様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

# 科学研究費助成事業　　研究成果報告書

平成　２９　年　　６　月　２１　日現在

機関番号：　３２６８９
研究種目：　挑戦的萌芽研究
研究期間：　2014～2016
課題番号：　２６５４０１５９

研究課題名（和文）類似ゲノムの差異を逃さないDe novoゲノム解析技術の開発

研究課題名（英文）De novo approach to find differentially appearing genome sequence patterns from the two NGS datasets.

研究代表者

清水　佳奈（Kana, Shimizu）

早稲田大学・理工学術院・准教授

研究者番号：６０３６７０５０

交付決定額（研究期間全体）：（直接経費）　　2,600,000円

研究成果の概要（和文）：近年の研究により，ゲノム配列は非常に多様であることが示唆された．しかし，現在主流となっている情報解析の手法では，シークエンサーから出力された断片配列をまずはじめに参照ゲノムに対して貼り付けて，その結果から統計情報を得る方策がとられているため，得られる解析結果は参照ゲノムの特徴に左右されて，ゲノムの多様性を見落としてしまう問題点があった．そこで本研究では，複数のデータセットを直接的に比較して，データセット間で異なる特徴を持つゲノム配列のパターンを発見する手法の設計及び実装を行った．

研究成果の概要（英文）：High-throughput sequencing technology enables to determine various genomes for a same species. Given such a variety of genomes, it is more natural to consider all of such variations. However, majority of analysis method conducts mapping against only a single reference genome in first, which leads to loss of important information caused by mis-mapping. In order to capture individual data’s feature, we developed new approach to analyze NGS data by comparing two different NGS data sets directly and discovering sequence patterns which appears either of the two datasets and do not appear in the other. The proposed approach can be applied to various problems such as finding breakpoints in cancer genomes.

研究分野：　バイオインフォマティクス

キーワード：　次世代シークエンサー　アルゴリズム　アラインメントフリー　ゲノム配列　パターン

１．研究開始当初の背景

国内外では，第二世代シークエンサー（以下 SGS）と呼ばれる超高速な DNA 配列決定装置の普及が進んでいる．SGS により，これまでは不可能であると考えられてきた個人レベル，さらには細胞レベルでの網羅的なゲノム解析が可能となった．これまでの研究では主として，個体差を生みだす個人レベルのゲノムの差異（一塩基多型や構造多型）に注目が集まっていたが，近年は，さらに同一個体における細胞レベルでのゲノムの差異にも注目が集まっている．実験研究の技術開発においては，細胞の一つ一つを分離して，個別にシークエンシングをする技術などの開発が進んでいる一方で，情報解析の技術開発においては，類似するゲノム間の重要な違いを捉える事のできる技術開発に成功しているとは言い難い．SGS のデータ解析の第一段階では，SGS から出力される DNA の断片配列（リード）を参照ゲノムに対してアラインメントするマッピングと呼ばれる情報処理が行われる．SGS では 1 回の実験で出力されるリードの数は数億にも上るため，計算量の問題から，参照するゲノムとほとんど完全に一致するリードのみしかアラインメントすることができない．このため，参照ゲノムには存在しないが実験で計測したゲノム上には存在する“真の配列”を含むリードの多くがマッピングから漏れてしまい，その後の解析に活かされることなく捨てられてしまうという問題が生じている．つまり，参照ゲノムに大きく依存する“まずマッピングありき”の解析手法には，本質的に SGS のデータが持つ豊富な情報量を十分に有効活用できず，その結果，ゲノムの多様性を見落としてしまう問題がある．

２．研究の目的

上記に述べた問題の解決のため，本研究ではシークエンサーから得られたデータのみからできる限りの情報を抽出し，必要に応じて既知ゲノムを利用するという新しい解析の方向性を模索し，ゲノムの多様性を見落とさない解析技術を開発することをめざす．

