利用課題名　次世代シーケンサーを用いたメタゲノム解析向けの超高速パイプラインの構築

[1]秋山泰、[2]黒川顕、[3]小西史一、[1]石田貴士、[1]鈴木脩司

1 東京工業大学 大学院情報理工学研究科 計算工学専攻

2 東京工業大学 大学院生命理工学研究科 生命情報専攻

3 東京工業大学 情報生命博士教育院

邦文抄録（500 字程度）

近年注目を集めているメタゲノム解析は未知の微生物に関するゲノム情報が得られるだけでなく、その環境中の共生系の理解や環境汚染の監視等に有用であるが、解析を進める上で計算量の大きな配列相同性検索処理がボトルネックの一つとなっている。この問題に対処するため、我々はGPUを用いた高速な配列相同性検索プログラムを開発し、TSUBAME2 上に大規模なメタゲノム解析向けの超高速パイプラインの構築をおこなった。

Keywords: メタゲノム解析、次世代シークエンサー、GPU 計算、配列相同性検索

1. 研究の背景

従来のゲノム解析は培養された単一の種のゲノム情報を明らかにするものであったが、近年シークエンサーの性能向上に伴い、土壌、海洋、ヒト体内等の環境中に生息する微生物のゲノムを分離培養せずにそのまま丸ごとシークエンスして解析することが可能となってきた。この解析手法はメタゲノム解析と呼ばれ、未知の微生物に関してのゲノム情報が得られるだけでなく、その環境中の共生系の理解や環境汚染の監視等に有用であり、大きな注目を集めている[1]。さらに近年では次世代シークエンサーと呼ばれる新型のシークエンサーの登場により、膨大な量のゲノム情報が短時間のうちに入手可能となっており、その大規模な情報を用いることでメタゲノム研究が更に進展することが期待されている。

しかし、このメタゲノムの解析では、サンプルに含まれる多くの種のゲノム情報がデータベースに登録されていないため、遠縁の種の配列データとの間で比較が可能となる高感度な検索手法が必要となる。配列相同性検索と呼ばれるこの検索処理は多くの計算を必要とする処理であり、その結果、メタゲノム解析を進める上でのボトルネックの一つとなってしまっている[2]。

そこで、本研究では次世代シークエンサーによる大量のメタゲノム情報を現実的な時間内に解析することを目的とし、TSUBAME2 の膨大な計算能力を利用可能とする大規模な全自動解析パイプラインを構築した。また、そのパイプライン中で行われる配列相同性検索そのものを高速化するため、従来のメタゲノム研究で標準的に利用されてきた配列相同性検索プログラム BLASTX[3]に加え、BLASTXと同等の検索感度を持つオリジナルのGPUプログラムである GHOSTM[4]を開発し、利用可能とした。これによって次世代シークエンサーの一度の読み取り実行によって得られるゲノム情報を数時間の内に処理することが可能となり、今後このパイプラインによって次世代シークエンサーによるメタゲノム解析が促進されることが期待される。

2. メタゲノムマッピング

次世代シークエンサーと呼ばれる現在のシークエンサーは非常に高いスループットを持ち、1度の読み取り実行で数千億塩基以上が解読可能となっている。しかし、その出力は 100 塩基（bp）程度の短い断片配列であるため、そこから意味のある情報を得るにはその出力に対してアセンブリやマッピングといった処理を計算機上で行う必要がある。

マッピング処理は既知のゲノム配列に対して断片配列を貼り付け、その一致する位置を同定する処理である。従来行われてきたような単一の生物種に対する解析ではリファレンスゲノムが存在するため、多くの不一致やギャップを

