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　３０　年　　６　月　１１　日現在

機関番号：　１２６０８
研究種目：　挑戦的萌芽研究
研究期間：　2016～2017
課題番号：　１６Ｋ１４６４１

研究課題名（和文）遺伝子構造解析に適した新規RNA-seq手法の開発およびアセンブル手法の開発

研究課題名（英文）Development of novel RNA-seq and assemble method for gene structure annotation

研究代表者
伊藤　武彦（ITOH, Takehiko）

東京工業大学・生命理工学院・教授

研究者番号：９０５０１１０６

交付決定額（研究期間全体）：（直接経費）　　2,900,000円

研究成果の概要（和文）：本研究では、実験／情報解析双方を用いた新規RNA-seq解析法を開発した。前者では、cDNAの5'を固定後Exo3処理を行うことで、5'は同一かつ3'は様々な配列を持つcDNAを取得し、これらをIlluminaにてシークエンスすることで5'配列をタグとして利用する実験手法を開発した。しかし、インサートが1000bp前後のライブラリ構築に失敗した。後者では前者の実験データを対象にしたアセンブルプログラムを開発することができた。初期の目的は達成できなかったが、本手法は同一cDNAの5'側3'側contigを同時に取得が可能なため、従来手法と組み合わせることで、効果的な遺伝子構造取得が可能になる。

研究成果の概要（英文）：In this research program, we have constructed a novel gene structure annotation protocol using both experimental and computational analyses. In the experimental analysis, the 5′ end of cDNA was protected. Then, the digestion reaction from the 3′ end was performed using Exo3. Finally, from the 5′ end, one sample was obtained, while from the 3′ end various samples were obtained. Next, we constructed an Illumina library from these samples, but we were not successful in the preparation of samples ranging from around 1,000 bp, even using various protocols. In the computational analysis, we succeeded in constructing a novel cDNA assembly program for the abovementioned data, and the performance of the simulation data was satisfactory. We did not succeed in achieving the initial target, but we could obtain 5′ and 3′ end contigs from the same cDNA; this information is very useful for gene identification. Hence, this program can be applied to improve the accuracy of gene identification.

研究分野：　バイオインフォマティクス

キーワード：　遺伝子構造予測　RNA-seq

１．研究開始当初の背景

本研究を開始した 2017 年当時においては、次世代シークエンサとりわけ Hiseq の登場・普及により、100-150bp 程度の短い配列ではあるものの、一度の稼働で数百 Gb にも及ぶ塩基配列が産出される時代になっており、塩基単価も極めて安いため、次世代シークエンサから出力される断片配列を用いて新規ゲノム配列を決定しようとする研究も数多く行われるようになっていた。特に Allpath-LG、SOAPdenovo といった denovo アセンブラプログラムの登場により、Gb オーダーの真核生物ゲノムの新規解読への適用も報告されるようになっていた。

このように次世代シークエンサを用いたゲノム解読は比較的容易になりつつあったが、ゲノム解読後に必要となる遺伝子領域の網羅的な同定は、ゲノムアセンブルほど容易ではなく相変わらず困難であり、その状況は一昔前から変わっていない。

一般的には RNA-seq を実施し、その結果から遺伝子配列を求める手法がよく用いられているが、RNA-seq をアセンブルする方法(Trinity/ SOAPdenovo-trans)、ゲノムにマッピングしてから遺伝子構造を明らかにする方法(TopHat/Cufflinks)どちらの結果も偽陽性の多さや、網羅性の低さの点から不十分である。これらの結果を補完するように、完全なコンピュータ予測である ab initio 法や近縁種の遺伝子情報に基づいたホモロジーベースの予測手法なども用いられているが、精度、感度は低いと言わざるを得ない。

このため遺伝子配列の網羅的な取得には、その生物種の RNA-seq からの遺伝子構築が一番適していると考えられるが、上述の通り既存手法のみでは十分な結果が得られない状況が続いていた。

２．研究の目的

次世代シークエンサを用いた新規ゲノム配列決定は、大型真核生物においても比較的容易に行える時代になっており、種々の研究を展開するにあたり、どのような研究においても必要となる情報の一つに遺伝子情報がある。ゲノム解読後にコードされている遺伝子領域を網羅的に明らかにすることは、避けて通ることのできない基本的な解析であるが、真核生物においてこの問題は、広く用いられている RNA-seq 解析によっても容易には解決できない現状がある。

そこで本研究では、遺伝子構造の再構築に適した新規 RNA-seq 実験手法の提案、および、その実験手法に応じた RNA-seq データのアセンブル手法、さらにはゲノム配列から構築した k-mer de bruijn グラフ上において RNA-seq からアセンブルした contig と合わせてアセンブルすることによる精度が高く網羅性の高い遺伝子構造抽出システムの完成を目的とする。

