructed phylogenetic networks with the original one to fully evaluate their method.

Recently, Moret et al. [8] proposed the first distance metric for phylogenetic networks, based on tripartitions. An edge in a phylogenetic network divides the leaves into three sets, or a tripartition: those leaves descended only from that edge, those descended from that edge but also reachable from the root by a path not containing that edge, and all other leaves. However, Cardona et al. [9] claim that this distance measure is not a metric by exhibiting two non-isomorphic networks with distance 0.

I hope to extend the geodesic distance between phylogenetic trees that I have been studying to a similar space-based distance between phylogenetic networks and find a practical algorithm to compute it. This distance would improve on the tripartition-based one by naturally incorporating the lengths of the network's edges, which can represent, for example, the number of mutations between species. Distances in a space of phylogenetic networks should reflect biological reality, so the structure of the space will depend on when two networks are considered

biologically close. For example, shrinking an edge in the network could fill the role played by rotations in the tree space. I will investigate whether the non-positive curvature property also holds in the new space, because if it does, the shortest path between two networks is unique, giving a well-defined average between networks.

If this research is successful, I plan to investigate the relevance of this distance to other networks, such as social or information networks. Another application might be to networks representing the evolution of computer code developed under both open and closed source models. I am very excited to start this new research next year, and hope to also work on other combinatorial and geometric problems as the opportunities arise.

References

- [1] W. F. Doolittle. Phylogenetic classification and the universal tree, *Science*, 284: 2124-9, 1999.
- [2] R. Mace and C.J. Holden. A phylogenetic approach to cultural evolution. *TRENDS in Ecology and Evolution*, 20:116-121, 2005.
- [3] M.K. Kuhner and J. Felsenstein. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Mol. Biol. Evol.*, 11:459-468, 1994.
- [4] L. Billera, S. Holmes, and K. Vogtmann. Geometry of the space of phylogenetic trees. *Advances in Applied Mathematics*, 27:733-767, 2001.
- [5] M. Dror, A. Efrat, A. Lubiw, and J. Mitchell. Touring a sequence of polygons. In *Proceedings of the 35th Annual ACM Symposium on Theory of Computing (STOC)*, 2003.
- [6] J. Canny and J. Reif. Lower bounds for shortest path and related problems. In *Proceedings of the 28th Annual Symposium on Foundations of Computer Science (FOCS)*, 1987.
- [7] C. Linder, B. Moret, L. Nakhleh, and T. Warnow. Network (Reticulate) evolution: Biology, models, and algorithms. In *Proceedings 9th Pacific Symp. on Biocomputing (PSB)*, tutorial, 2004.
- [8] B. Moret, L. Nakhleh, T. Warnow, C. Linder, A. Thoise, A. Padolina, J. Sun, and R. Timme. Phylogenetic networks: modeling, reconstructibility, and accuracy. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 1:13-23, 2004.
- [9] R. Cardona, F. Rossello, and G. Valiente. Tripartitions do not always discriminate phylogenetic networks. arXiv:0707.2376v1, 2007. Alternative Introduction: