GPU を用いた時系列データ解析向けの Smith-Waterman 法の高速化手法の提案

須戸 里織
Saori SUDO

1　はじめに

近年，脳機能解析の発達により，fNIRS（functional Near-InfraRed Spectroscopy）や fMRI のデータ解析に注目が集まっている．特に，fNIRS のデータ解析においては，アミノ酸配列や塩基配列の相同性検索に利用される SW（Smith-Waterman）法を利用する取り組みが行われている [1]．

一方で，SW 法は並列性が高く，処理に時間を要するという特徴があり，GPU（Graphics Processing Unit）をはじめとした並列計算アーキテクチャにより高速化が盛んに行われている．しかし，既存の GPU 向け SW 法アプリケーションは，時系列データの解析には適していない．

本研究報告では，時系列データ解析向けの SW 法を提案し，GPU を用いて高速化を行う．

2　SW 法

SW 法は 2 つの文字列の共通部分列を抽出するための手法である．SW 法は比較する文字列と同じサイズの行列を用いて類似度を計算する．この行列をアラインメント行列と呼ぶ．SW 法は様々な改良アルゴリズムが存在しており，本章では一般的に用いられる SWG（SW-Gotoh）法と，本研究報告で取り扱う PSSW[2]（Parallel Scan SW）法について述べる．

2.1　SWG 法

一般的に，FPGA や GPU など，CPU と比べて複雑な命令を苦手とするアーキテクチャでは，SWG 法が用いられる．

入力に長さ m，n の 2 つの文字列（データベースシーケンス $D = d_0d_1d_2...d_{m-1}$ とクエリシーケンス $Q = q_0q_1q_2...q_{m-1}$）が与えられたとする．$sub(d_i, q_j)$ を，BLOSUM62 をはじめとするアミノ酸置換テーブルにおける d_i と q_j の相同性スコアとする．相同性スコアとは，2 つの文字の類似度を数値で示したものである．さらに，G_{init} と G_{ext} を，それぞれギャップ開始ペナルティ，ギャップ伸縮ペナルティとする．以上の定義により，SWG 法による (m, n) のアラインメント行列は次の 3 つの式によって定義される．

$$E_{i,j} = \max\{H_{i,j-1} - G_{init}, E_{i,j-1} - G_{ext}\}$$
$$F_{i,j} = \max\{H_{i-1,j} - G_{init}, F_{i-1,j} - G_{ext}\}$$
$$H_{i,j} = \max\{H_{i-1,j-1} + sub(d_i, q_j),\ E_{i,j},\ F_{i,j},\ 0\}$$

$0 \leq i \leq m$ および $0 \leq j \leq n$ において，$H(i,0) = H(0,j) = E(i,0) = E(0,j) = F(i,0) = F(0,j) = 0$ と定義される．上記の計算をアラインメント行列のすべての要素に対して行ったあと，行列中で最大の $H(i,j)$ を D, Q 間のアラインメントスコアとする．

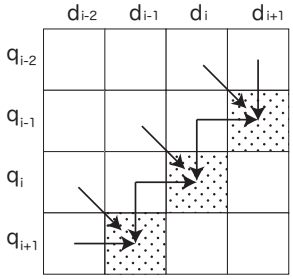

Fig.1　SWG 法のアライメント行列におけるデータの依存関係．各々のセルは H・F・E の値を保持し，矢印方向に示す依存関係がある．

上記の式より，$H(i,j)$ の値は Fig. 1 に示す依存関係を持つ．アラインメント行列の計算は，要素間の依存関係の矢印と直交する斜め方向の要素間で並列性をもつ．よって，Fig. 1 でドットで示した要素は同時に計算が可能である．

この手法はアラインメント行列の左上と右下に近づくにつれて，同時に計算可能な要素が減少する．また，比較する 2 つ文字列長が異なると，同時に計算可能な最大要素数は，2 つの入力文字列長の短い方になるという特徴を持つ．この特徴を考慮して，効率的に SWG 法を実行するためには，最低でも文字列長が 10^4 以上でなければならない．fNIRS の解析に用いられるデータは，一般的に 10^4 より短いため，SWG 法を用いて解析を行うことは効率的ではない．

2.2　PSSW 法

PSSW 法は，SWG 法の並列性を改良した手法である．PSSW 法を用いると，SW 法を行に沿って並列化することができる．PSSW 法による (m, n) のアラインメント行列は次の 4 つの式によって定義される．

$$F_{i,j} = \max\{H_{i-1,j} - G_{init}, F_{i-1,j}\} - G_{ext}$$
$$\tilde{H}_{i,j} = \max\{H_{i-1,j-1} + sub(d_i, q_j),\ F_{i,j},\ 0\}$$
$$\tilde{E}_{i,j} = \max_{0<k<j}\left\{\tilde{H}_{i,j-k} - kG_{ext}\right\}$$
$$H_{i,j} = \max\left\{\tilde{H}_{i,j}, \tilde{E}_{i,j} - G_{init}\right\}$$

上記の式より，$H(i,j)$ の値は Fig. 2 に示す依存関係を持つ．アラインメント行列の計算は，各行の要素間で並列性をもつ．よって，Fig. 2 でドットで示した要素は同時に計算が可能である．

