Steven Lee, Alejandro Matero, Harshil Patel, Ben Sauerwald, Adam Waleed, & Jacob Wong
June 10, 2022
ECES 450/650
Final Report

Incremental E-Value For MMSeqs2

Abstract

The Incremental E-Value For MMSeqs2 project that was conducted was essentially an attempt at recreating a similar method that was previously implemented by BLAST, another search algorithm/tool. 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, which increases the amount of data on hand drastically year-over-year. To counter this problem, BLAST implemented an incremental e-value, which ultimately allowed for the opportunity to search against the newly added portion of the database as opposed to loading and searching through the entire database every query as new data was added. The project and corresponding research regards the application of 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).

The actual execution of the project required the conduction of an MMSeqs2 search, which must have a query and target database to carry out the task. A full sequence was queried against various size databases ranging from 20-100% (full database). The e-values of all of the hits are then extracted and can be analyzed. A high-level interpretation of the results essentially presented an observation that as the target database grew, the number of search hits increased, also concurrently decreasing the e-value. The results obtained accurately corresponded with what was hypothesized by the group prior to execution. With the presentation of a few irregularities in the output data, further analysis was necessary to be conducted to provide a clear-cut understanding of implementing an incremental e-value for MMSeqs2, which can be found in the *Discussion*. As mentioned prior, research regarding incremental e-values holds great significance due to the fact that sequence databases will only continue to grow at a faster rate from here.

Materials & Methods

Many-against-Many Sequence searching (MMSeqs2) is an open-source software/searching tool written in C++. MMSeqs2 was the tool that was utilized in the project to practically recreate the methodology that was constructed through BLAST prior. An MMSeqs2 search needs a query and a target database, for the query to search through. The input data that was to be analyzed was found on the Drexel University Picotte cluster through the path /ifs/groups/eces450650Grp/data/incremental_data .

Since the effect of the Karlin-Altschul e-value correction equation is being analyzed in the project, the full sequence will be utilized as the query and successively searched through each percentage (20-100%) of the whole database. Proper execution of MMSeqs2 creates resultant databases from its search. From there, the outputted data can be even further reduced to a best results database. This presents a sorted data output by e-value database. The database can then be converted to a partitioned table giving the e-value for all the hits obtained through the MMSeqs2 searching algorithm.

To complete all of this, the following procedure was implemented. After installing the MMSeqs2 codebase, the query database was created using the "createdb" command. Similarly the target database was created using the same command. The target database can then be indexed to allow for a fast read-in using the "createindex" command. This becomes more helpful if the target database is kept the same for successive searches from different queries. However, for the execution conducted in the project, the query database is kept the same and the target database is the one altered with each search. The search command was then used to create the list of resultant databases. This ultimately sorted the recently acquired output data into the best results databases.

Finally, the best results databases are converted to the BLAST tab formatted file using the "convertalis" command. From the MMSeqs2 user guide, the tabs of the file are as follows: (1,2) identifiers for query and target sequences/profiles, (3) sequence identity, (4) alignment length, (5) number of mismatches, (6) number of gap openings, (7-8, 9-10) domain start and end-position in query and in target, (11) e-value, and (12) bit score. After each search and tabbed file creation, the output data required was saved separately to allow for data analysis of each incremental search.

Results

Due to the fact that the sequence databases are sufficiently large, the search result hits are numbered in the tens of thousands. Given this, only discrete values will be shown with the unabridged data being attached through the github repository. As will be shown in the analysis, an e-value score of zero was selected as the criteria to be a hit. Additionally, a top hit was selected from each incremental search. The reasoning behind the selection will be detailed in the analysis.

Table 1: Incremental Search Hits vs. Corrected Hits

Portion	Hits	Corrected Hits
20%	541	30794
40%	1358	36426
60%	1353	36413
80%	1366	36407
100%	1488	38443

