Drexel University – ECES450

iDiamond Alignment Study

Integrating Diamond with iBLAST Incremental Search Framework

Justin Morgan, Ayaz Noor, Martina Ross, Vishnu Sekar, Sean Reilly

3-19-2024

Table of Contents

ABSTRACT

As genetic databases expand exponentially, the efficacy of traditional tools like the Basic Local Alignment Search Tool (BLAST) diminishes due to prolonged query times. In response, iBLAST and its variants, such as DIAMOND, have emerged, leveraging previous alignments to expedite searches. These tools have significantly reduced query times from days/hours to mere minutes, revolutionizing research efficiency (Buchfink, Reuter, & Drost, 2021). This study aims to assess not only the time efficiency of Diamond but also its efficacy compared to BLAST. The research involves partitioning the Astral Dataset and aligning sequences using Diamond as the database size increases. Subsequently, output files are utilized to simulate BLAST alignments through iBLAST code. By merging the output files, researchers gauge the number of successful hits obtained from each alignment. This method can also analyze what alignment was unique to the database size and what was shared between others. Out of the 1426 sequences, Diamond yielded 4 resultant hits, while the iBLAST simulation produced 17 hits. Through this comprehensive approach, the research endeavors to provide insights into the comparative performance of Diamond and BLAST across varying database sizes. Such findings are crucial for optimizing sequence alignment protocols in genetic research, ensuring swift and accurate analysis amidst the expanding genomic data landscape.

METHODS AND MATERIALS

Design

The objective of this study is to investigate the Diamond Alignment tool, replicate BLAST, and compare the results in terms of runtime and accuracy. The study begins with an exploration of BLAST, a historically reliable method for querying DNA sequences. However, with data sets growing larger, longer run times became an issue, leading to the creation of iBLAST. The study is structured into three tasks, each aimed at achieving a specific objective. The results of each task are essential for subsequent sections.

The study measures the efficiency of the alignment tool by analyzing the time taken to complete a query. Additionally, accuracy is assessed based on the number of alignments returned. Merging these alignment output files simulates BLAST and comparing them against Diamond alignments provides a measure of accuracy. This methodology enables a comprehensive evaluation of the performance and effectiveness of Diamond Alignment compared to traditional methods like BLAST and its improved version, iBLAST.

Materials

- Picotte high performance computer cluster – available through Drexel University
- Access to the ASTRAL SCOPe database – this database is available on Berkeley Lab's website.
- DIAMOND high performance tool – available from respective GitHub repositories
- iBLAST – available from GitHub
- Git and GitHub – storage repository of all documentation and code
- Statistical software: Python w/SciPy and NumPy

Experimental Procedures

TASK 1

Task 1 involved the application of the Diamond Alignment Tool on the Astral Data set, initially divided into ten batches. The Diamond program, in this context, searches a sequence database with a query sequence to return sequences from the target database that show significant alignment, known as hits. Within these alignments, numerous pairwise locally optimal gapped alignments exist, referred to as high scoring pairs (HSPs). For the task, nine out of these ten batches were sequentially merged into databases in a cumulative manner: the first database contained data from the first batch, the second database included data from the first two batches, and so forth. The tenth batch served as the query input, processed through each of these incrementally compiled databases using Diamond, to generate records of alignment matches. Additionally, a separate file was produced to document the execution time of Diamond. Originally, the output was in TSV format; however, to ensure compatibility with future tasks, the output format was changed to XML through code adjustments.

TASK 2

During the second phase of the project, we utilized the Diamond Alignment Tool for sequence alignment, preparing our data for subsequent analysis with iBLAST. This preparation was crucial for ensuring compatibility between the two tools, especially for statistical corrections. iBLAST facilitated the application of the Spouge statistic correction automatically during our analyses, which adjusted the expectation values (e-values) of our sequence alignments. E-values are critical for evaluating the significance of sequence alignments within protein databases; they estimate the likelihood of an alignment occurring by chance, with lower e-values indicating more significant matches (Rahman, Hines, & Feng, 2021).

