C-009

Smith-Waterman 法計算における省メモリ化アーキテクチャの開発
Development of an architecture to reduce memory space for Smith-Waterman method

杉江 崇繁 *
Takashige Sugie

伊藤 智義 †
Tomoyoshi Ito

1. まえがき

ncRNA に未知の重要な新機能が多く存在することが知られている [1][2]．また，非転写領域もゲノム上に存在し，それらの割合は高度な知的生物ほど高くなる傾向にある．ヒトにおいてはコード領域は数パーセント程度であるため，解析領域は膨大になる．解析には様々な手法があり，それらの中で最も重要な解析手法の内の一つに類似度検索がある．類似度検索は複数の生物配列間において類似性の強弱を求めるものであり，生物配列長が $O(n)$ である 2 本の生物配列間の検索では，Smith-Waterman 法のようなダイナミックプログラミング (DP) 法を用いるとその計算量は $O(3n^2)$ になる．加えて，RNA の解析では二次構造を考慮して類似度検索を行う必要がある．これには多次元の DP 法を用いるため，解析時間は非現実的な時間を要する．そこで我々は現在主流であるストリーミング命令に特化した計算機では DP 法に対して不向きである点に注目し，計算機科学の視点から超高速な類似度検索システムの開発を行ってきた [3]．Smith-Waterman 法を用いたペアワイズアラインメントによるグローバルマッチングでは，Intel 社製 Core2 Duo 2.66GHz を搭載した Personal Computer(PC) に対して約 200 倍の高速化に成功している．

二次構造予測を視野に入れた解析には，ペアワイズアラインメントに比べて時間的及び空間的複雑さが指数的に増加する．例えば Sankoff アルゴリズムを用いた共通二次構造予測では 4 次元の DP 行列が必要になるため，生物配列長を $O(n)$ とすると時間複雑さは $O(n^6)$ となり，空間複雑さは $O(n^4)$ になる．どちらも現実的に破綻しているが，特に空間複雑さは現在の計算機技術では実装することすら困難にしている．本論文ではこれまでに開発しているアーキテクチャのメモリアクセス部分において演算部と異なる数式を適用することにより，メモリをほぼ半減させるアーキテクチャを開発したので報告する．

2. Smith-Waterman 法

ペアワイズアラインメントでは比較対象となる 2 本の生物配列を用いて，図 1(a) のような 2 次元の DP ネットワークを構成する．ここでは，ペアワイズアラインメントを行う 2 本の生物配列は”GKFD” と”GFSD” である．左上のノードから任意のノードまで，関与する全てのノードに対して次式を適用することで，そこまでの類似度を得ることができる．

$$S^H_{i,j} = Max(S^H_{i,j-1}, S^D_{i-1,j-1} + g, S^V_{i-1,j} + g) + r \quad (1)$$

$$S^D_{i,j} = Max(S^H_{i,j-1}, S^D_{i-1,j-1}, S^V_{i-1,j}) + W_{i,j} \quad (2)$$

$$S^V_{i,j} = Max(S^H_{i,j-1} + g, S^D_{i-1,j-1} + g, S^V_{i-1,j}) + r \quad (3)$$

