CATEGORY: **LIFE & MATERIAL SCIENCE - LS08**

POSTER **P5273** | CONTACT NAME **Shuji Suzuki: suzuki@bi.cs.titech.ac.jp**

GPU TECHNOLOGY CONFERENCE

GHOSTZ-GPU: Fast Protein Sequence Homology Search on GPUs

Shuji Suzuki[1], Masanori Kakuta[1], Takashi Ishida[1] and Yutaka Akiyama[1]

[1] Graduate School of Information Science and Engineering, Tokyo Institute of Technology, Japan

TOKYO TECH Pursuing Excellence

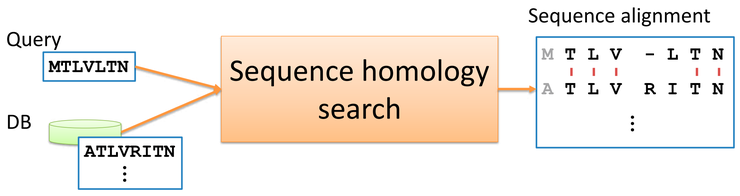
Tokyo Institute of Technology—TOKYO TECH—develops distinctive students with outstanding qualities of creativity and leadership. TOKYO TECH is making significant contributions to science and technology in many fields of expertise, creating new and powerful synergies. TOKYO TECH, being a research-based university, is dedicated to education and research, and to exploring knowledge in science and technology. Pursuing excellence, TOKYO TECH serves society and the world.

Abstract

Protein sequence homology searches are often used in various biological fields. Currently, protein sequence homology searches require large computational time. To accelerate protein sequence homology search, we mapped our original GHOSTZ algorithm, which is one of the fastest protein sequence homology search algorithms for metagenomic analysis, onto GPUs and implemented it as GHOSTZ-GPU. As results, GHOSTZ-GPU with 12 CPU threads and 3 GPUs was an approximately 7.1-fold faster than GHOSTZ with 12 CPU threads.

Sequence homology search

Sequence homology search is a method to find biologically similar sequences to a query from a sequence database.
This method is essential for identifying evolutionary relationships among species and can be also used for estimating potential functions and structures of biomolecules. To find similar sequences, the measurement of sequence similarity is required. In biological analyses, the highest score of the sequence alignment for query and database sequences used in general.



The computation time

O(|Q||D|)
|Q|:query length
|D|: the total length of sequences in a database

Current major sequence homology search

BLAST, BLAT, RAPSearch, GHOSTX, GHOSTZ [1]

Research Aim:

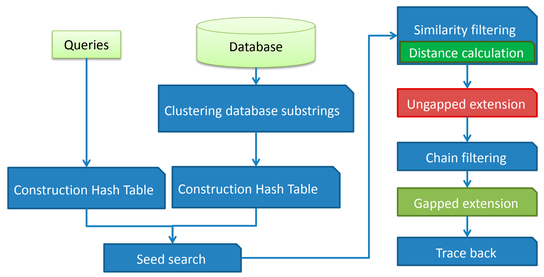
Huge number of protein sequence homology searches become a bottleneck in various biological fields, including metagenomic analysis

1. Large computation of sensitive homology searches
2. A huge amount of genomic data
   The size of DNA sequencer (Hiseq2500) result is about 1TB / run

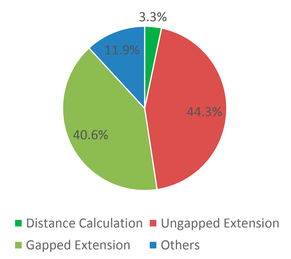
To accelerate protein sequence homology search, we developed a protein sequence homology search algorithm with GPUs based on GHOSTZ algorithm and we implemented it as **GHOSTZ-GPU**.

GHOSTZ: Faster sequence homology searches by clustering subsequences

GHOSTZ [1] is one of the state of the art protein sequence homology search tools for metagenomic analysis. GHOSTZ clusters similar subsequences from a database to perform an efficient seed search and ungapped extension by reducing alignment candidates based on triangle inequality. GHOSTZ with 1 CPU thread is an approximately 261-fold faster than BLAST with 1 CPU thread [1].