For the experimental setup, we executed two distinct runs of Diamond alignments, utilizing the tenth batch of our dataset as the query in both instances. The first run employed a database created from batch one, and the second run utilized a database derived from batch two. After obtaining the alignment results in XML format from both runs, we merged these using iBLAST, which automatically applied the necessary statistical corrections. This merging process allowed us to inspect the effects of incremental database updates on the alignment process, particularly focusing on variations in the corrected e-values. We analyzed these effects by comparing the consolidated outcomes with those achieved when the tenth batch was queried against a combined database of batches one and two (taken from the results of Task 1), aiming to assess the incremental update approach’s efficiency.

TASK 3

We conducted an experiment using Diamond to run batch ten queries against nine separate databases, much like in Task 1. However, these databases only consisted of one batch each, so database one would only contain batch one, database two would only contain batch two, and so on. This generated nine distinct result sets (Result1 through Result9). We progressively merged these results using iBLAST, starting with Result1 and Result2 to form Result12, and continued this process until all results were combined into Result123456789. This result represented an alignment against an incrementally updated database. We compared Result123456789 to the last result of Task 1, where batch ten queries were aligned against a single database comprised of batches one through nine. This comparison allowed us to study the nuances of incremental

database updates and their effects on the alignment process, especially in terms of alignment scores and the statistical significance as indicated by e-values.

Data Analysis

Metrics:

- Time to Run the Alignment Sequence
- Amount of Hits between each Diamond Sequence

In Task 1, the time in seconds required to query the sequence was measured, along with the number of hits returned, serving as a measure of accuracy for that alignment. For Tasks 2 and 3, the metrics included both the total number of hits and unique hits. Each alignment output was merged using iBlast code, and the final hit count was determined. Unique hits specific to each fold were tallied, as well as the total count of hits when all files were combined. To assess accuracy, these figures were compared against the number of hits from a Diamond alignment of equal size. This comparative analysis provided insights into the accuracy and effectiveness of the alignment processes.

RESULTS

In Tasks 1 to 3, our methodology involved documenting the outcomes of each database query in XML format, establishing a hierarchical structure that commences with each iteration phase. Within these iterations, detailed insights into each query, drawn from the comprehensive dataset of 1426 sequences in batch 10, were systematically organized. Upon DIAMOND's identification of an alignment—termed a 'hit'—between a query sequence and its database counterpart, a 'hit block' was instantiated within the iteration framework. This section meticulously recorded essential details such as the hit definition, the sequences of both the query and the hit, HSP e-values, and additional pertinent HSP-related metrics. This arrangement of data is pivotal for conducting nuanced comparative analyses across datasets.

In Task 1, we passed batch 10 through each of the cumulative databases, with the following results. Each database provided between two and five hits. Most iterations in a file did not produce hits, and each iteration that did, did not produce more than one. Most hits were recurrent throughout the files (noted in the GitHub as matched hits), however some databases produced unique results. This can be seen in the figure shown below.

