David Rich
Research Methods
Sept. 11, 2018
Research Abstract

Abstract

In the field of computational biology, protein and DNA homology search is a task of critical importance. Two of the premiere software suites for this task are HMMER and MMSEQS. Both utilize a pipeline of a few different algorithms, which procedurally filter down the search space as the algorithms used become more complex. Both search pipelines begin with their own custom pre-filters, then use ungapped and gapped Viterbi algorithms, which eliminate a huge portion of unlikely matches. At this point, MMSEQS concludes and returns the alignments which are above a given threshold score. But HMMER has a final filter, called Forward-Backward. The Forward-Backward algorithm makes HMMER more accurate in the majority of test cases, but at costs to its runtime. Forward-Backward is implemented by summing over all possible paths through a HMM model. We propose to implement a new heuristic version of this algorithm. Rather than computing all cells of the dynamic programming matrix, this algorithm would prune paths which fell below a given factor of the strongest scoring path. By pruning low-probability paths, we should be able to dramatically reduce the runtime of the algorithm, with only minor costs to accuracy. We believe this could be used to either boost the accuracy of MMSEQS without dramatic harm to its runtime, or make HMMER run faster with little detriment to accuraccy.