The flow of GHOSTZ



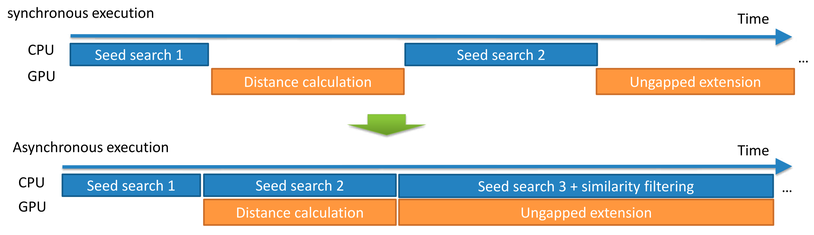
Execution profile of GHOSTZ

GPU Implementation

To accelerate protein sequence homology search, we developed a protein sequence homology search algorithm with GPUs based on GHOSTZ algorithm and implemented it as GHOSTZ-GPU. We have mapped the following steps of GHOSTZ algorithm to GPUs: **distance calculation**, **ungapped extension** and **gapped extension**. For the GPU implementation, we used NVIDIA's CUDA 6.0. To increase utilization efficiency of GPUs, we used following techniques.

(1) Asynchronous execution CPU and GPU

GHOSTZ-GPU reduces inactive threads in gapped extension and uses asynchronous executions on CPU and GPU.

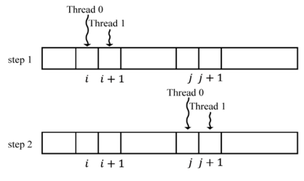


(2) Optimization of reading database

While the other threads perform protein sequence homology search against a database chunk, the special thread reads the next database chunk.

(3) Optimizing sequence memory access

Distance calculation, ungapped extension and gapped extension in GHOSTZ require many accesses for sequence data in global memory. Therefore, GHOSTZ-GPU is optimized the data accesses in these memories on GPU.



(4) Load balancing of gapped extension

The query length affects the computing time of gapped extension. For better load balancing, GHOSTZ-GPU sorts seeds by the query length and then assigns a seed to a thread on GPU in order.

Results

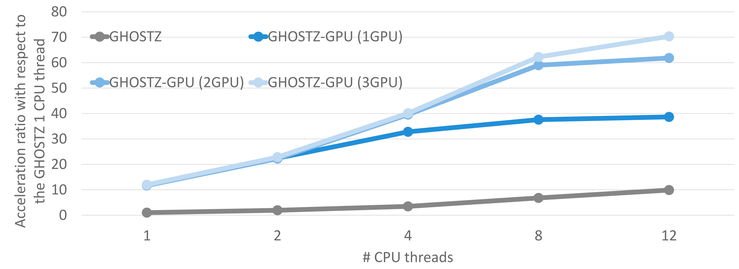
We performed GHOSTZ-GPU to evaluate the performance of our system. We used metagenomic data sampled from soil microbiome metagenomic sequences and obtained by using a next-generation sequencer.

Computational environment:
TSUBAME2.5
(1 Thin node)

CPU: Intel Xeon 2.93 GHz (6 cores) x 2
GPU: NVIDIA Tesla K20X x 3
Memory: 54GB

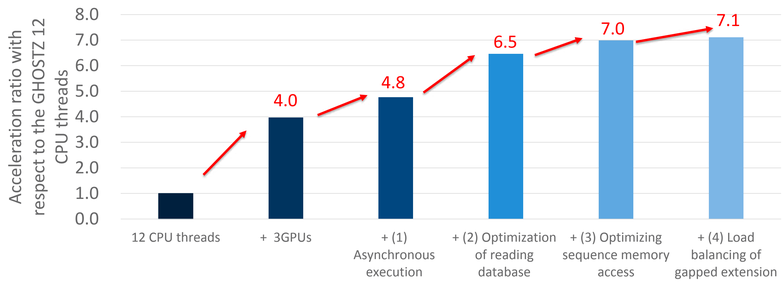
Data set:
Queries:
soil microbiome metagenomic sequences (SRR407548, 150 bp, 1,000,000 reads)

Database:
Kyoto Encyclopedia of Genes and Genomes (KEGG) GENES database (May 2013, 10 million sequences, 3.9GB)



- GHOSTZ-GPU (12 CPU threads + 2 GPUs) was an approximately **6.2-fold faster** than GHOSTZ (12 CPU threads).
- GHOSTZ-GPU (12 CPU threads + 3 GPUs) was an approximately **7.1-fold faster** than GHOSTZ (12 CPU threads).

GHOSTZ-GPU (12 CPU threads + 3 GPUs) is estimated to be an approximately **1853-fold (= 261 × 7.1) faster** than BLAST (12 CPU threads).



References

[1] Shuji Suzuki, Masanori Kakuta, Takashi Ishida, and Yutaka Akiyama, Faster sequence homology searches by clustering subsequences, *Bioinformatics* (in press) doi:10.1093/bioinformatics/btu780

Acknowledgments

This research was supported by a Grant-in-Aid for JSPS Fellows (248766) and HPCI Strategic Program Computational Life Science, Application in Drug Discovery and Medical Development by MEXT of Japan and the CUDA COE Program by NVIDIA.