LH_005

ピーク性能で計算速度比10,000倍を達成する相同性検索専用PCクラスタシステム

PC cluster system which is 10,000 times faster than a PC at peak performance by special-purpose computers for similarity search

杉江 崇繁* 戎崎 俊一* 青見 文博† 増田 信之† 伊藤 智義† 高田 直樹‡ 下馬場 朋禄§
Takashige Sugie Toshikazu Ebisuzaki Fumihiro Aomi Nobuyuki Masuda Tomoyoshi Ito Naoki Takada Tomoyoshi Shimobaba

1. まえがき

ヒトゲノムが解読されてからおよそ3年が経ち，ポストゲノム計画と呼ばれている解読から解析へ研究対象が移行した現在，FANTOMコンソーシアムと理化学研究所 遺伝子構造・機能研究グループらによって新な事実がもたらされた．これまでヒトゲノムの90%以上は意味を成さないジャンクDNAと呼ばれてきたが，これらの領域にも何らかの機能を果している可能性があることが示された[1]．つまり，3Gベース長にも及ぶ塩基配列から網羅的に検索する必要性を示唆している．一方，現在のジーンバンクには多種多様な生物のゲノム配列が登録されており，それらの総数は100Gベースをも超えている．しかし，日々増大し続けている生物配列に対して我々は十分な解析能力を有していない．そこで，バイオインフォマティクスにおける相同性検索に対して，我々は計算機科学の面から解析能力を向上させるべく研究に取り組んできた．

相同性検索ではBLASTなどの計算を間引くアルゴリズムが解析ツールの主流となっているが，DP法を採用しているSmith-Waterman法の方が検索感度が良いことが知られている．これまでに，我々はグローバルマッチングによるSmith-Waterman法を用いた相同性検索専用計算機: Bioler(BIOLogical sequence explorER)を開発している[2]．Bioler はペアワイズアラインメントにのみ対応し，特にBioler-3ではIntel社製Pentium4 3.4GHzを搭載したPersonal Computer(PC)に対して600倍の高速化に成功している．今回，新たにセミグローバルマッチングによる演算回路を開発し，ユーザが利用可能な演算回路の選択肢を増やした．また，相同性検索専用計算機を用いたPCクラスタシステムを構築し，高感度かつ高速演算を実現したので報告する．

2. 相同性検索専用計算機: Bioler-3

ペアワイズアラインメントでは図1(a)のような2次元のDPネットワークを構成し，左上のノードから任意のノードまでに関与する全てのノードに対して次式を適用することで，そこまでの相同性を示す類似度を得ることができる．

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

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

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

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た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．

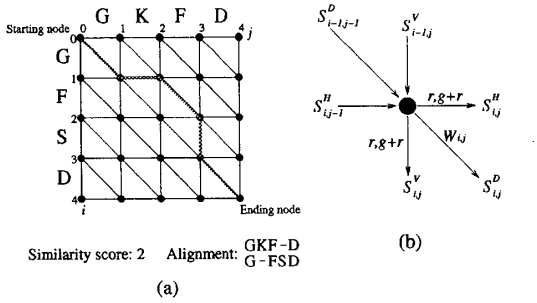

図1: DPネットワークとノードにおける得点の流れ

ところで，次のアラインメントされた文字列は1文字の不一致だけで，大変良く類似した文字列である．し

```
CAGCA-CTTGGATTCTCGG
---CAGCGTGG--------
```

かし，グローバルアラインメントを行うと以下のようにアラインメントされる．これは末端の連続ギャップを計

```
CAGCACTTGGATTCTCGG
CAGC-----G-T----GG
```

算に取り入れているために起きる現象である．生物学的には後者よりも前者の結果を得ることが望ましい．したがって，末端の連続ギャップには異なるギャップ値を用いる必要がある．一般にはそのギャップ値に0を用いる．このようなアラインメントをセミグローバルアラインメントと呼ぶ．

図1(b)から伺えるように，ノードの計算には左上に位置する3ノードからの計算結果が必要である．これらのノードは行列の位置が異っており，多くの計算機アーキテクチャに採用されている擬似ベクトルを用いた演算の高速化が非常に困難な構造をしている．唯一，ローカル

*理化学研究所, RIKEN
†千葉大学, Chiba Univ.
‡湘北短期大学, Shohoku Col.
§山形大学, Yamagata Univ.

