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 25 年　3 月 31 日現在

機関番号：12601
研究種目：若手研究（Ｂ）
研究期間：2010～2012
課題番号：22700307
研究課題名（和文）次世代ＤＮＡシークエンサーを用いたゲノム解読のためのアルゴリズム開発
研究課題名（英文）Development of algorithms for genome sequencing using next-generation DNA sequencers
研究代表者
笠原　雅弘（KASAHARA MASAHIRO）
東京大学・大学院新領域創成科学研究科・講師
研究者番号：60376605

研究成果の概要（和文）：次世代ＤＮＡシークエンサーと呼ばれる高速・安価にＤＮＡ配列を読み取ることができる装置を用いてゲノム解読を行う場合には、解読・復元されたゲノム配列が断片化することが多く、各断片がどの染色体上のどの位置に存在するかは分からなかった。そこで、次世代ＤＮＡシークエンサーを用いて短期間に比較的少ない手間で遺伝学的地図を構築するアルゴリズムを開発し、染色体上に解読ゲノム断片配列を整列するシステムを開発した。

研究成果の概要（英文）：Genome sequencing using so-called next-generation DNA sequencers often results in rather fragmented sequences, and the positions of the fragmented sequences on chromosomes are usually unknown. We developed a system/algorithm that constructs a genetic map in a short time with less labor using next-generation DNA sequencers. Genetic maps generated by our algorithm can arrange fragmented sequences on chromosomes.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
２０１０年度	1,600,000	480,000	2,080,000
２０１１年度	700,000	210,000	910,000
２０１２年度	700,000	210,000	910,000
総　計	3,000,000	900,000	3,900,000

研究分野：
科研費の分科・細目：
キーワード：ゲノム，アルゴリズム，ゲノムアセンブリ，遺伝学的地図

１．研究開始当初の背景

大きなゲノムサイズを持つ生物について、ゲノム配列を新規に解読する際には全ゲノムショットガン法が事実上の標準として用いられてきた。特にゲノムサイズが数億塩基対以上である場合には、クローンバイクローン法などの他の手法では時間・労力・費用の問題があまりにも大きく、ヒトゲノムやマウスゲノムなど産業上の大きなニーズがある種を除いて全ゲノムショットガン法が唯一のゲノム解読法といって良い状態であった。

全ゲノムショットガン法ではゲノム配列をランダムに裁断し、裁断された断片配列を並列にＤＮＡ配列シークエンサーで読んでいく。ＤＮＡ配列シークエンサーから出力された配列はゲノム上のランダムな断片配列に相当し、これらの数百万から数千億本の断片配列から類似する配列をコンピューターアルゴリズムによって探し出して結合し、最終的な「解読配列」を出力する。

また、近年のＤＮＡ配列シークエンサーにおける技術改良は目覚ましい。これまで長らく使われてきたサンガー法に基づくＤＮＡ配列シークエンサーと比べ、Illumina 社の

HiSeq2000 などをはじめとする次世代ＤＮＡ配列シークエンサーは数桁高い出力スループットを実現し、ランニングコストも出力塩基あたりで数桁下がってきていた。

このような状況にあって、大きなゲノムを解読する際に次世代ＤＮＡ配列シークエンサーを用いた全ゲノムショットガン法を用いようとするのは必然の流れであると言えよう。しかし、次世代ＤＮＡ配列シークエンサーを用いた全ゲノムショットガン法によるゲノム解読には様々な問題が山積している。

問題のうちの一つはリード長が短いことである。サンガー法に基づいた旧来のＤＮＡ配列シークエンサーは一千塩基対程度の長さのリード（解読配列）を出力できるのに対して、研究開始当時の次世代ＤＮＡ配列シークエンサーのリード長は（スループットの低い一部の機種を除いて）数十塩基対程度にとどまっていた。全ゲノムショットガン法において、リード長が短い場合には高い信頼性を持って連続していると考えられるゲノム解読断片（以降、コンティグと呼ぶ）を長くすることが難しく、高度に断片化したバラバラのゲノム配列しか解読できなくなる。

