様式Ｃ－１９

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成２４年　６月　８日現在

機関番号：８２６２６
研究種目：若手研究（Ｂ）
研究期間：2010～2011
課題番号：２２７００３１９

研究課題名（和文）　ギガシークエンスデータの高速解析技術の開発

研究課題名（英文）　Developing fast algorithm for analyzing Giga-sequence data.

研究代表者
清水　佳奈　（SHIMIZU KANA）
独立行政法人産業技術総合研究所・生命情報工学研究センター・研究員
研究者番号：６０３６７０５０

研究成果の概要（和文）：ギガシークエンサーは，短い断片配列（リード）を大量に出力するため，高速な解析技術の開発が急務となっている．本研究では，オフセット付き鳩ノ巣原理を応用し，大量のリードから超高速に類似配列を発見するアルゴリズム SlideSort を開発した．SlideSort は従来手法と比較して，同程度のメモリで 1000 倍以上の速度向上を達成した．考案したアルゴリズムの応用例として，最小全域木を構築するソフトウェアの開発も行った．類似ペア検索の応用範囲は広く，上記に述べたクラスタリングの他にも，共通パターンの発見，アセンブリの効率化などに役立つと期待される．

研究成果の概要（英文）：
Next Generation Sequencing (NGS) technology calls for fast and accurate algorithms that can evaluate sequence similarity for a huge amount data. In this study, we designed and implemented exact algorithm SlideSort that finds all similar pairs whose edit-distance does not exceed a given threshold from NGS data, which helps many important analyses, such as de novo genome assembly, identification of frequently appearing sequence patterns and accurate clustering. In comparison to state-of-the-art methods, our method is much faster in finding remote matches, scaling easily to tens of millions of sequences. Our software has an additional function of single link clustering, which is useful in summarizing NGS data for further processing.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
２０１０年度	1,300,000	390,000	1,690,000
２０１１年度	1,200,000	360,000	1,560,000
総　計	2,500,000	750,000	3,250,000

研究分野：情報学
科研費の分科・細目：生体生命情報学
キーワード：ゲノム，ギガシークエンサー，アルゴリズム

１．研究開始当初の背景
近年，ゲノム解読の速度が飛躍的に向上し，医療，環境，エネルギー問題など，様々な分野での応用が強く期待されている．ギガシー

クエンサーは，従来型と比較して約 300～1000 倍の塩基配列決定能力を有する．技術改革は現在も進んでおり，数年以内にさらに 1000 倍の性能を実現すると予測されている．このデータ量は圧倒的であり，バイオインフォマティクスの分野で蓄積されてきた様々な配列解析技術の多くでは扱うことができない．現状では得られたデータの約 3～4 割程度が何の解析もされずに捨てられている．また，解析の対象となった配列もゲノムに貼り付けるなどの簡単な処理がされるにとどまっている．そのため，生物学的な解析に重要な情報量を落とすことなく，効果的にデータを絞り込む技術の開発が非常に重要になる．

２．研究の目的

本研究では大量のリードを高速にかつ効果的にクラスタリングするアルゴリズムを開発することにより，クラスタの代表点のみに対する解析から，大量データの解析で得られるのと等価値な結果を得ることを目的とする．データセットの本質を保持しつつも解析対象となるデータの量を大幅に減らすことによって，様々な配列解析技術の適用が可能となり，ギガシークエンサーから得られる生物学的な情報量は大幅に増すであろうと期待できる．

３．研究の方法

本研究では 2 年間で，高速クラスタリングアルゴリズムの開発，及び開発したアルゴリズムの実装と実問題への応用を行った．初年度は，ギガシークエンサーのデータの特徴を十分に検討し，アルゴリズムの開発を行った．また，プロトタイプを実装し，実データ上でテストを行いながら改良を重ねた．次年度は，開発したアルゴリズムをソフトウェアとして完成された形に実装した．ユーザーインターフェースの開発や並列化への対応を行った．完成したソフトウェアは web にて公開した．アルゴリズムの開発，ソフトウェアの実装について，適宜，学会や論文誌などで発表し，研究成果の普及にも努めた．

４．研究成果

本研究では(1)大量のギガシークエンスデータから超高速にペアを発見するアルゴリズム SlideSort を考案し，その応用例として(2)最少全域木を構築するプログラムを開発した．また，これらをまとめて(3)スタンドアロンソフトウェア，C++ライブリ，GUI で操作可能なソフトウェアとして実装し，web 上で公開した．

(1) 高速ペア発見アルゴリズムの開発

考案したアルゴリズムは，編集距離による検索を行う．通常，配列数が N のデータセットから全ての類似ペアを列挙する場合は，N^2 の編集距離計算が必要とされる．本研究では，あらかじめデータセット全体から類似配列群を絞り込んでおき，同じ類似配列群に含まれる配列間にのみ編集距離計算を行うことによって，コストの高い編集距離計算の回数を大幅に削減する戦略をとった．

まず，編集距離が閾値 d 以内の配列間の特徴について詳細に検討し，次に述べるオフセット付き鳩ノ巣原理を明らかにした．

【オフセット付き鳩ノ巣原理】*編集距離が閾値 d 以内の配列間では，配列を b 個（b>d）に分割した場合，少なくとも b-d 個において文字列のパターンがずれ幅 d/2 以内で一致する．*