マッチングにおいて，ある特定の条件を満したときに擬似ベクトル演算が可能であることが報告されているが，6 倍程度の高速化に留まっている [3]．そこで，我々は専用計算機を開発することにより，大幅な計算速度の向上を行った．

専用計算機のハードウェアには我々が開発した東京エレクトロンデバイス社製 TD-BD-Bioler3 を採用した．これは PCI デバイスとして動作し，演算プロセッサに Xilinx 社製 FPGA XC2VP70-5FF1517C(700 万ゲート相当) を 4 チップ利用できる他，演算回路を動的再構成可能であるため，異なるアプリケーションに対して個々に最適化された回路を用いることができる．論理合成ツールには Synplicity 社製 Synplify-8.2 を，配置配線ツールには Xilinx 社製 ISE-8.1.03i を用いた．グローバルマッチングでは 110 パイプラインを，セミグローバルマッチングでは 90 パイプラインを 1 チップに構成し，133MHz で動作させることができた．OS に FedoraCore1 をインストールした PC に対して専用計算機単体ではそれぞれ 600 倍，500 倍の計算速度である．

3. Bioler-3 を用いた PC クラスタシステム

構築した PC クラスタシステムの構成を図 2 に示す．計算ノード (図 2 における *Node1-Node4*) には Intel 社製 Pentium4 3.4GHz メインメモリ 2Gbyte の PC を 4 台用意し，Bioler-3 をそれぞれに 4 枚インストールした．制御ノード (図 2 における *Node0*) には Intel 社製 Pentium4 2.8GHz メインメモリ 1Gbyte の PC を用意した．これら 5 台のノードは 1Gbps のイーサネットを用いて接続される．OS には FedoraCore1 を採用し，カーネルのみ linux-2.4.32 に差し替えた上で，mpich-1.2.7p1 を用いてクラスタを制御した．制御ノードから計算ノードへの MPI(Message Passing Interface) を用いた通信帯域は 60Mbyte/sec，計算ノードから制御ノードへは 43Mbyte/sec である．本システムでは情報量の多いデータベース配列はそれぞれの計算ノードが保有することにして通信量の削減を図った．したがって，イーサネットにおける通信コストはクエリ配列と類似度である．これらは計算時間に対して十分に小さな値であるため，ノード間通信のオーバーヘッドは無視できる．

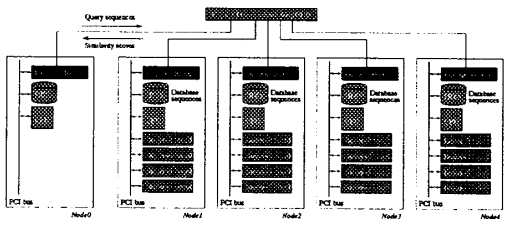

図 2: Bioler-3 を用いた PC クラスタシステム

図 3，図 4 に計算時間を示す．グローバルマッチングでのピーク性能は PC の 11,000 倍に達し，16K ベース長のクエリ配列をヒトゲノム全域に対して検索してもおよそ 10 日で計算が終了する．セミグローバルマッチングのピーク性能においては PC の 9,000 倍となり，同様の検索時間はおよそ 12 日となる．

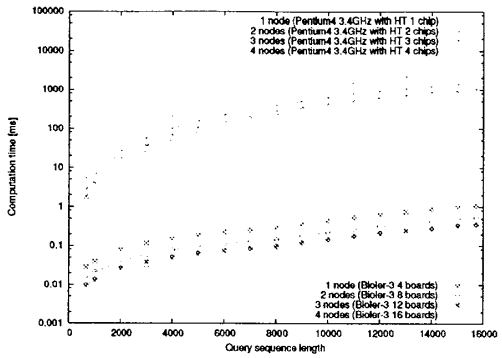

図 3: グローバルマッチングの計算時間

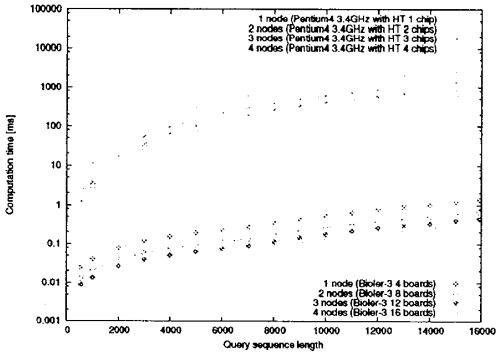

図 4: セミグローバルマッチングの計算時間

4. まとめ

セミグローバルマッチングの演算回路を開発したことにより，幅広い相同性検索を高速に行うことが可能になった．相同性検索専用計算機の PC クラスタ化により PC の約 10,000 倍という圧倒的な計算速度を実現した．今後は 3 つの比較器の増設と類似度検出回路の増強などの改良によってローカルマッチングの回路を開発するなど，さらなる発展を遂げたい．

参考文献

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

[2] 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.

[3] Benner S.A., Cohen M.A. and Gonnet G.H., Six-fold speed-up of Smith-Waterman sequence database searches using parallel processing on common microprocessors, *Bioinformatics* **16(8)** pp.699-706, 2000.