3　実装

SWG 法より効率的に，fNIRS のデータ解析を行うことができる PSSW 法を用いて GPU による高速化を行う．

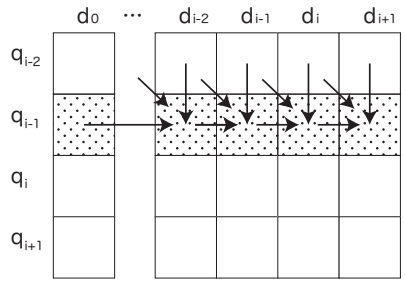

Fig.2　PSSW 法のアライメント行列におけるデータの依存関係．各々のセルは $H \cdot \tilde{H} \cdot F \cdot \tilde{E}$ の値を保持し，矢印方向に示す依存関係がある．

3.1　CUDA

本研究報告では，NVIDIA 社が提供している GPU 向けの統合開発環境である CUDA を用いた．CUDA を用いることで GPU を並列計算機として利用することができる．CUDA では，C 言語の構文に加えて，ホストとデバイス (GPU とビデオメモリ) のメモリ確保, 解放, データ転送といったデバイスの操作に関する機能の拡張が行われている．

CUDA を利用すると GPU は **grid**，**block**，**thread** の 3 つの単位で管理される．複数の **thread** の集まりを **block** といい，複数の **block** の集まりを **grid** と定義する．1 つの **grid** は 1 つの GPU に対応する．GPU は複数の Streaming Multi Processor(以降 SM) からなり，1 つの SM は 1 つの **block** に対応する．SM は複数の Streaming Processor(以降 SP) からなり，1 つの SP は 1 つの **thread** に対応している．各 **thread** に割り当てられた処理は同時に実行され，並列処理が実現される．

CUDA では実行の際に Warp という単位で同じ命令が行われ同時に処理される．Warp は 32 個の **thread** からなるため，**thread** 数は 32 の倍数の場合が望ましい．Warp 内の各 **thread** が分岐命令でそれぞれが違う方向に分岐を繰り返していくと，並列処理が行われないため GPU の性能低下を招く．

3.2　PSSW 法の CUDA 実装

PSSW 法で用いるアラインメント行列の行を，Warp である 32 個の **thread** に割り当てて，各々 $H \cdot \tilde{H} \cdot F \cdot \tilde{E}$ について計算する．データベースシーケンス長が 14400 であれば，1 個の **thread** で 450 要素を並列に計算する．

Table1　使用マシン構成

OS	Ubuntu 12.10
CPU	Core i5-2400 3.10GHz
Main Memory	8GB
GPU	GeForce GTX 480
開発環境	CUDA 5.0
コンパイルオプション	-O3

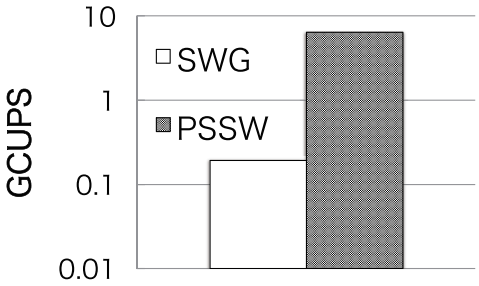

Fig.3　SWG, PSSW 法における性能評価

4　評価

SW 法の性能評価は GCUPS（Giga Cell Updates Per Second）という指標が広く用いられている．クエリーシーケンスの長さを $|Q|$, データベースシーケンスの長さを $|D|$, 実行時間を t とすると，GCUPS は $(|Q| \times |D|)/(t \times 10^9)$ によって与えられる．

本研究では，実行時間 t に，データベースシーケンスとクエリーシーケンスを転送する時間，アラインメントスコアを求める時間，および計算結果をホスト側に返す時間を含んでいる．

性能測定は，fNIRS の出力データを文字列化したものについて行う．データベースシーケンスにシーケンス長 14440 のものを，クエリーシーケンスにシーケンス長 600 のものを用いた．ギャップペナルティーは $G_{init} = 5$ と $G_{ext} = 5$ とし，置換スコア *sbt* は，文字が一致したならば 1，不一致ならば –1 とした．これらのパラメータは，fNIRS のデータ解析に最適とされているものである．

評価を取得したマシンの構成を Table 1 に示す．Fig. 3 に SWG 法と PSSW 法の評価を示す．Fig. 3 より SWG 法に比べて PSSW 法は約 33 倍の性能向上がみられた．

5　まとめと今後の展望

本研究報告では, 時系列データ向けの SW 法を GPU を用いて高速に計算する手法について述べた．また，GPU を用いて，SWG 法と比較して PSSW 法が約 33 倍の性能向上がみられた．今後の展望としては，fNIRS のデータの特徴に焦点を当てた SW 法の高速化手法の提案が考えられる．

参考文献

1) 廣安知之, 西井琢真, 吉見真聡, 三木光範, 横内久猛. 相同性検索を用いた 2 つの時系列データからの類似部分抽出手法と DTW による類似部分の評価. 情報処理学会研究報告. MPS, 数理モデル化と問題解決研究報告, Vol. 2010, No. 24, pp. 1–6, 2010-09-21.

2) Ali Khajeh-Saeed, Stephen Poole, and J. Blair Perot. Acceleration of the smith-waterman algorithm using single and multiple graphics processors. *J. Comput. Phys.*, Vol. 229, No. 11, pp. 4247–4258, June 2010.