この原理より，b-d 個の部分文字列が一致する配列群を類似配列群とすれば，編集距離が閾値 d 以内のペアを全て発見することが可能なことが明らかとなった．提案手法では，ブロックに対して深さ優先でソートを繰り返すことにより，効果的に類似配列群を発見する．ソートの計算コストは O(N)であるため，計算性能は飛躍的に向上した．具体的なアルゴリズムを図１に示す．

Algorithm 1 SlideSort

```
1.  function SlideSort
2.      SlideSortRecursive(φ, φ)
3.  end function
4.  function SlideSortRecursive(y , X)
5.      if y = φ then
6.          m ← 1
7.          go to line 26
8.      end if
9.      if |I(X)| < 2 then                          ▷ Pruning by frequency
10.         return
11.     end if
12.     if no strings in I(X)match X with zero offset then
13.         return
14.     end if
15.     if |y| = b − d then
16.         for (i, j) ∈ P_X do
17.             if (i, j) is canonical then          ▷ See equation 2
18.                 if EditDist(s_i, s_j) ≤ d then
19.                     Report (i, j)
20.                 end if
21.             end if
22.         end for
23.         return
24.     end if
25.     m = max(y) + 1
26.     for z = m, ··· , d + |y| do
27.         R ← φ
28.         for −⌊d/2⌋ ≤ r ≤ ⌊d/2⌋ do   ▷ Generate a string pool
29.             R ← R ∪ {s[q_z + r, q_z + r + w_z − 1] | s ∈ I(X)}
30.         end for
31.         Sort and scan R to find the set of new elements X
32.         for all (x_new, z) ∈ X do
33.             SlideSortRecursive(y + [z], X + {(x_new, z)})
34.         end for
35.     end for
36. end function
```