一般的に、リード長より短い反復配列がゲノム中に存在していてもそれをアルゴリズム的に解決することが可能であるが、数億塩基対以上のサイズのゲノムを持つ生物では、ゲノムに桁違いに多くの反復配列が含まれていることがほとんどである。また、そのような反復配列は短くなればなるほどゲノム中の出現頻度は指数的に増えることが、幾つかのゲノム配列による調査から知られている。

すなわち、全ゲノムショットガン法を用いてゲノムを解読する場合には、ゲノム中に存在する反復配列を乗り越えて正しい復元配列を出力することが重要であるが、高いスループットを実現した次世代ＤＮＡ配列シークエンサーを用いる場合には、その出力するリード長が短いために、乗り越えるべき反復配列の量が指数関数的に「増えて」しまいゲノム配列の復元が難しくなっていた。

実際に、次世代ＤＮＡ配列シークエンサーを用いて大きなゲノム配列を解読する研究が世界各地で行われていたが、数億塩基対以上のゲノムサイズを持つ生物に対して長いコンティグやスキャッフォルド（複数のコンティグをギャップにより結合したもの）を得ることはできていなかった。

２．研究の目的

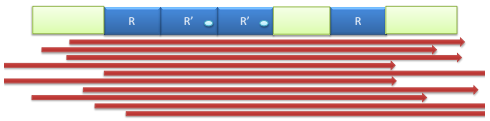
本研究の目的は、数億塩基対以上のゲノムサイズを持つ生物に対して、次世代ＤＮＡ配列シークエンサーを用いて短期間でなるべく安価に新規のゲノム配列解読を行い、高い連続性（コンティグレベル・スキャッフォルドレベル）を持った解読ゲノム配列を得るために有益な様々なゲノム情報処理アルゴリズムを開発することである。

３．研究の方法

本研究では、ゲノム解読のためのアルゴリズムとして大きく分けて二種類のアルゴリズム開発を大きな柱として開発を進めた。

一つ目の大きな柱は、比較的低い一致率のペアワイズアラインメントを高速に発見するアルゴリズムの開発である。次世代ＤＮＡ配列シークエンサーのなかには、PacificBioscience 社の開発するリアルタイム一分子ＤＮＡ配列シークエンサーなど、将来的には非常に長いリード長を実現できると考えられている製品もあるが、そのようなＤＮＡ配列シークエンサーは一分子の計測に基づいているためシグナル／ノイズ比が小さく、出力する塩基列に高い割合でエラーが含まれている。出力塩基におけるエラーの割合は第二世代ＤＮＡ配列シークエンサーと比べて１０倍以上高く、おおよそ１５％程度と見積もられている。このように高いエラー率を考慮し、大規模ゲノム解読に使えるほどの高速なアラインメントツールは研究開始当初には存在しなかった。また、このような一分子シークエンサーの出力に含まれるエラーは従来のエラーと比べて性質が大きく異なる。Illumina HiSeq 等の出力に含まれるエラーは置換エラーが主体であったが、一分子シークエンサーの出力には特に挿入・欠失型のエラーが非常に多い。しかし、従来の高速ペアワイズアラインメントアルゴリズムでは置換エラーを主に考慮したエラーモデルを用いており、挿入・欠失型の高頻度エラーの存在下では急速にパフォーマンスが劣化していた。

一分子シークエンサーから出力されるエラー率は高いが長いリードを、Illumina 等のＤＮＡ配列シークエンサーからアセンブリしたコンティグに高速・堅牢にペアワイズアラインメントすることができるアルゴリズムを開発すれば長いコンティグ・スキャッフォルドの作成に貢献できると考えられた（下図）。