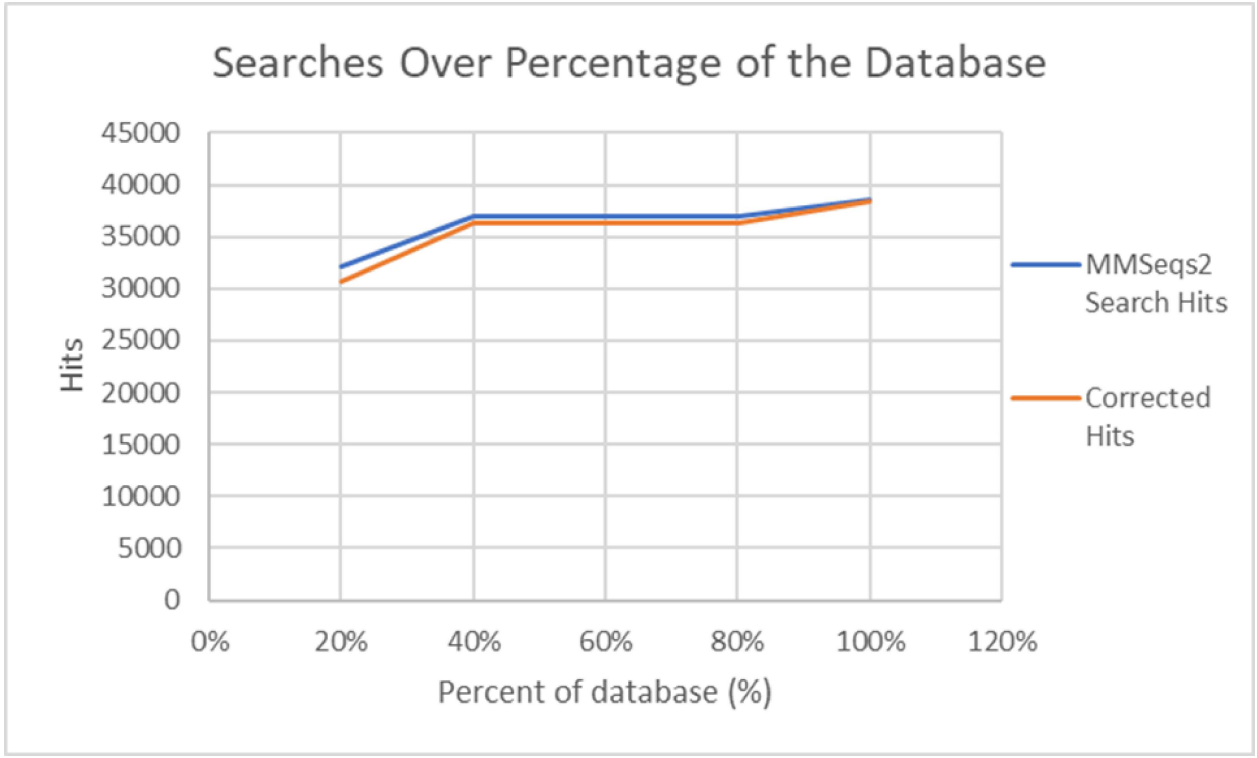

Figure 1: Database Size vs Search Hits & Corrected Hits

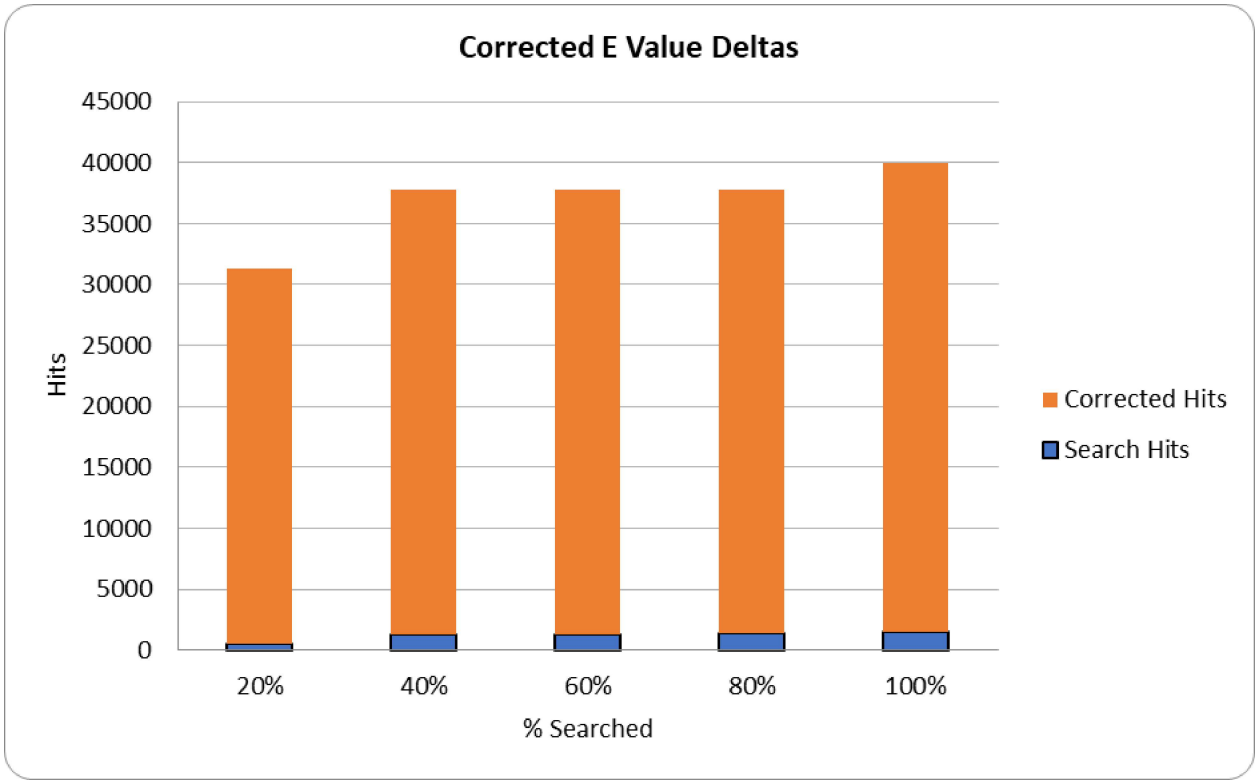

Figure 2: Incrementally Searched Corrected E-Value Deltas

Table 2: Top Hit of Each Search

20 %		40 %		60 %		80 %		100 %	
Query	Target	Query	Target	Query	Target	Query	Target	Query	Target
d1r27c2	d1r27a2	d1ti4k2	d1vleu2	d1vldu2	d1ti2e2	d1vldu2	d1ti6i2	d1ti2g2	d1ti4e2
Bit Score		Bit Score		Bit Score		Bit Score		Bit Score	
2236		1593		1593		1593		1593	

As stated before, an e-value as close to zero is the best possible value. When performing the data analysis it was noted that the correction equation produced significantly more zero e-value hits than the original search. Given this data point, it was selected that only zero e-values be constituted as hits to show the implementation of the equation. The correction equation should only decrease the value or maintain it the same. When looking at the data, although it does produce significantly more hits, it does follow the search algorithm. From the 40% to 60%

search, hits went down and the corrected hits followed suit. Since there was an exorbitant amount of zero e-values, other criteria must be analyzed.

The table above gives both bit score and number of mismatches and both of these were used to determine the top hit. The table was sorted by highest bit score and the highest value was selected that exhibited zero instances of mismatch. They are displayed in *Table 2* and raise some interesting questions. Only twice did the same query sequence appear in the top hit, and it was matched to different target sequences. The bit score gives a rule of thumb of homology of the sequences. Each top hit bit score was high enough to be considered significant. However, it is important to note that the highest bit score with no mismatches was identical for the 40-100% searches. This could be due to the nature of the MMSeqs2 search. As the database was incremented through, it chose better and better sequences to produce the best match.