３．研究の方法

本研究では，参照ゲノムを用いずに，類似リードの頻度による解析を行い，構造変異解析等に応用可能な解析技術を開発した．SGS では，正確性を高めるためにゲノム上の同じ場所を重複して読むため，同一，もしくは非常に類似する配列を持つリードが複数出力される．本研究では，各リードについて類似するリードの本数を数え上げ，様々な解析に応用する．以降，類似リードの頻度を degree と呼ぶこととする．ここでは，がんゲノムを例にとって，ゲノムの構造変異を解析する手法を説明する．がんゲノムでは図 1-(a)の様にゲノムの特定の領域が入れ替わるリアレンジメント呼ばれる現象が起きる．リアレンジメントの起きている場所を特定することは，がんゲノムの解析において重要な課題の一つである．ここで，正常ゲノムから得たリードのセットと，がんゲノムから得たリードのセットを直接比較することを考える．図 1-(b)の最上段が示すように，正常ゲノムにもがんゲノムにも存在する領域から読まれたリードの場合，degree はほぼ同値となる．しかし，リアレンジメントの起きている境目に重なるリードは，がんゲノムにのみに存在する．（図 1-(b)の 3-4 段目）また，シークエンシングエラーなどが原因でゲノム上に存在しない配列が読まれた場合は，双方のゲノムにおいて degree の値は小さくなる．（図 1-(b)の最下段）このように，類似リードの頻度を計算して比較することにより，既知ゲノムへのマッピングをせずに，リアレンジメントを含むリードを特定することができる．

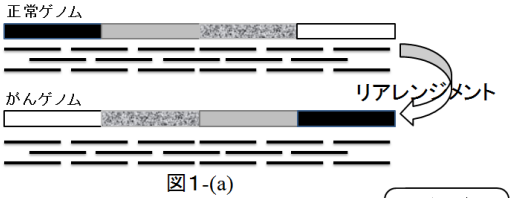


図1-(a)

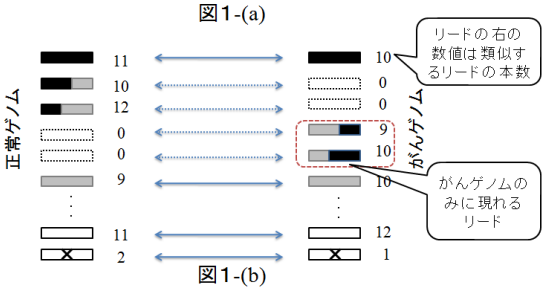


図1-(b)

４．研究成果

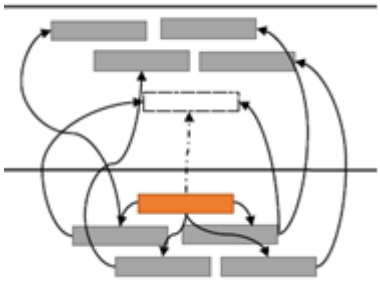
本研究では，上記に述べた類似リードの頻度解析による解析技術を開発した．類似リードの頻度解析を行うためには，膨大なサイズのリードのセットから類似するリードを発見して比較する必要がある．

（1）　SlideSort 法によるアプローチ

まずは，SlideSort 法の応用によって，セットに含まれるリード数に線形な計算量で頻度解析を行うことのできる手法を開発した．SlideSort 法は定められた閾値 $\theta$ 以内の編集距離のペアを入力データに対して線形時間で数え上げる事のできる手法である．SlideSort 法で編集距離が $\theta$ 以内の配列を逐次発見し，類似する配列のパターンをカウントしていくことにより，各セットで頻出する配列のパターンを発見することができる．また，各々のセットで頻出する配列のパターンを数え上げたのち，セットをまたぐ配列間で SlideSort 法により類似するペアを発見することにより，セット間で出現頻度に差異のあるパターンを発見することが可能となる．

セットをまたぐ配列間で類似ペアを探す際には，下記の図で示すように複数の配列パタ

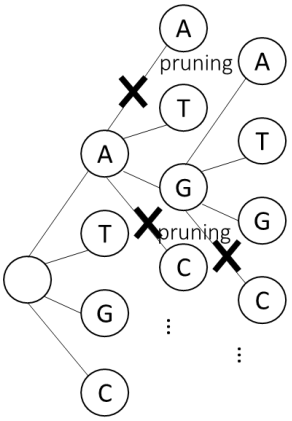
ーンが重複してセット内で類似している配列とペアを組む可能性がある．本研究では，こういった情報を統合して本質的に必要なパターンの数え上げを行うため，セット内の類似パターンの情報をセット間でのパターン比較にも利用する手法を考案した．



提案手法を実装し，複数のデータセットの比較を直接的に行って（例えば，同一患者のガン細胞由来のデータと正常細胞由来のデータの比較）データセット間で差異のある配列パターンを抽出することができるシステムを構築した．