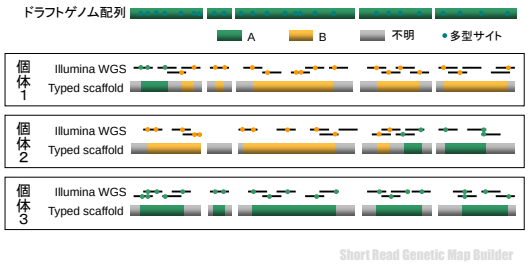
反復配列（R, R′）の集中するゲノム領域を長いリードにより復元する。

二つ目の開発の柱は、遺伝学的地図を用いたウルトラコンティグ（あるいはスキャッフォルド）作成支援アルゴリズムである。ゲノ

ム解読において長いスキャッフォルドを得るためには、既知のゲノム上距離を持ったリードのペアであるメイトペアを読み、コンティグを連結する方法が広く活用されてきた。メイトペアは第二世代以降のＤＮＡ配列シークエンサーでも読むことができるが、長い距離のメイトペアを数多く得るためには非常に長いＤＮＡ断片のみを傷つけることなく大量に精製したり、得られた長いＤＮＡ断片をアダプター付加ののちに自己環状化したりする必要などがあり、そのゲノム解読における効力は限定的であった。特に、５万塩基対を超える長さのメイトペアを次世代シークエンサーで大量に読むことは、研究開始当初だけではなく報告書執筆時点においても実験的に非常に難しく、５万塩基対を超える接続性を別の実験的情報から得る必要があった。

そこで、遺伝学的地図を次世代ＤＮＡ配列シークエンサーにより高速・安価・簡便に作成する方法を提案する。提案手法では、以下に説明する３種類のデータが存在することを仮定する。まず、何らかの手法により得られたドラフトゲノム配列が必要である。ドラフトゲノム配列は次世代ＤＮＡ配列シークエンサーを用いた全ゲノムショットガン法で得られたコンティグ・スキャッフォルド配列で良く、連続性（スキャッフォルドの長さ）は比較的低くて良い。また、対象とする生物について、掛け合わせが可能な二つの系統が存在し、その F_1 個体（コケ等、一倍体が大きい生物種の場合）ないし F_2 個体（一般的な二倍体の生物の場合）が２００個体程度得られることが必要である。

これらの各個体から DNA を抽出し、Illumina シークエンサーを用いて、各個体の非常に薄い（ゲノムサイズの 0.05 倍程度でよい）全ゲノムショットガンシークエンスを行う。Illumina シークエンサーではインデックスランを用いることで１レーンに最大１２サンプルを同時にシークエンスすることができるため、これを利用して合計３ラン未満で各個体の薄いショットガンシークエンスができる計算となる。



得られた薄いショットガンリードをすべてまとめてドラフトゲノムにアラインメントを行い、多型候補サイトを抽出する。また、両ストランドからのサポート、分離比、アラインメントの厚み、両ストランドにおけるアラインメント厚み比など、様々な指標を用いて連鎖解析に耐えうる信頼性の高い多型サイトのみを抽出する。

その後、ドラフトゲノム上では短い距離（おおむね５万塩基対未満程度）でアセンブリのミスが少ないことを仮定して各多型ローカスにおけるジェノタイプの欠損値を補完する。本手法では１個体あたりのショットガンリードが非常に薄いため、半分以上のローカスにおいてジェノタイプが欠損値となるために、欠損値補完のステップは非常に重要となる。

最後に、得られた多型サイトをマーカーとして連鎖解析を行い、マーカーを連鎖群に落とす。従来の連鎖地図と比べてマーカー数が２桁程度大きくなるため、高速な連鎖地図の作成アルゴリズムを開発し対処する。得られた連鎖地図を用いて既存のドラフトアセンブリにおけるミスアセンブリを修正し、連鎖群にスキャッフォルドを載せることで長い連続性を持ったアセンブリを出力することを目指す。

４．研究成果

１分子ＤＮＡ配列シークエンサー情報を活用する高速検索アルゴリズムとして、YASRAT アルゴリズムを開発した。従来のペアワイズアラインメントアルゴリズムでは連続パターン、あるいは非連続な固定パターンをシードとしてハッシュテーブルを作り、検索用のインデックスとしていたが、YASRAT では挿入・欠失を許した特殊な形の穴あきシード群とソートを組み合わせることで従来の検索インデックスを用いるより高速・高精度なシーディングアルゴリズムを開発した。一般的に使われている次世代ＤＮＡシークエンサー用のアルゴリズムと比べると、データベース配列とクエリー配列の双方を同時にソートするため、データベース配列・クエリー配列双方の量がどれだけ増加してもシードヒット数に対してほぼ線形で検索が終了するのが大きな利点であると考えられる。

