LITERATURE REVIEW: GPU-Based Accelerations for Protein Sequence Alignment Using BLAST

Eric Arezza
School of Computer Science
Carleton University
Ottawa, Canada K1S 5B6
ericarezza@cmail.carleton.ca

October 22, 2020

1 Introduction

Parallel computing using a Graphics Processing Unit (GPU) is a commonly used hardware acceleration method to reduce the runtime of computations that would otherwise be performed serially. An advantage to only using a GPU instead of a computing cluster is that multiple machines are not required, so no message-passing for remote memory access is needed and the system can work standalone. A GPU is also cheaper, simpler, and provides a more general-purpose programmable option than employing FPGAs for parallel applications. Additionally, GPUs have been gaining more attention and development in recent years making them more ubiquitous, thus providing greater support and functionality. For these reasons, using a GPU to speedup computations renders it a feasible and practical application for parallelization.

A useful application that can benefit from GPUs is computing sequence alignments of biological data such as protein's amino acid sequences. The Basic Local Alignment Search Tool (BLAST) is a standard tool in bioinformatics used by the National Center for Biotechnology Information (NCBI) that performs this function [1, 2]. Specifically, BLASTP is the program for protein sequence alignment and will herein be synonymously referred to as BLAST unless otherwise stated. In general, sequence alignment involves aligning a query sequence to subject sequence and calculating a similarity score between matching amino acids of the sequences. These alignments can help biologists identify functionally similar regions between sequences and understand evolutionary relatedness between organisms or cellular components [3]. A brief description of the BLAST process is explained in the following section to better understand its computational algorithm.

In many cases, multiple proteins may be desired for querying against entire databases of proteins, as done with BLAST, which requires each pairwise alignment to be performed and extends the time to obtain results. More significantly, as next-generation sequencing technologies improved sequencing throughput, protein databases have grown exponentially [4]. Due to this growth, parallel computing will be necessary to minimize the computing times for sequence alignment. Therefore, this paper addresses this topic by independently investigating GPU implementations of BLAST for accelerating protein sequence alignment.

2 Literature Review

2.1 BLAST Algorithm Overview

The BLAST algorithm can be divided into four sections with various opportunities for parallelization: seeding/hit detection, ungapped extension, gapped extension, and traceback.

1. To begin, proteins are parsed into a list of *k*-mer words of their amino acid sequence. For example, using the default *k* value of 3 for the sequence LMDKN would produce a word list of LMD, MDK, and DKN. The subject sequence from a protein database is then searched for matches to each word from the query, resulting in "hits". Clearly, this step of hit detection can benefit from parallelization of all *k*-mer words and extended to the next steps. Every hit becomes a seed for extending amino acid characters on each side of the word for both sequences.
2. Ungapped extension is performed whereby the seed is iteratively extended on each side and the resulting alignment between query and subject hit is scored based on a fixed scoring matrix and a given threshold. As an aside, in PSI-BLAST [2], this scoring matrix is re-calculated after additional iterations of the BLAST alignment providing a Position-Specific Scoring Matrix (PSSM). If the seed extension increases the score then the seed continues extending, but if the score begins to drop off then the seed is no longer extended.
3. Gapped extension is then performed in a similar way allowing gaps in the seed extension and the resulting alignment (ungapped and/or gapped) becomes a local alignment hit between the query and subject sequences with an associated similarity score. If the score is above a given threshold, the alignment is saved as a High Scoring Pair (HSP). Figure 1 below illustrates an example of seed extension using the word IYP from a query against a subject sequence.
4. Finally, these HSPs are traced back for sequence information, saved, and formatted for user output.

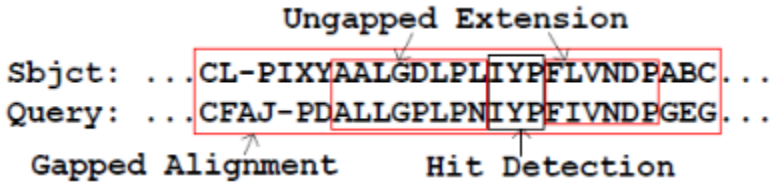

Figure 1. Illustration of BLAST seeding and extension for protein alignment [5].

2.2 GPU-Accelerated Approaches

GPU-based BLAST algorithms have been published as early as 2010 [6] and have continuously been investigated. A summary of GPU strategies for accelerating certain stages of the BLASTP algorithm was presented by Said et al. [7] for CUDA-BLASTP [8], NCBI-GPU-BLASTP [9], GPU-BLASTP [5], and CuBLASTP [10] methods while proposing a strategy for parallelizing PSI-BLAST. Another earlier analysis of GPU accelerated methods for BLAST was conducted by Glasco [11]. It is apparent that this topic is recurring as new approaches are presented and the need for efficient BLAST acceleration grows.

Although a GPU-enabled cluster can certainly improve speedup as proposed in CLUS_GPU-BLASTP [12], this solution will typically offer a roughly linear speedup with cluster size relative to the GPU implementation used. Consequently, this paper will only investigate the uses of a GPU on a single machine for BLAST. The most recent GPU implementation for BLAST protein alignment was developed by Ye et al. named H-BLAST [13] utilizing NCBI-BLAST base code and functions from GPU-BLAST and will be compared with previously mentioned methods.

Table 1. Summary of GPU Methods for BLASTP.