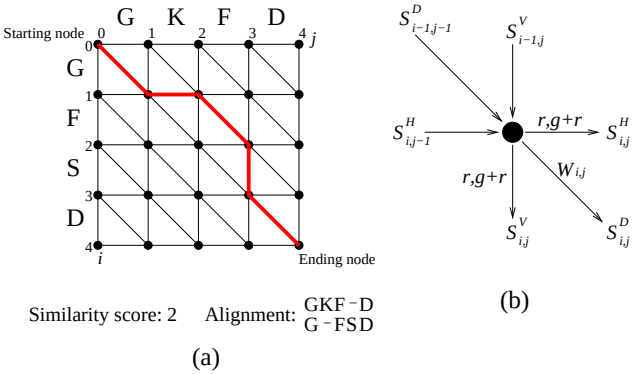

図 1: DP ネットワークとノードにおける類似度の流れ

ここで S は類似度を表し，上付き文字は進む経路の方向 (水平:H, 対角:D, 垂直:V) を，下付き文字はノードの位置を示している．g は開始ギャップ値，r は伸長ギャップ値である．W はスコア行列と呼ばれる，予め定められた類似度が格納されている 32×32 のテーブルである．本来 i, j 番目の文字によって決まる値であるが，便宜上 $W_{i,j}$ と表記する．Max 関数はその中で最も高い類似度を返す．図 1(b) はノードにおける各変数の対応を示している．

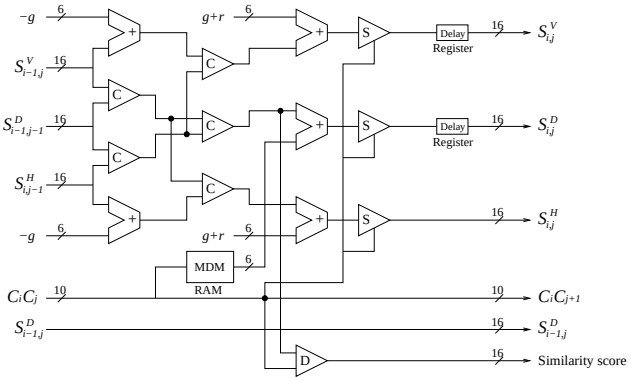

図 2: 演算部のパイプラインブロック図

数式 1，数式 2，数式 3 をパイプライン化すると，図 2 のようになる．図中の C は比較器を示しており，入力された 2 値のうち大きい値を出力するものである．S は選択器を示しており，アルゴリズムによって DP ネットワーク境界における整合性や類似度のゼロクリアを行う．D は検出器で，目的となる類似度の検出を行う．このアー

*東京工科大学, Tokyo University of Technology
†千葉大学, Chiba University

キテクチャは図 1(a) において斜め方向に計算を進めたときに，大部分のノードにおいてパイプラインストールを発生させない．類似度を検索する生物配列長が長くなると，演算結果を Random Access Memory(RAM) などの記憶装置に保存しておく必要が生じる．図 2 のアーキテクチャでは $S^H_{i,j}, S^D_{i,j}, S^V_{i,j}, S^D_{i-1,j}$ が該当する．つまり，生物配列長を n とすると $4n \times 16bit$ の空間が要求される．

3. アーキテクチャ

ここで数式 1，数式 2，数式 3 を熟考してみる．これらの数式はノードから出力される全ての方向への計算式である．3 方向への計算結果が同時に出力されるため常に計算可能状態を維持することができ，効率の良いベクトル演算器を構成することができる．しかしながら，必然的に 3 方向の解を一時的に格納する RAM が必要になる．そこで，3 方向の解を計算するのではなく，本来求めるべきノードの類似度までを計算することにする．この計算式は数式 4 で与えられる．

$$S_{i,j} = Max(S^H_{i,j-1}, S^D_{i-1,j-1}, S^V_{i-1,j}) \quad (4)$$

次に，数式 4 から数式 1，数式 2，数式 3 と同等の類似度を復元する．始めに数式 1 と数式 3 について考える．これらはノードに到達する経路の方向によってギャップコストが決定される．したがって，数式 4 を計算すると同時に，それぞれの Max 関数において最大値を獲得する経路を記憶しておけばよい．これはノードを通過する際の，方向の変化情報だけなので 1bit である．

数式 2 では Max 関数の後にスコア行列から得られる類似度を加算する．先の数式 1，数式 3 の復元を考えると，演算前の類似度を RAM に格納することの方が得策である．以上から RAM へのアクセス部分には図 3 のようなアーキテクチャを開発した．