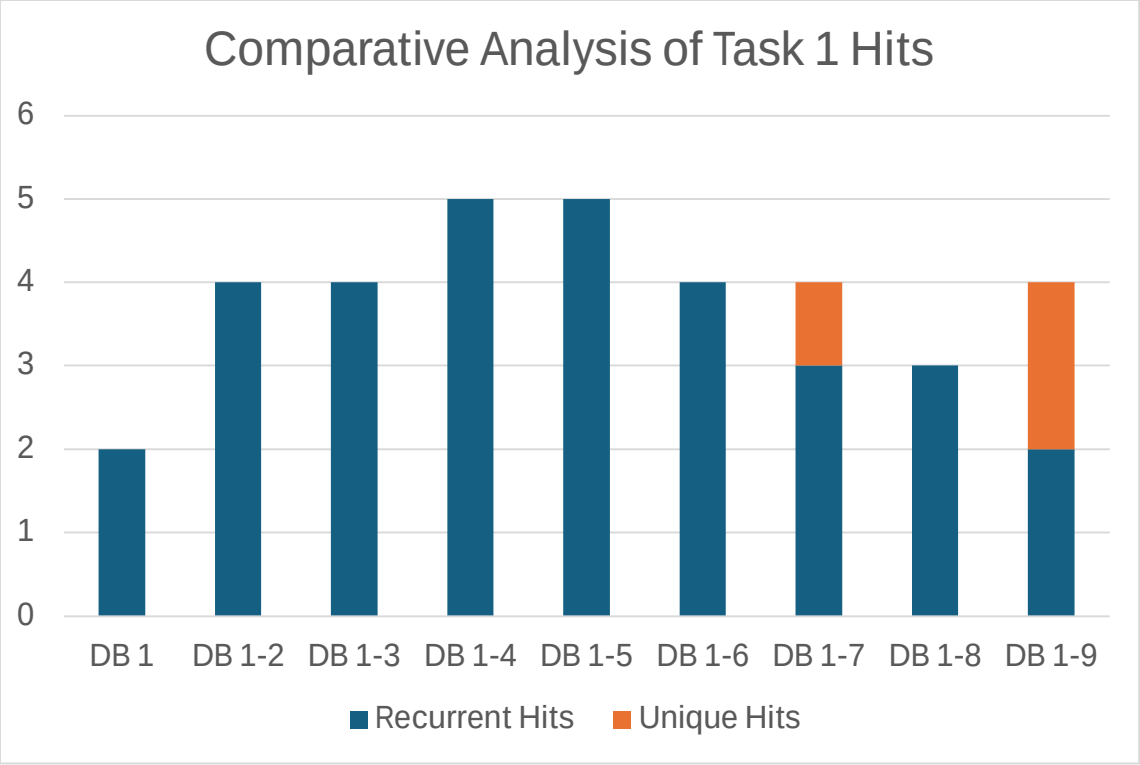


Figure 1 showcases the number of hits found in each database run, as well as how many are unique.

Additionally, using python commands, we calculated the time it took to run batch ten through each database. We can see the time took in the next figure. Each run took about three seconds.

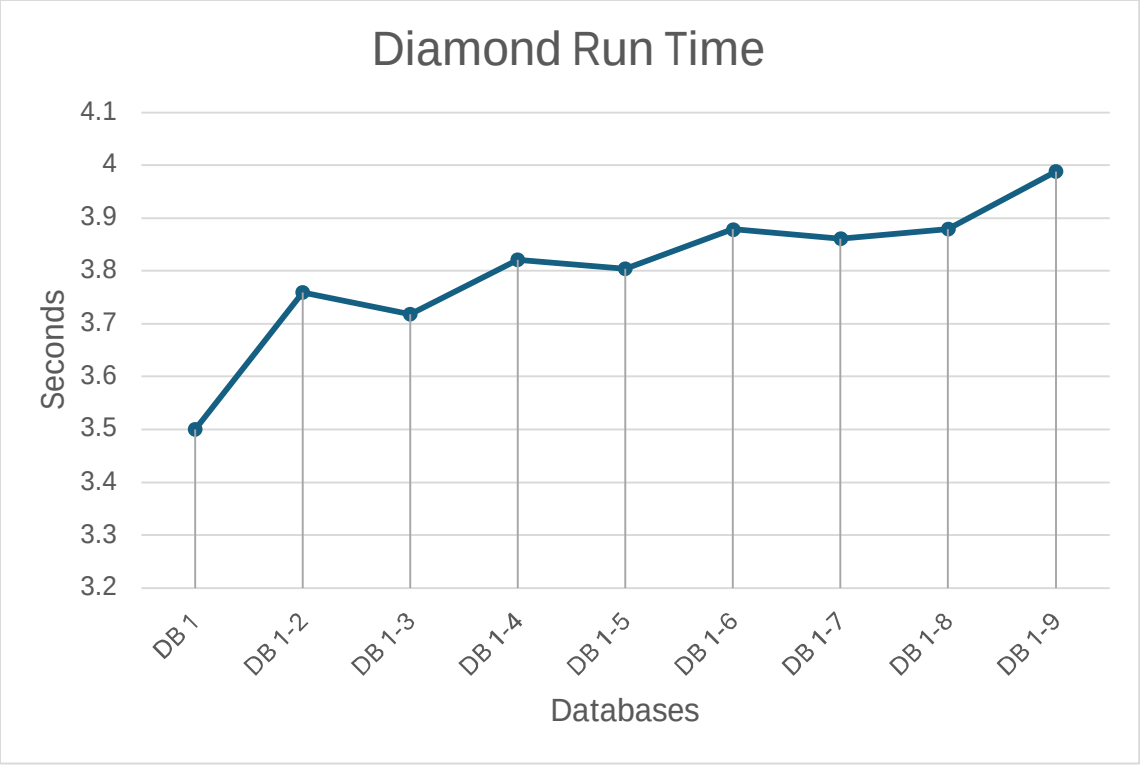


Figure 2 displays run times of batch 10 through each cumulative database.

