GTC Workshop Japan 2011 (2011 年 7 月 22 日)

マルチ CPU コアと GPU を用いた高速な配列相同性検索

A fast homology search tool with multi-CPU cores and GPUs

鈴木　脩司 [1)], 石田　貴士 [1)], 秋山　泰 [1)]
Shuji SUZUKI, Takashi ISHIDA and Yutaka AKIYAMA

1) 東京工業大学　大学院情報理工学研究科 （〒 152-8552 東京都目黒区大岡山 2-12-1 W8-76）

Key Words: *CUDA, sequence homology search, next-generation sequencer*

1　はじめに

クエリ配列に類似する配列をデータベースから探索する相同性検索は，生物学における重要な解析手法である．しかし，近年次世代シーケンサーの登場により DNA 配列の読み取り速度が劇的に向上したことで，従来よく用いられてきた相同性検索ツールの BLAST ではこの解析を行うことが困難になってきた．例えば，最新のシーケンサーでは 1 度の実行で約 600G 塩基分の情報を得ることができるため，BLAST ではメタゲノム配列の相同性検索に約 25,000CPU 日が必要になる．我々はこの問題を解決するため，以前に GPU を用いた相同性検索ツールである GHOSTM[1] の開発を行った．GHOSTM の処理は seed 探索と伸長という 2 段階に大きく分かれており，GHOSTM ではこの 2 つの処理を共に GPU によって処理していた．しかし，seed 探索は GPU よる高速化が難しい処理であり，そこがボトルネックとなっていた．一方で，GHOSTM では 1 CPU コア対 1GPU という割合で計算資源を使用する設計であったため，マルチコア CPU の場合，活用できていない CPU コアが多く残されていた．このため，本研究では GPU による高速化の寄与が大きい伸長の処理は GPU 上で高速に処理する一方，seed 探索は CPU 上で実行することで GPU では処理が難しいが高速なアルゴリズム [2] を採用し，さらにその処理を並列化して複数の CPU コアを使って高速化した．こうすることで，1 ノードの全ての計算資源を有効に活用する事が可能となり，GPU のみを用いた場合に比べ更なる高速化を達成した．

2　手法

まずデータベースとクエリ，それぞれについて効率的に文字列検索を行うデータ構造である suffix array を構築し，局所的に類似度の高い位置 (seed) を探索する．そして，探索して見つかった seed を中心に伸長し，類似度が閾値以上の配列をデータベースの中から探し出すことで相同性検索を行う．

ここで，seed 探索は多くのメモリを使用するため，GPU 上での処理は困難である．そのため seed 探索は CPU で行い，更にマルチスレッドによる並列処理によって高速化した．また，配列の長さが長くなるにつれて伸長の計算に多くの時間が掛かるが，GPU によって 1 CPU の 10 倍の高速化が可能である．このため，伸長の計算に GPU を使用している．

また，CPU と GPU は非同期で通信，計算が可能であり，seed 探索をした後，GPU で伸長を計算させている最中に CPU が GPU の計算を終了を待つ必要はない．このため，図 1 のように seed 探索と伸長をオーバーラップした．

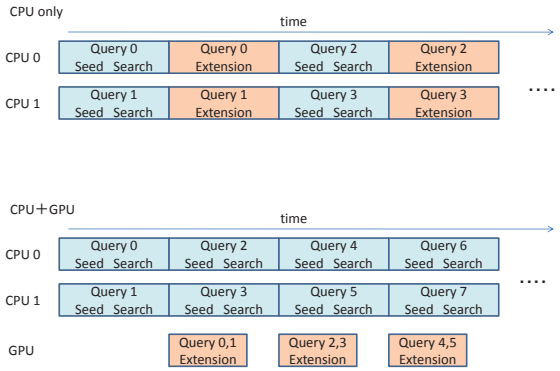

図 1: CPU の場合 (上) と CPU+GPU の場合 (下) の処理の流れ

3　結果

提案手法の計算速度を TSUBAME2 の Thin ノード上で NCBI BLAST(version 2.2.24) の実行速度と比較することで評価を行った．このノードには，12CPU コアと Tesla M2050 が 3 つ搭載されており，1GPU あたり 4 つの CPU コアを持っている．

提案手法は BLAST と比べると 1CPU コア + 1GPU を利用した時，最大で約 260 倍の高速化を達成した．また，複数の CPU コアを利用して 4CPU コア + 1GPU で計算を行ったところ，1CPU コア + 1GPU で計算した場合に比べ，約 1.73 倍の速度向上が得られた．そのため，1 ノード全ての計算機資源 12CPU コア + 3GPU を利用すると，その計算速度は BLAST に比べ約 1350 倍となっており，大幅な高速化を達成することができた．GPU は高速に計算ができるが，GPU だけでなく CPU のコアをすべて活用することで，更なる高速化が達成できる．

参考文献

[1] 鈴木 脩司 , 石田 貴士, 秋山 泰: GPU による DNA 断片配列の高速マッピング, 情報処理学会研究報告バイオ情報学 (BIO), 21(30), pp.1-6, 2010.

[2] 鈴木　脩司 , 石田 貴士 , 秋山 泰: FM-index を用いた高速な配列相同性検索ツールの開発, 情報処理学会研究報告バイオ情報学 (BIO), 23(21), pp.1-6, 2010.