また、YASRAT アルゴリズムの採用する複数の穴あきシードセットを用いた場合に、シードの一致率について理論的な下限値保証を与えられることを示した。このシーディングアルゴリズムは比較的単純なソートをベースとしているため、分散環境でも実装することができた。

また、ソートステップにおける並列化手法を工夫し、総実行命令数は増えるもののＣＰＵのキャッシュメモリへのヒット率を上げる計算手法を採り入れることによってＣＰＵソケット間の通信帯域をより有効に利用

し、更に高速で実用的なアラインメントアルゴリズムとして用いることができることを確認した。

遺伝学的地図を次世代ＤＮＡ配列シークエンサーを用いることで短期間に作製するためのアルゴリズムを開発した。アルゴリズムの性能確認のために実際の実験データを用いて試験を行った。

基礎生物学研究所の長谷部光泰教授および金沢大学の西山智明助教の協力によりヒメツリガネゴケの２系統、Gransden と Villersexel およびそれらを掛け合わせた分離集団約２００個体の提供を受け、東京大学の菅野純夫・鈴木穣研究室にて Illumina Genome Analyzer によるショットガンシークエンシングを実施し、出力リード配列を我々が解析した。シークエンシングにはインデクシングランを用いて合計でドラフトゲノムサイズの 78.7 倍の塩基配列が得られた。シークエンシングに用いていないインデックスに分類されるリード数は１％未満であり、インデクシングのミスはほとんど考える必要がないことを見いだした。

また、得られたペアエンドシーケンスを、以前にＪＧＩにより発表されていたドラフトゲノム配列と比較して SNP を検出するパイプラインを開発し、連鎖解析に耐えうる頻度・精度の SNP が検出できる条件を見いだした。また、この条件によるフィルター後の SNP 数は５８万となり、ゲノム平均で千塩基対に１カ所の超高密度マーカーを得ることができた。

また、このようにして得られた分離集団個体のアリルパターンをグラフィカルに表示するために、東京大学森下研究室で開発されているＵＴゲノムブラウザを拡張し、ゲノムブラウザ上で分離パターンが見られるようなソフトウェアを開発した。この成果はＵＴゲノムブラウザの本家にマージされ、一般に公開されている。

また、各個体におけるジェノタイプ欠損値を周辺のジェノタイプからノイズを考慮しながら推定するアルゴリズムを開発した。また、このアルゴリズムによる推定結果をＵＴゲノムブラウザを用いて表示・検討したところ、既存のＪＧＩによるヒメツリガネゴケアセンブリには大域的ミスアセンブリが存在していることを見いだした。また、ドラフトゲノム中に存在する汚染（コンタミネーション）配列上には１：１に分離するＳＮＰが検出されないために、ドラフトゲノム中で汚染配列も発見し取り除くことができることを見いだした。

次に、ドラフトゲノム上における近接マーカーはほとんど同じジェノタイプを持っていることを仮定して欠損ジェノタイプの補完を行うアルゴリズムを作成した。また、マーカー間の組み替え価が非常に高く大域ミスアセンブリと思われる領域を切断し、反復的アルゴリズムにより遺伝的に距離が近いマーカーを持つスキャッフォルドをまとめてウルトラコンティグを形成するアルゴリズムを開発した。

この結果、ヒメツリガネゴケのドラフトゲノム中で１００マーカー以上持つスキャッフォルドを３３の連鎖群に結合することができた。大きな連鎖群から順に２６群を取り出すとドラフトゲノムの９０％をカバーすることができ、ヒメツリガネゴケの染色体数は２７とされていることと十分に整合性が高いことが分かった。

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）

〔学会発表〕（計２件）

(1)Masahiro Kasahara, YASRAT: Yet Another Short Read Alignment Tool, The Biology of Genomes, 2011 May 10, Woodbury, The United States of America

(2)Masahiro Kasahara, YASRAT: Yet-Another Sequence Alignment Tool, Joint Cold Spring Harbor Laboratory/Wellcome Trust Conference on Genome Informatics, 2010 Sep. 16, Hinxton, The United Kingdom

６．研究組織
(1)研究代表者
笠原　雅弘（KASAHARA MASAHIRO）
東京大学・大学院新領域創成科学研究科・講師
研究者番号：60376605