図１　SlideSort アルゴリズム

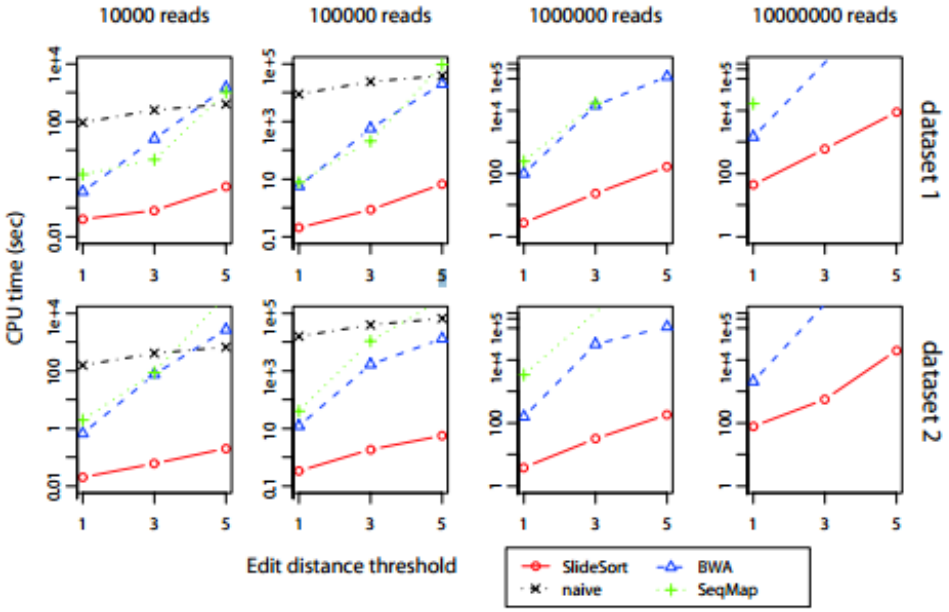

図２　計算時間の比較

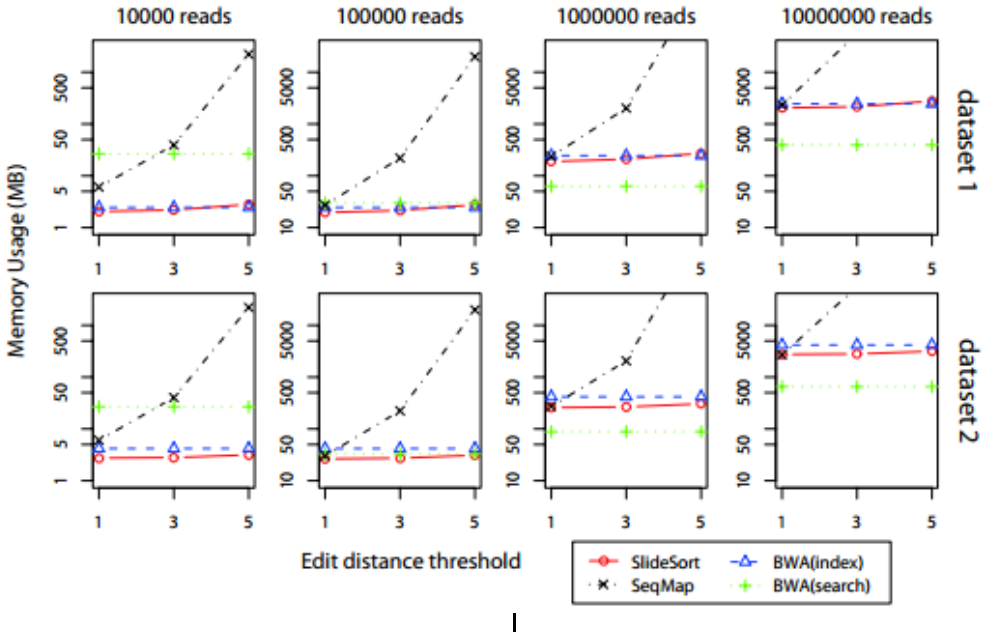

図３　メモリ使用量の比較

また，オフセット付き鳩ノ巣原理に対して改良を加えることにより，通常の編集距離に加え，生物配列の比較に重要なギャップ開始，伸長コストを考慮した検索が可能なアルゴリズムも設計した．
10,000～10,000,000 配列のデータに対して実験を行ったところ，従来手法と比較して数十～数千倍の速度を達成した．（図２）
また，メモリ使用量はメモリ効率が良いとされる Suffix Array を用いた従来手法と同程度であることが確認できた．（図３）データの並列化に対応する手法や，SSE などのビット並列化の技術も導入し，さらに高速に計算可能な仕組みを実装した．
ペアの列挙は，下記に述べるクラスタリングの他にも，類似配列の数え上げや，ゲノムアセンブリなどにも用いられるため，応用範囲の広い基礎的なアルゴリズムを構築できたと考えている．

(2)クラスタリングへの応用
SlideSort の効果的な応用例として，最少全域木を高速に構築するプログラムも開発した．最小全域木の構築では短連結法によるクラスタリングと同じ結果が得られるため，ギガシークエンスデータの解析に直接役立つ．開発したプログラムは，オンラインアルゴリズムを用いており，隣接行列を保持せず，逐次形状を更新しながら全域木を構築するため，メモリを大量に消費しない．リード数 1000 万の実データに対して計算を行ったところ，3G 程度のメモリを用いて 10 分以内に最小全域木を構築することができた．図４に実際に構築された木の例を示す．

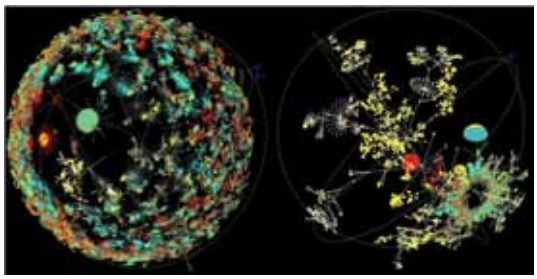

図４　SlideSort によって構築された最少全域木

(3)ソフトウェアの開発
SlideSort を利便性の高いソフトウェアとして実装した．スタンドアロンプログラム，C++言語用のライブラリのほか，GUI により操作可能なソフトウェアも開発した．図５にインターフェースの一部を示す．GUI バージョンは，計算結果を数値だけでなく，グラフなど，視覚的に分かりやすい形で出力する．これらは，Linux, Windows の両方で動作する．

図５　GUI バージョンの SlideSort

５．主な発表論文等
（研究代表者，研究分担者及び連携研究者には下線）

〔雑誌論文〕（計　1 件）
①Kana Shimizu, Koji Tsuda, SlideSort: All Pairs Similarity Search for Short Reads, Bioinformatics, 査読有, 27(4), 2011, 464-470
DOI:10.1093/bioinformatics/btq677

〔学会発表〕（計 4 件）
①Kana Shimizu, Koji Tsuda, Fast and exact algorithm for Next Generation Sequencing data analysis, ISMB/ECCB 2011 Highlights Track, 2011 年 7 月 18 日, Vienna

②Kana Shimizu, Koji Tsuda, SlideSort: A fast and exact tool for finding all similar pairs from next-generation sequencing data, RECOMB 2011, 2011 年 3 月 29 日, Vancouver

③Kana Shimizu, Koji Tsuda, SLIDESORT: All pairs similarity search for short reads, 2010 年日本バイオインフォマティクス学会 年会, 2010 年 12 月 15 日, 九州

④Kana Shimizu, Koji Tsuda, Developing an exact method to find similar pairs with small edit-distance, ISMB 2010, 2010 年 7 月 12 日, Boston

〔産業財産権〕

○出願状況（計１ 件）

名称：配列解析装置，配列解析方法およびコンピュータプログラム
発明者：清水佳奈，津田宏治
権利者：産業技術総合研究所
種類：特許
番号：2010-156342
出願年月日：2010 年 7 月 9 日
国内外の別：国内

〔その他〕
ホームページ等
http://www.cbrc.jp/~shimizu/slidesort/index.php

６．研究組織
(1)研究代表者
清水　佳奈（SHIMIZU KANA）
独立行政法人産業技術総合研究所・生命情報工学研究センター・研究員
研究者番号：60367050