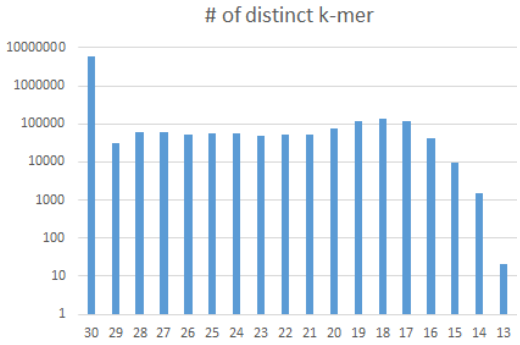
（2） k-mer 数え上げによるアプローチ

SlideSort 法を応用した基本手法ではパターンの数え上げに見落としがない一方で，計算時間が比較的重くなってしまう問題点があった．そこで，大規模なデータに対しても実用的な速度で動作するように，同様の解析をより高速に行うことのできる手法の開発にも取り組んだ．改良した手法では，二つのデータセットでそれぞれ k-mer の数え上げを繰り返し行うことによって，出現頻度に大きな差異のある可変長の部分文字列（以下 differential k-mer）を発見する．各繰り返しでは k-mer の長さを少しずつ伸長していくが，毎回の数え上げの際にデータセット間で頻度の比較を行い，不要なパターンを枝刈りすることにより高速化を図っている．




また，アルゴリズムでの改良の他，並列化の実装を行うなどの工夫も行った．Differential k-mer を用いた方法では，枝刈りの際に，セット間で頻度に際のあるパターンを見逃してしまう可能性も生じる．提案手法をシミュレーションデータ（Chr21，HG19 より 1000 万本のリードを疑似的に生成したデータセットと，Chr21，HG19 に疑似的にリアレンジメントを発生させた人工的なゲノムから 1000 万本のリードを疑似的に生成したデータセットを用意．シークエンシングエラーを想定し，1%の置換/Indel を発生させた．）により評価したところ，90%以上の精度で差異のある領域に重複する配列パターンを発見することができたことを確認した．また，擬陽性は 0.1%以下であった．このように提案手法では十分な精度で複数の NGS データセットの差異を発見することができる事が示された．

開発したプログラムを用いて，実データ（同一患者のガン細胞由来のゲノムと通常細胞由来のゲノム）で評価を行ったところ，以下のグラフのように，データセット間で差異のあるパターンが多数発見された．




本手法では，リードデータを走査しては k-mer を発見し，枝刈りの後に k-mer の長さを伸ばしていく手法を取るため，メモリの量と計算量及びファイル I/O のトレードオフが生じる．開発したソフトウェアでは，合計 700G の fastq によるデータを解析したところ，CPU30 コアの利用で，約 2.5 日程度の時間が必要であった．ピークメモリは 200G 程度であった．本研究では，k-mer を木構造で扱うが，巨大な入力配列に対して，データセット間に絞った配列のパターンは多くの場合で非常に小さくなるため，データの重要な上のみに絞った圧縮ソフトウェアとみなすこともできる．上述の 700G の fastq の場合，解析の結果得られた木構造のデータはわずか 4G バイト程度であった．また，differential k-mer を含むリードはそれぞれのデータセットで 1.7G，2.6G であった．このプログラムの出力である differential k-mer tree や，tree に現れるパターンのみを含むリードに絞って解析することによって，膨大な量のリードを扱うことなく，様々な知見が得るための詳細な解析を行うことが期待できる．

上記ソフトウェアの他，提案手法の基盤技術となる SlideSort 法のプログラム群を整備し，ユーザーが利用可能なように web 上に公開するなどの作業も行った..

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）
Wijaya E, Shimizu K, Asai K, Hamada M, Reference-free Prediction of Rearrangement Breakpoint Reads. Bioinformatics, 2014: 27 (18): 2559-2567.（査読有）

〔学会発表〕（計　3　件）
1. Kana Shimizu, Privacy-preserving genome sequence search, 2016 International Workshop on Spatial and Temporal Modeling from Statistical, Machine Learning and Engineering perspectives (STM2016), July 23, 2016, 統計数理研究所（東京都立川市）

2. Kana Shimizu, SlideSort for comparative NGS data and privacy preserving search for sensitive data, RDF summit, 2016, 2 月 23 日, 東北メディカル・メガバンク機構（宮城県仙台市）

3. 清水 佳奈, 参照ゲノムをなるべく用いない解析/ゲノム秘匿検索, NGS 現場の会 第四回研究会, 2015, 7 月 2 日, つくば国際会議場（茨城県つくば市）

〔その他〕
ホームページ等
https://github.com/iskana/SlideSort

６．研究組織
(1) 研究代表者
清水 佳奈(SHIMIZU, Kana)
早稲田大学 理工学術院 准教授
研究者番号：60367050