The following section will provide a brief examination of a particular finding from Task 2, focusing on query 7. The purpose here is not to underscore its critical significance, but to present a detailed snapshot of the information typically contained within a result file, thereby offering insight into the nature of our analytical process. The query was characterized by a sequence from Helicobacter pylori [TaxId: 85962], specifically identified as d4ltta1, which spans amino acids 1-88. This annotation, "d4ltta1 d.298.1.2 (A:1-88)," signifies a distinct protein structure attributed to Helicobacter pylori. This bacterium is recognized for its potential to colonize the human stomach lining, occasionally leading to ulcers or stomach cancer. The descriptive nomenclature encapsulates the database entry, the structural classification of the protein, and the segment of the protein sequence under examination. The hit matched with was delineated as nine-heme cytochrome c from Desulfovibrio desulfuricans ATCC 27774 [TaxId: 876], under the accession a.138.1.1. This denotes a cytochrome c protein variant found in Desulfovibrio desulfuricans, a bacterium implicated in sulfate reduction processes, highlighting the cytochrome c family's significant roles in cellular respiration and energy generation.

The discerned alignment between the Helicobacter pylori and Desulfovibrio desulfuricans sequences unveils potential evolutionary or functional correlations meriting further investigation. The Helicobacter pylori sequence segment analyzed was:

KSFQKDFDKLLLNGFDDSVLNEVILTLRKKEPLDPQFQDH-----
ALKGKWKPFRECHIKPDVLLVYLVKDDEL

This was closely matched by the Desulfovibrio desulfuricans sequence:

EQMQKGINGTLLPGDNEALAAETVLAQKTVEPVSPMLAPYKVVIDALADKYEPSNFTHRRHLTSLME
RIKDDKL

These alignments illuminate conserved and divergent regions, potentially indicative of structural or functional motifs significant to the proteins' roles in bacterial physiology and biochemistry.

It is also pertinent to discuss the overall findings from Task 2. To recap, the approach involved utilizing DIAMOND to process batch ten through two separate databases—firstly, containing batch one, and secondly, batch two. Subsequently, the outcomes were consolidated using iBLAST, followed by a comparison of these merged results against those obtained from executing batch ten through a combined database of batches one and two (obtained from Task 1 results). Interestingly, the alignment hits were consistent across the different result files, with identical iterations yielding hits and the discovered hits themselves being the same in both files. Nonetheless, a slight variation was observed in their e-values. Focusing on query 7, for instance, the e-value for the alignment between Helicobacter pylori and Desulfovibrio desulfuricans, as determined solely by DIAMOND, was 2.54e-04. Conversely, the e-value in the consolidated iBLAST file exhibited a minor reduction, registering at 2.53e-04.