許容する必要がなく、通常の文字列検索に近い処理によって解析が可能であった。現在では BWA[5]や Bowtie[6]といった高速なプログラムが開発されており、数台のワークステーションがあれば次世代シークエンサーの出力に対しても十分に対処が可能となっている。
その一方、メタゲノム解析では環境中に含まれる微生物のすべてのゲノム配列が既知であることは稀であるため、マッピングにおいては近縁の種のゲノム情報を参照する必要があり、曖昧な一致まで検知可能となるような高い感度の検索が必要となる。そのような多くの不一致やギャップを許容する検索は一般に配列相同性検索と呼ばれ、従来のゲノムマッピングに比べて非常に多くの計算を必要とする。さらにメタゲノム解析ではより高い検索感度を得るため、A, T, G, C の 4 文字で表現される DNA 配列のままで検索を行なわず、それらをコドン表に従い 20 種類のアミノ酸からなるタンパク質配列に翻訳してから解析を行う。DNA 配列の比較では塩基の差は一致か不一致の 2 状態でしか区別しないことが多いのに対し、タンパク質配列ではアミノ酸間で性質の類似度の違いから、アミノ酸置換毎に異なるスコアを用いるため、計算量は更に大きなものとなる。
このようなンパク質配列間での曖昧な一致も含めた検索を行うため、現在では比較的高速かつ高感度で配列比較が可能な近似的手法 BLAST[3]が利用されてきた[7]。しかし、次世代シークエンサーは膨大な量の DNA 断片配列を出力するため，すべての DNA 断片配列をアミノ酸配列に翻訳してからマッピングを行うのは BLASTX プログラムを用いても長い計算時間が必要となる。現在、最新の Illumina 社の Hiseq 2000 DNA シークエンサーの出力は 1 度の読み取りで合計 600G 塩基にも達し、そのデータを解析するには約 25,000CPU 日が必要となってしまう。そのため、現在ではメタゲノムの解析においては、このマッピング処理が大きなボトルネックとなっており、その高速化が必要とされている。

3. GPU 配列相同性検索プログラム GHOSTM

我々はメタゲノム解析において処理のボトルネックとなっている配列相同性検索を高速化するため、GPU を利用した高速な配列相同性解析プログラムである GHOSTM[4]を開発した。GHOSTM のアルゴリズムは BLAST と類似したものであるが、GPU 上での実行により適したものとなっており、メタゲノムの解析に必要十分な検索感度を持つように設計されている。GHOSTM は NVIDIA 社の CUDA を用いて実装されており、CUDA バージョン 2.2 以上が利用可能な NVIDIA 社製の GPU が搭載された計算機で利用可能である。一部の処理を簡略化し、処理の大部分を GPU 化することで大幅な高速化を達成しており、詳細は後述するが 1GPU を用いた場合、1CPU コア上で実行された BLASTX に対して約 130 倍の高速化を達成している。

3. 1 アルゴリズム

図1に示すように、GHOSTM では BLAST と同様にまず K 文字の部分文字列の一致を探索する事でアラインメントの候補となる部位を同定し、その後に各アライメント候補について Smith-Waterman アルゴリズム[8]により局所的アラインメントを行う事で詳細なアラインメントとそのスコアを計算している。そして、最終的に全てのアラインメントのスコアがソートされ、そのスコア上位のヒットが検索結果として出力される。
GHOSTM ではこれらの処理のうち、計算の大部分を占めるアラインメント候補探索と局所的アラインメントの双方を GPU 上で処理している。GPU は多くのコアを搭載しており、処理を適切に並列化することで大きな高速化が可能であるが、GHOSTM では大量の断片配列を処理する必要があるため、まず、アラインメント候補の探索の際には各断片配列の処理を各 GPU スレッドで行い、局所的アラインメントの際には各アラインメント候補に対する処理を GPU スレッドに割り当てて処理することで並列化を行っている。

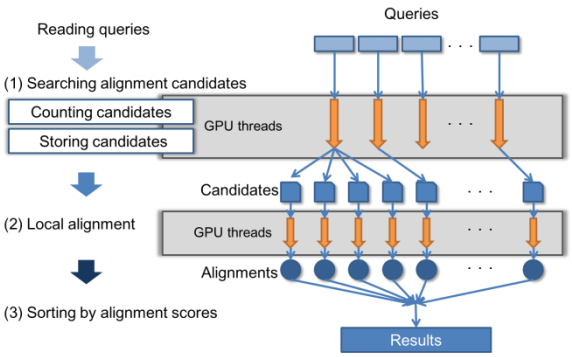

図 1　GHOSTM の処理の流れ

3. 2 検索速度と検索感度

メタゲノム解析では高い検索感度が必要となるため、BLAT[9]のような BLAST より高速であるがに検索感度の低いアルゴリズムを利用する事は困難であった。一方

GHOSTM はメタゲノム解析に十分な検索感度を有しており、図2で示すように動的計画法によりアライメントを行うSSEARCH[10]の結果を正解としたテストでは BLAST には劣るものの、BLAT よりも高い検索感度を示し、特に実際に利用される Bit スコア 50 以上の領域では BLAST とほぼ同程度の検索感度を示しており、BLAST の代替として利用可能な事が示されている。

