ENGINEERING SCIENCES & APPLIED MATH--FEATURED SPEAKER

NJMerge: A New Divide-and-Conquer Method for Scaling Phylogeny Estimation to Large Datasets

Tandy Warnow

PhD in Mathematics, UIUC

October 29, 2018 • 4 pm – 5pm

M416 (ESAM Conference Room), Tech

Phylogenies - also known as evolutionary trees - are graphical models that describe how a set of species have evolved from a common ancestor. Yet the estimation of these phylogenies is a challenging problem, and the most accurate of the phylogeny estimation methods are based on attempts to solve NP-hard optimization problems. One approach to improving the scalability of phylogeny estimation has been divide-and-conquer, in which the species set is divided into overlapping subsets, trees are constructed on the subsets, and then these subset trees are combined using a supertree method. These approaches have shown promise on moderately large datasets, yet the use of supertree methods, which typically also attempt to solve NP-hard optimization problems, limits the scalability of these approaches. In this talk I will present a new divide-and-conquer approach that does not require supertree estimation: we divide the species set into disjoint subsets, construct trees on the subsets, and then merge the subset trees using a new method, NJMerge, which uses a distance matrix computed on the full species set. We find that NJMerge provides dramatic improvements in running time without sacrificing accuracy and sometimes even improves accuracy. Furthermore, although NJMerge can sometimes fail to return a tree, the failure rate in our experiments is less than 1%. Together, these results suggest that NJMerge is a valuable technique for scaling computationally intensive methods to larger datasets, especially when computational resources are limited. This is joint work with Erin Molloy, a PhD student at the University of Illinois.

Note: Cookies will be served at 3:30