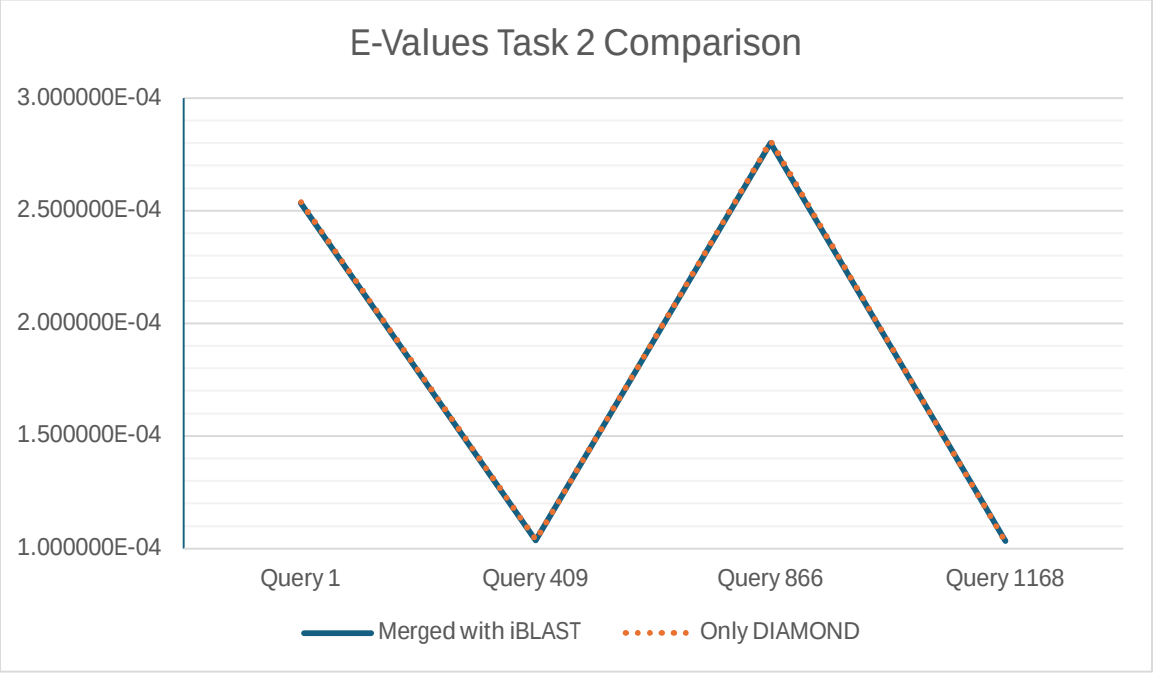
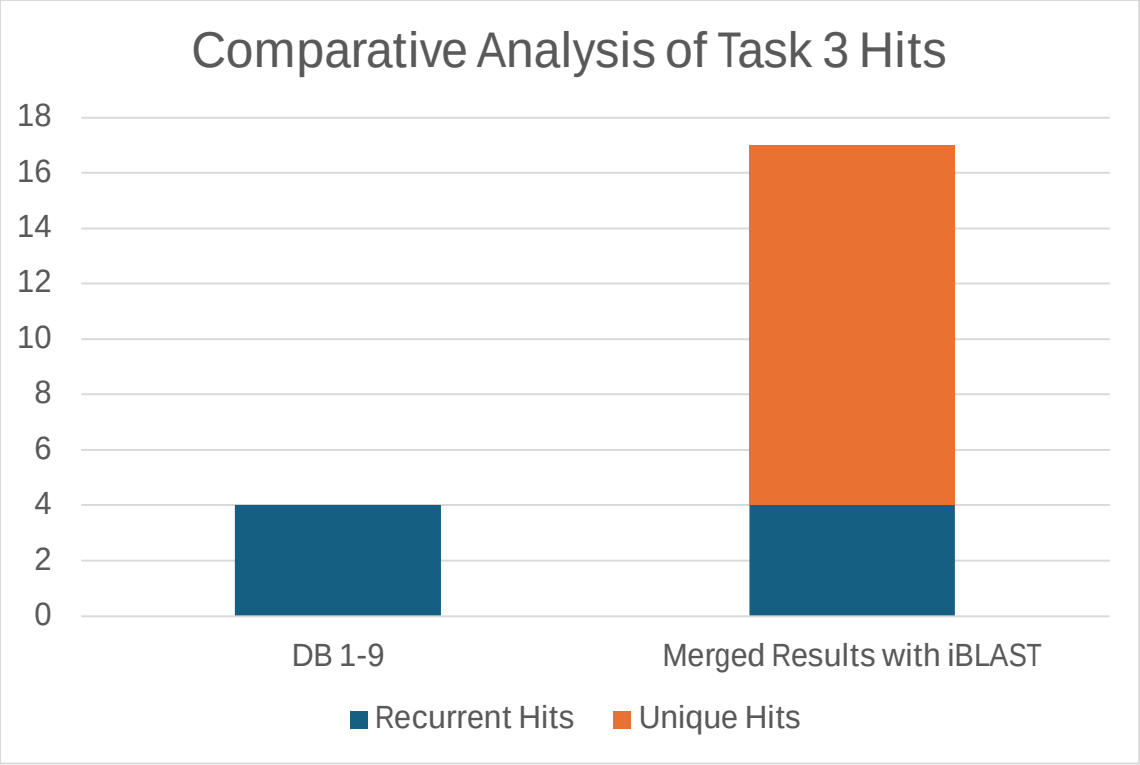


Figure 3 compares the e-values found in the results files of task 2. Though they are quite similar, the values from the results merged with iBLAST are slightly smaller.