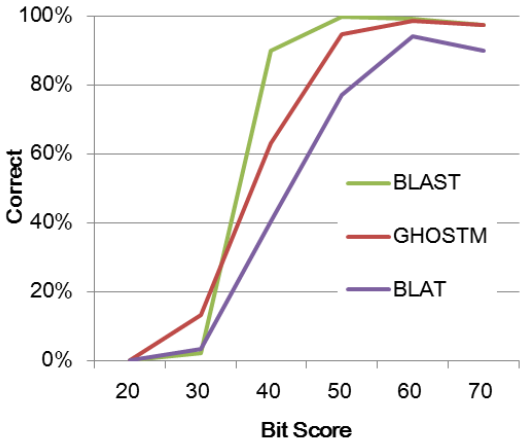

図2　GHOSTM の検索感度

また、表1は約 75bp のリード 100,000 本をクエリとして、KEGG の genes.pep タンパク質配列データベース（約 2.5GB）に対して相同性検索を実行した際の実行時間であるが、GHOSTM の検索速度は1GPU を用いた場合、1CPUコア上で実行されたBLASTXに対して約130倍、4GPUを用いた際には約400倍の高速化を達成している。これは検索感度落とすことで高速化を実現しているBLAT の高速化率約 40 倍に比べてもより高速であり、GHOSTM は高感度と高速化を同時に実現している。

表 1. 実行速度の比較

Program	#GPUs	Time (sec.)	Acceleration ratio
GHOSTM	1	2855	129.5
GHOSTM	4	909	406.7
BLAT		9898	37.3
BLASTX (1 thread)		369678	1
BLASTX (4 threads)		102255	3.6

4.　大規模全自動解析パイプライン

我々はメタゲノム解析におけるボトルネックである配列相同性解析を高速化するため、GPU を利用した高速なプログラムである GHOSTM を作成した。しかし、一台の計算機では、この GHOSTM を用いても次世代シークエンサーの出力を解析するには不十分である。そのため、我々は TSUBAME2 の大量の計算ノードを解析に利用することで、次世代シークエンサーによるメタゲノム解析を現実的な時間で行うことを目指し、TSUBAME2 上に全自動のメタゲノム解析パイプラインを構築した。多くの計算ノードを利用する大規模な計算ではデータベースのコピー、計算結果の書き込みといったファイルの入出力に関する処理が問題となるが、このパイプラインではノード間でコピーを2分木状に行う工夫を行う事で入出力に依存するボトルネックを解消している。

5.　グランドチャレンジ制度で得られた結果

我々は本パイプラインについてTSUBAMEグランドチャレンジ（超大規模アプリケーション）制度を利用し、次世代シークエンサーによって得られた大規模なデータに対する解析速度の検証を行った。解析に用いたデータは汚染土壌に関するメタゲノムデータである。メタゲノム原データは各 75bp の DNA リード 224 million 本であるが、実際に配列相同性検索が行われたのは、ここから低品質なデータを除去後した 71million 本の DNA リードであり、これらがクエリとして 4.2GB のタンパク質配列情報が含まれた NCBI nr データベースに対する検索が行われた。パイプラインの相同性検索エンジンとしては CPU 上で動作する BLASTX と我々の開発した GPU 上で動作する GHOSTM のそれぞれについて実効性能の比較を行った。

パイプラインは計算コア数に対してほぼ線形の速度向上を示し、図3のように BLASTX を相同性検索に用いた場合、TSUBAME2 の CPU 16,008 コア（1,334 ノード）用いた際に 1 時間あたり約 24 million の DNA リードの処理を実現した。また、GHOSTM を用いた場合では図4のように、2,520 GPU（840 ノード）を用いることで 1 時間あたり約 60 million の DNA リードの処理を実現した。GHOSTM と 2,520 GPU を用いた場合では 1,260GPU を用いた場合に対して速度の向上が 2 割程度に留まっているが、これは計算資源に対して今回用いたデータのサ

イズが小さくなりすぎてしまったため、負荷分散等に失敗したことが原因であり、最新のシークエンサーの出力や、複数回のシークエンシングの結果を同時に処理すればGPU 数に対する速度向上はほぼ線形となると考えられる。