	NCBI-GPU-BLASTP	**CUDA-BLASTP**	**GPU-BLASTP**	**CuBLASTP**	**H-BLAST**
Database (proteins size)	env_nr (6,031,291)	GenBank nr (9,230,955)	Ncbi nr (9,874,397)	env_nr (6,000,000), swissprot (300,000)	Ncbi nr (14,324,397)
Query Proteins	51 mouse (UniProt)	P14144, P42018, Q52TG9, Q52KR2, P08678	4 proteins	3 proteins	250 (swissprot), 6 groups (100 to 600)
Query Lengths/Amino Acids	Up to 4498	127, 254, 517, 1054, 2026	1000 to 4000	127, 517, 1054	100 to 5000, maximum 9000
FSA or NCBI Base Code	NCBI	NCBI	FSA	FSA	NCBI
Reported Comparisons	BLASTP	BLASTP	BLASTP, CUDA-BLASTP	BLASTP, FSA-BLAST, CUDA-BLASTP, GPU-BLASTP	BLASTP, GPU-BLASTP
GPU Used	Fermi C2050	GeForce GTX 280, GeForce GTX 295	Tesla, C1060, Fermi C2050	Kepler K20c	K20x, K40m
Available	http://archimedes.cheme.cmu.edu/?q=gpublast	https://sites.google.com/site/liuweiguohome/software	N/A	https://github.com/vtsynergy/cuBLASTP	https://github.com/Yeyke/H-BLAST

BLAST's heuristic approach for sequence alignment was an initial advantage to faster processing times than the more accurate Smith-Waterman algorithm and led to its adoption as a standard tool. An enhancement of BLAST was further developed, termed Faster Search Algorithm (FSA-BLAST) to reduce computational complexity and time by about half per query while maintaining nearly identical functionality and output as BLAST [14]. Some GPU strategies implement the FSA-BLAST base code along with their parallel approach. For example, GPU-BLAST is built over FSA-BLAST while NCBI-GPU-BLAST is built over the original BLAST code. Thus, these methods should be compared appropriately.

Often, results are reported relative to each other which may use different NCBI-BLAST or FSA-BLAST base code, compared under different database and query searches, and of course using different hardware as seen in Table 1 above. This investigation will re-evaluate previous results of such approaches providing an independent comparison of their performance. Further analysis of the results will be discussed in the following sections regarding BLAST algorithm speedup.

References

[1] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990 Oct 5;215(3):403-10. doi: 10.1016/S0022-2836(05)80360-2. PMID: 2231712.

[2] Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997 Sep 1;25(17):3389-402. doi: 10.1093/nar/25.17.3389. PMID: 9254694; PMCID: PMC146917.

[3] Xiong, J. (2006). SEQUENCE ALIGNMENT. In Essential Bioinformatics (pp. 29-94). Cambridge: Cambridge University Press.

[4] The UniProt Consortium, UniProt: a worldwide hub of protein knowledge, Nucleic Acids Research, Volume 47, Issue D1, 08 January 2019, Pages D506–D515, https://doi.org/10.1093/nar/gky1049

[5] S. Xiao, H. Lin and W. Feng, "Accelerating Protein Sequence Search in a Heterogeneous Computing System," 2011 IEEE International Parallel & Distributed Processing Symposium, Anchorage, AK, 2011, pp. 1212-1222, doi: 10.1109/IPDPS.2011.115.

[6] Ling, Cheng, and Khaled Benkrid. “Design and Implementation of a CUDA-Compatible GPU-Based Core for Gapped BLAST Algorithm.” Procedia computer science 1.1 (2010): 495–504. https://doi.org/10.1016/j.procs.2010.04.053

[7] M. Said, M. Safar, M. Taher and A. Wahba, "Accelerating iterative protein sequence alignment on a heterogeneous GPU-CPU platform," 2016 International Conference on High Performance Computing & Simulation (HPCS), Innsbruck, 2016, pp. 403-410, doi: 10.1109/HPCSim.2016.7568363.

[8] Liu W, Schmidt B, Müller-Wittig W. CUDA-BLASTP: accelerating BLASTP on CUDA-enabled graphics hardware. IEEE/ACM Trans Comput Biol Bioinform. 2011 Nov-Dec;8(6):1678-84. doi: 10.1109/TCBB.2011.33. PMID: 21339531.

[9] Vouzis, P. D., & Sahinidis, N. V. (2011). GPU-BLAST: using graphics processors to accelerate protein sequence alignment. Bioinformatics (Oxford, England), 27(2), 182–188. https://doi.org/10.1093/bioinformatics/btq644

[10] Zhang J, Wang H, Feng WC. cuBLASTP: Fine-Grained Parallelization of Protein Sequence Search on CPU+GPU. IEEE/ACM Trans Comput Biol Bioinform. 2017 Jul-Aug;14(4):830-843. doi: 10.1109/TCBB.2015.2489662. Epub 2015 Oct 12. PMID: 26469393.

[11] Glasco, D. (2012). An Analysis of BLASTP Implementation on NVIDIA GPUs. [Online] Available: https://www.semanticscholar.org/paper/An-Analysis-of-BLASTP-Implementation-on-NVIDIA-GPUs-Glasco/42bb3ca76542c08566547de2d828b2e3e61af4f3

[12] Rani, S., Gupta, O.P. CLUS_GPU-BLASTP: accelerated protein sequence alignment using GPU-enabled cluster. J Supercomput 73, 4580–4595 (2017). https://doi-org.proxy.library.carleton.ca/10.1007/s11227-017-2036-4

[13] Weicai Ye, Ying Chen, Yongdong Zhang, Yuesheng Xu, H-BLAST: a fast protein sequence alignment toolkit on heterogeneous computers with GPUs, Bioinformatics, Volume 33, Issue 8, 15 April 2017, Pages 1130–1138, https://doi.org/10.1093/bioinformatics/btw769

[14] M. Cameron, H.E. Williams, and A. Cannane, ``Improved Gapped Alignment in BLAST'', IEEE/ACM Transactions on Computational Biology and Bioinformatics, 1(3), 116-129, 2004.