Finally, Task 3's results are perhaps the most interesting. While Task 2 yielded the same hits for both result files, that was not the case for Task 3. Using the iBLAST merging method, Task 3 yielded seventeen hits, compared to only four hits from the DIAMOND only method.



This observation adds an intriguing dimension to the analysis of Task 3's results. It suggests that DIAMOND's efficiency in identifying hits may be inversely related to database size, performing optimally when analyzing smaller datasets. Consequently, when results from DIAMOND are merged with those from iBLAST, the combined method capitalizes on DIAMOND's strength in a smaller database context, thereby enhancing the overall hit detection capability. This synergy between DIAMOND's precision in more contained datasets and iBLAST's merging algorithm potentially offers a more comprehensive approach to uncovering relevant alignments. Such a dynamic indicates the importance of strategic tool selection and combination in bioinformatics to maximize data extraction and analysis efficacy, particularly in projects where database size and complexity vary. This insight not only sheds light on the tools' operational characteristics but also informs future research strategies, emphasizing the need for methodological adaptability in response to database scale.

Discussion

IMPLICATIONS

The findings display the importance of optimizing sequence alignment algorithms as genetic databases continue to expand. DIAMOND's efficiency in handling large datasets offers significant advantages over BLAST in situations calling for quick analysis. However, the discrepancy in hit counts between DIAMOND and BLAST warrants further investigation into their sensitivity and specificity, as accuracy is crucial for sequence analysis. The implications of these findings extend beyond the realm of basic sequence alignment. Accurate alignment results serve as the foundation for downstream analyses in fields such as evolutionary biology, functional genomics, and drug discovery. Any compromise in accuracy at the alignment stage could

propagate errors throughout subsequent analyses, potentially leading to incorrect conclusions and interpretations.

FUTURE DIRECTIONS

Current studies are investigating the use of Diamond for querying orthologs, genes diverging at specific evolutionary events and found across multiple species. This approach aims to enhance efficiency compared to traditional methods, which rely solely on nearest sequence matches, by reducing processing time and computational demands. Ortholog identification is crucial for understanding evolutionary relationships and gene functions across species but often faces bottlenecks due to complexity. By harnessing Diamond's efficient sequence alignment algorithms, researchers aim to expedite ortholog queries without sacrificing accuracy. However, ensuring result reliability requires careful validation and parameter optimization. Overall, Diamond holds promise for revolutionizing evolutionary and comparative genomics research, potentially advancing our understanding of gene evolution and function across diverse species.

Other studies like ours are being done on a much more massive scale. For example, a study has been done at the Max Planck Institute for Developmental Biology, Tübingen, Germany that queries 40 million sequences against the 280 million sequences found in NCBI using diamond. Studies such as this show the true magnitude of how accurate DIAMOND is, and if the massive increase in processing speed can outweigh the accuracy lost compared to BLAST.

UNRESOLVED QUESTIONS

Exploring the optimal balance between speed and accuracy in alignment algorithms remains a pertinent question. While DIAMOND offers expedited searches, ensuring that speed enhancements do not compromise the accuracy of results is crucial. Achieving this balance requires a nuanced understanding of the trade-offs involved and may necessitate further refinement of alignment algorithms. Something that may be a future goal but still a question, is how machine learning techniques be integrated with DIAMOND to increase speed and accuracy as well. Using these ML/AI techniques, there is a high possibility that speed, and accuracy can both be maintained by well written models.

WORKS CITED

Buchfink, B., Reuter, K., & Drost, H.-G. (2021). Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nature Methods*, 366–368. doi:10.1038/s41592-021-01101-x

Rahman, S. R., Hines, H. M., & Feng, W.-c. (2021). iBLAST: Incremental BLAST of New Sequences via Automated E-Value Correction. *PLOS ONE*. doi:10.1371/journal.pone.0249410