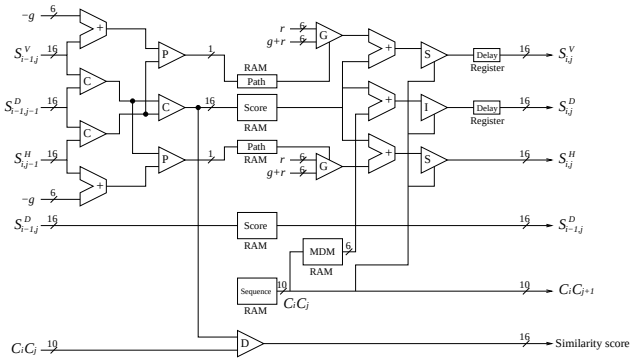

図 3: メモリアクセス部のパイプラインブロック図

図中の P は比較器を示しており，入力された 2 値のうちどちらか一方を基準にして，他方より大きい時は 1 を，小さい時は 0 を出力するものである．G は選択器を示しており，入力された経路パターンにより適切なギャップコストを選択して出力する．I は S の機能に加えて，DP ネットワークの初期ノード (左上) を検出して類似度の初期値を出力することができる．生物配列長を n とすると一時的に格納する類似度は，$2n + 2n \times 16bit$ になる．

このアーキテクチャはメモリアクセス部分にのみ適用すればよいので，演算部のパイプラインはそのまま利用することが可能である．したがって，演算部のアーキテクチャを崩すことなく，パイプライン処理を実現しながら計算に必要なメモリ空間をほぼ半減することに成功した．

本論文のアーキテクチャによる計算量への影響は次のようになる．図 2 のプロセッサエレメントの数を N とすると，これまでのアーキテクチャでは約 N 倍の高速化を獲得することができる．しかしながら，本論文で提案しているアーキテクチャではスコア行列を参照する前の時点で類似度を RAM へ格納してしまうため，スコア行列の参照に必要な比較対象の文字を失ってしまう．従来のアーキテクチャであれば RAM へアクセスする前に一連の計算手順が完結しているため，計算を開始するノードを N だけ進めることができた．本論文で提案しているアーキテクチャを用いると，スコア行列の参照に必要な比較対象の文字を再度，RAM から獲得しなければならない．したがって，生物配列が格納されている RAM から取り出す文字は N だけ進めた位置のものではなく，$N-1$ だけ進んだものになる．つまり，$N-1$ 倍の高速化に計算性能が低下する．このように欠点も存在するが，メモリ空間を半減できる効果の方が遙かに大きい．ただし，スコア行列に用いられている類似度の語長の RAM(図 2 や図 3 では $n \times 6bit$) を用意して，スコア行列の参照結果を格納しておけば N 倍の計算性能のままである．

4. まとめ

本論文で提案しているアーキテクチャを開発したことにより，DP 法計算に必要なメモリ空間をほぼ半減することに成功した．これは非常に大きなメモリ空間を必要とする DP 法に対して有効であり，二次構造を考慮に入れた類似度検索などの，多次元 DP 法計算機の実現に向けての前進ということができる．今後は指数的に増大する時間的及び空間的複雑さに対してさらに効果を発揮するアーキテクチャを開発し，ncRNA や非転写領域などに存在すると言われている未知の新機能の発見へ継げていきたい．

参考文献

[1] The FANTOM Consortium, The Transcriptional Landscape of the Mammalian Genome, *Science* **309** pp.1559-1563, 2005.

[2] RIKEN Genome Exploration Research Group and Genome Science Group (Genome Network Project Core Group) and the FANTOM Consortium. Antisense Transcription in the Mammalian Transcriptome. *Science*, Vol. 309, pp. 1564-1566, 2005.

[3] T. Sugie, T. Ito and T. Ebisuzaki, A Special-Purpose Computer for exploring similar biological sequences: Bioler-2 with multi-pipeline and multi-sequence architecture, *Comput. Phys. Comm.* **162(1)** pp.37-50, 2004.