

Mathematical & Computational Biology Seminar

Organizer: Valerie Hower

Wednesday, 2:00–3:00pm, 939 Evans

Sept. 23 **John Kececioglu**, University of Arizona

Learning models for aligning protein sequences with predicted secondary structure

Accurately aligning distant protein sequences is notoriously difficult. A recent approach to improving alignment accuracy is to use additional information such as predicted secondary structure. We introduce several new models for scoring alignments of protein sequences with predicted secondary structure, which use the predictions and their confidences to modify both the substitution and gap scoring functions, and develop efficient algorithms for computing optimal alignments under these models. We also review an approach to learning the values of the parameters in these models called inverse alignment. We then evaluate the accuracy of these models by studying how well an optimal alignment under the model recovers known benchmark reference alignments. Our experiments show that using parameters learned by inverse alignment, these new secondary-structure-based models provide a significant boost in alignment accuracy for distant sequences. The best model improves upon the accuracy of the standard sequence alignment model for pairwise alignment by as much as 15% for sequences with less than 25% identity, and improves the accuracy of multiple alignment by 20% for difficult benchmarks whose average accuracy under standard tools is at most 40%.

This is joint work with Eagu Kim and Travis Wheeler.