Department Master's Proposal

David Rich

Statement of Purpose

In the realm of bioinformatics, the task of sequence alignment is a critical component of many research goals, such as gene annotation or homology search. HMMER [1] and MMseqs [2] are two of the most commonly used sequence alignment suites designed for such tasks. These tools both excel in different aspects: MMseqs generally performs the sequence alignment much faster, while HMMER is generally much more accurate.

Broadly speaking, both tools function in much the same way: by taking an entire query and target sequence model, and passing them through a pipeline of increasingly complex and stringent filtering algorithms. Both of these pipelines utilize the gapped and ungapped Viterbi algorithm. However, a major differentiator between the two is HMMER's inclusion of the Forward-Backward algorithm. While more computationally expensive, Forward-Backward has been shown to find noisy alignments with low identity which algorithms like Viterbi alone cannot.

As a compromise, we propose a heuristic version of the Forward-Backward algorithm, which we have termed the Cloud Search. By pruning the search space of Forward-Backward, we hope to drastically reduce the search space, and thus the runtime of the algorithm. Traditionally, Forward-Backward computes the sum of all probability scores for all possible alignments through a given model. By contrast, at each step through the model, Cloud Search will compute the maximal scoring state; any states with a score below some ratio of the max score will be pruned and all paths passing through this point will be calculated no further. Due to state dependencies, this will be accomplished by navigating the dynamic programming matrices in an anti-diagonal fashion, as proposed by Wazniak [3]. Through this judicious pruning of only very low probability areas, we believe that these savings should come at minimal cost to accuracy. This heuristic pruning method has been explored in different algorithms like beam-search[4] and x-drop (used by BLAST) [5]. For a comparative analysis, I will benchmark my algorithm integrated into the HMMER pipeline against current HMMER, MMSEQS, and other state of the art alignment software using a variety of datasets with queries of varying length and identity.

Timeline

For my timeline, I plan to do the following:

- November – December 2019: Complete my initial code implementation.
- January – March 2020: Completion of testing and iterative code optimization and the writing of my thesis and paper.
- March – May 2020: I plan to have my research concluded, my thesis written and ready to defend.

Results/Deliverables

At the conclusion of my research, I will produce the following deliverables:

- For my software product, I will have a version integrated into a branch of the HMMER pipeline.
- My research results will be compiled into a thesis and a paper, which will be submitted for publish to a scientific journal.

References

[1]
Eddy SR (2008) A Probabilistic Model of Local Sequence Alignment That Simplifies Statistical Significance Estimation. PLoS Comput Biol 4(5): e1000069. doi:10.1371/journal.pcbi.1000069

[2]
Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology* **35,** 1026–1028 (2017).

[3]
Rognes, T. (2011). Faster Smith-Waterman database searches with inter-sequence SIMD parallelisation. *BMC Bioinformatics*, 12(1).

[4]
Ploetz, Thomas & Fink, Gernot. (2003). Towards Faster Profile HMM Evaluation.

[5]
Frith, M., Hamada, M. and Horton, P. (2010). Parameters for accurate genome alignment. *BMC Bioinformatics*, 11(1).