IEValue for MMSeqs2

Steven Lee, Alejandro Matero, Harshil Patel, Ben Sauerwald, Adam Waleed, & Jacob Wong
Drexel University, Dept. of Electrical & Computer Engineering

May 24, 2022

Contents

Abstract

Sequence Similarity Searching is an important method that allows for searching input sequences against sequence databases. An overwhelming problem regarding the technique results from the evolution of technology in the field that increases the amount of data on hand drastically year over year. To counter this problem, a prior method based on an incremental e-value has been studied by iBLAST. The project and corresponding research is regarding to applying an incremental e-value correction to MMSeqs2. This will be done by sequence searching utilizing MMSeqs2 and implementing the incremental e-value equation that is performed in the iBLAST paper (Dash, et al).

Introduction

Sequence Similarity Searching is a conventional technique of searching sequence databases by using alignment to a query sequence. This method statistically evaluates how well the query matches the databases, which can lead to deducing the homology between sequences and transferring information to the query sequence. The most widely used algorithm is BLAST - basic local alignment search tool (Altschul et al., 1990). Sequence similarity is an effective and reliable strategy for identifying homologs, however, many similarity searchers seek further conclusions (Pearson, 2022). Algorithms like BLAST, FASTA, and HMMER are often used to determine related sequences with similar functions.

Through the years, the exponential increase in research and investigation in the genomics field has led to the expansion of datasets due to the incremental addition of newly found data. In turn, as new data is introduced the results for a specific query will change over time. The changes in the dataset force scientists to rerun BLAST against the updated datasets in order to achieve the revised results. This extra work is not only time-consuming, but it comes with wasted execution time, money, and computational resources (Dash, Rahman, Hines and Feng, 2022). This problem is addressed by iBLAST, an incremental BLAST of new sequences via automated e-value correction. iBLAST uses past results to conduct the same query search but this time only on the newly added data. Subsequently, the results from both searches, old and new, are converged to produce updated search results. The main distinction between NCBI BLAST and iBLAST is the ability of the latter to enable efficient biological discovery at a much

faster speed with a substantially reduced computational cost (Dash, Rahman, Hines and Feng, 2022).

Currently, despite some sequence alignment programs like HMMER and DIAMOND being successful at achieving similar output to NCBI BLAST at an improved computational speed, the problem of large-scale sequence alignment being a computational burden still remains. The introduction of incremental value aims to find a solution by trying to save both time and money in large-scale projects using subsequent merging of results from two separate databases, thus allowing recycling of previous results (Dash, Rahman, Hines and Feng, 2022).

1. Overview of the Project and Goals

The main goal of the project is to implement an incremental e-value similarly to how it was done in the reference paper for iBLAST (Dash, et al). The incremental e-value is an extremely efficient method regarding sequence searching since the entire database doesn't need to be updated constantly as many different inputs are added. The implementation of the incremental e-value was successfully done on BLAST. To perform this implementation in accordance to MMSeqs2, the program will be run for searching utilizing the input data provided in the directory in the Picotte cluster and then applying the Karlin-Altschul Equation to implement e-value correction.

Week	Status	Milestones
May 9-13	✓	Install and test MMSeqs2
May 16-20	✓	Perform initial test with project data
May 23-27	IP	Correct errors and implement E-value equation
May 30 - June 3	TBC	Perform final test/Compare Results
June 6-10	TBC	Conclude on Results

Figure 1: Timeline for deep learning using recursive neural networks in metagenomics.
(legend: X – completed; IP – in progress; TBC – to be completed)

2. Obstacles Encountered

Aside from the main issues of determining how to use the base of mmseqs2, the group was met with various other obstacles so far throughout the process. The first being the confusion on whether or not we are using mmseqs2 for clustering, alignment, or something else (which we later determined for our purposes will be used for searching). The next being clarity in the placement of data on the picotte server. The input data paths given were supposed to be in various locations however all locations aside from one were empty. The next hurdle was figuring out the input file type. The input files are required to be a FASTA file with the group being provided 5 FASTA files sequences. Finally, another point of clarity was needed when determining which databases were the query database and which were the target. This was resolved in determining the batch-0 is to be used as the query database.

3. Current state of the project

3.1 Current Implementation and Next Stages of the Project

The data provided contains 5 FASTA file sequences, batch-0 and batch-4-1 through batch-4-4. These files contain the metadata and sequences that we will be using to search through. From the naming conventions, batch-0 is used as the query database and the other batches as the target databases. MMSeqs2 has a search function that allows for alignment between a query and target database. The resulting database is a partitioned table that contains the E-value score of each match. The goal of this project is to implement the Karlin-Altschul equation for E-value correction. It is unknown at this time how the correction will change the results and so a comparison will need to be made.

$$p(x \geq S) = 1 - e^{-E}$$ [Karlin-Altschul Equation]

3.2 Preliminary results

MMSeqs2 is an algorithm located in a github repository. The first goal of installing the algorithm and allocating the data to be searched to its directory has been achieved. A test of the search functionality was performed using the supplied data from the MMSeqs2 repository. The

resulting E-values were then put through the correction equation. The resulting values were either a positive 1 or 0. The data from the project was then put through the search algorithm. batch-0 was used as the query database and batch-4-1 was used as the test target database. It seemed to have worked the same but the resulting E-values were all extremely small. When placed into the correction equation they all resulted in a 0. Upon inspection of the example data from MMSeqs2, it is widely different from the provided batch files. Further investigation is needed to determine if the correct query and target databases are being used.

4. Future Work

To complete this project, the data will need to be manipulated correctly. In order to achieve this the following goals are required to be met.

- Determine correct query and target databases for search
- Perform uncorrected search results on all target databases
- Perform E-value correction and re perform search incrementally.
- Compare corrected and uncorrected results
- Conclude on the effectiveness of the E-value correction

Completing the previously mentioned requirements will allow for a proper implementation of the incremental e-value regarding MMSeqs2, similarly to how it was done regarding BLAST. It is expected that with proper implementation, the preliminary results of the e-value will then be updated from 0. An evaluation from there will then be taken to measure the validity of results alongside the performance.

References

[1] Altschul, Stephen F., Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman. "Basic Local Alignment Search Tool." *Journal of Molecular Biology* 215, no. 3 (1990): 403–10. https://doi.org/10.1016/s0022-2836(05)80360-2.

[2] Pearson, W., 2022. *An Introduction to Sequence Similarity ("Homology") Searching*.

[3] Dash, S., Rahman, S., Hines, H. and Feng, W., 2022. *iBLAST: Incremental BLAST of new sequences via automated e-value correction*.