Discussion

Overall, the project can be deemed a success. A recreation of what was implemented by BLAST, regarding an incremental e-value for sequence searching, was able to be adapted to another search algorithm by the name of MMSeqs2. At a high-level abstraction, this was able to be executed by running a search on MMSeqs2 with a query and target database and then extracting the e-value and correcting it with the Karlin-Altschul equation.

From the execution of the project, one of the main ideas found was that as the target database increased, search hits increased as well, while also decreasing the e-value. In a broad sense, this was expected by the group prior to the conduction of the project. A decreasing e-value practically minimizes the chances of randomness involved in search sequence hits. Ideally the e-value is wanted to be 0 to maximize the probability of accuracy in the sequence search.

An interesting observation that occurred was the leveling of search and corrected hits between the 40-80% database shown in *Figure 1*. This was an observation that required further analysis. The bit score and corresponding number of mismatches were then analyzed. The results concluded that each top hit bit score was high enough to be considered significant. There was also an identical highest bit score with no mismatches for the 40-100% of the database.

To further progress the research centered around an incremental e-value, a recreation of the project that was conducted could be implemented onto a new search tool and a full comparison between iBLAST, e-value for MMSeqs2, and the new tool could be analyzed.

References

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

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.

Pearson, W., 2022. *An Introduction to Sequence Similarity (“Homology”) Searching*.