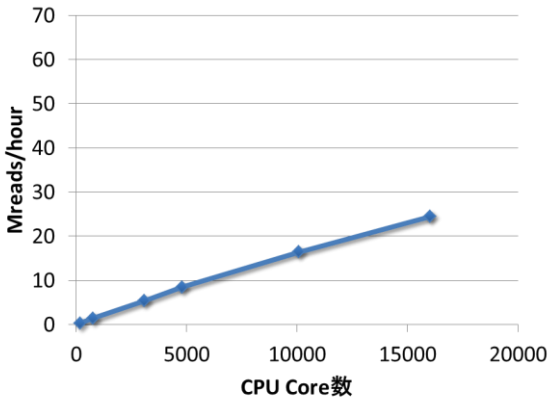

図3. BLASTX を用いた際の CPU コア数に対する処理速度の向上

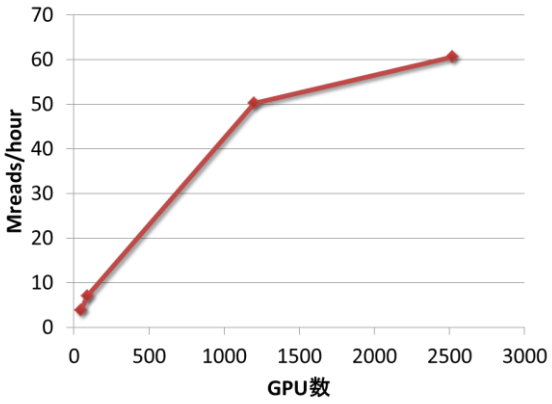

図4. GHOSTM を用いた際の GPU 枚数に対する処理速度の向上

6. まとめ

我々は次世代シークエンサーから得られるメタゲノムの解析において現在ボトルネックとなっている相同性検索を高速化し、現実的な時間で解析を行うことを目的とし、GPU による効率的な相同性検索プログラム GHOSTM を開発し、また TSUBAME2 の大量の計算機資源を利用可能とする大規模な解析パイプラインの構築を行った。解析パイプラインは使用した CPU コア数、GPU 数に対してほぼ線形の速度向上を示し、TSUBAME2 のほぼ全体を利用した場合、1 時間あたり約 60 million の DNA リードの処理が可能であった。これは次世代シークエンサーの 1 度の読み取り実行から得られる出力に対する解析が数時間以内に可能となる事を示しており、今後我々のパイプラインによって次世代シークエンサーによるメタゲノム解析が促進されると考えている。

参考文献

[1] M. Arumugam et al., “Enterotypes of the human gut microbiome.,” *Nature*, vol. 473, no. 7346, pp. 174-80, May 2011.

[2] J. C. Wooley, A. Godzik, and I. Friedberg, “A primer on metagenomics.,” *PLoS computational biology*, vol. 6, no. 2, p. e1000667, Jan. 2010.

[3] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, “Basic local alignment search tool.,” *Journal of molecular biology*, vol. 215, no. 3, pp. 403-10, Oct. 1990.

[4] S. Suzuki, T. Ishida, K. Kurokawa, and Y. Akiyama, “GHOSTM: A GPU-Accelerated Homology Search Tool for Metagenomics,” *PLoS ONE*, vol. 7, no. 5, p. e36060, May 2012.

[5] H. Li and R. Durbin, “Fast and accurate short read alignment with Burrows-Wheeler transform.,” *Bioinformatics (Oxford, England)*, vol. 25, no. 14, pp. 1754-60, Jul. 2009.

[6] B. Langmead, C. Trapnell, M. Pop, and S. L. Salzberg, “Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.,” *Genome biology*, vol. 10, no. 3, p. R25, Jan. 2009.

[7] P. J. Turnbaugh, R. E. Ley, M. a Mahowald, V. Magrini, E. R. Mardis, and J. I. Gordon, “An obesity-associated gut microbiome with

increased capacity for energy harvest.," *Nature*, vol. 444, no. 7122, pp. 1027-31, Dec. 2006.

[8] T. F. Smith and M. S. Waterman, "Identification of common molecular subsequences.," *Journal of molecular biology*, vol. 147, no. 1, pp. 195-7, Mar. 1981.

[9] W. J. Kent, "BLAT--the BLAST-like alignment tool.," *Genome research*, vol. 12, no. 4, pp. 656-64, Apr. 2002.

[10] W. R. Pearson, "Searching protein sequence libraries: comparison of the sensitivity and selectivity of the Smith-Waterman and FASTA algorithms.," *Genomics*, vol. 11, no. 3, pp. 635-50, Nov. 1991.