３．研究の方法

本研究は、新規 RNA-seq 実験手法の開発と、得られた RNA-seq 情報解析手法の開発の大きく二つからなる。前者は、RNA より逆転写された cDNA 配列に対して ND 法を応用し、様々な長さの断片配列を生成、その両端をシークエンスすることで、得られた配列がどの cDNA 由来かを区別することを可能にする実験手法である。

既存の RNA-seq 実験手法はいずれも発現量解析をターゲットとしており、遺伝子構造解析を狙ったものではない。遺伝子構造解析を高精度に実現するためには、遺伝子間で共通に持つ配列の影響を抑えるために個々の遺伝子ごとにシークエンスを実施し、個別にアセンブルすることが好ましい。

しかし、Illumina など次世代シークエンサの利用時に遺伝子ごとにライブラリ調整することは、非現実的である。そこで、以下の様な手法を適用することで、擬似的に遺伝子ごとにタグ付けされた配列の取得を試みる。

まず、RNA に対して逆転写により cDNA を構築する。得られた各 cDNA の 5’末端を修飾保護し、3’側から Exo3 を用いた消化反応を実施する。サンプルを一定時間ごとに回収することで、少しずつ 3’側が削られた様々な長さの cDNA 断片が得られる。これらの cDNA 断片に対して、両端を Illumina により pair-end シークエンスすると、5’端より得られる配列は各々の cDNA に対して全て同じ配列となる。これを「タグ」として read を分類することで、遺伝子ごとに read をクラスタ化することが可能となる。クラスタ内には同一遺伝子の 3’側から少しずつ削られた配列群が分類されており、遺伝子ごとの配列データが取得できるという手法である。

後者は前者で得られたデータに特化したアセンブル部とゲノムを「ガイド」として利用する de brujin グラフベースの RNA-seq アセンブル部からなり、どちらも新規に開発される。この際には、当研究室で開発された高ヘテロ接合性に対応したゲノムアセンブラ Platanus のアルゴリズムを基盤とする。

まず開発された実験手法に基づいて得られた RNA-seq データに対するアセンブル法を開発する。得られた pair-end データを 5’側の配列データに基づいてクラスタリングし、Illumina データを遺伝子ごとに分類するアルゴリズムを開発し、続いて、各クラスタ内に含まれる 3’側の配列データをアセンブルするアルゴリズムを開発する。3’側データは Exo3 による消化時間に応じて、少しずつずれた配列群からなっているため、Illumina データで一般的に用いられている de bruijn

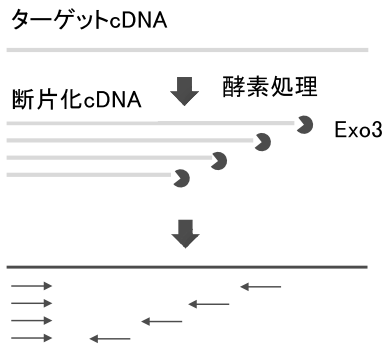
グラフベースではなく、Overlap Layout Consensus アルゴリズムの適用が適していると考えられる。この実現により、各クラスタから数本程度にまではアセンブルにより繋がった、RNA-seq contig が得られると期待される。

続いて、予測対象となるゲノムデータを一旦k-merからなるグラフに分解し、そのk-merグラフ上にある程度の長さにアセンブルされた RNA-seq contig から得られる k-mer グラフを配置して行くことで、ゲノムを「ガイド」として利用した RNA-seq アセンブルアルゴリズムを開発する。

これらの実現により、本研究の目的である高等真核生物を対象とした新規遺伝子構造抽出手法の開発を行う。

４．研究成果

まず、前者の実験手法についての成果を示す。RNA-seq からの遺伝子構造構築にあたり、問題を難しくしているのは、どのシークエンスデータが同一の cDNA 由来の配列であるのかを判断する方法がないことが大きな要因である。その問題を解決するために本手法では、各 cDNA に対して 5’側を修飾保護し、3’側を Exo3 で削り込む事により、5’端は揃った上で 3’側が様々な長さの cDNA 断片を作成し、その両端をシークエンスする事で、5’側の配列を cDNA 毎にユニークなタグ情報として利用する手法を採用している（下図）。

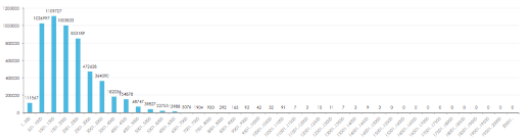


この実験系がうまく働くかを検証する目的で、初めに *S.pombe* ゲノムを EcoRI, HindIII で断片化した二種類の DNA 断片に対してそれぞれ Exo3 ヌクレアーゼ処理を行い、一定時間毎にサンプルを回収する事で、5’側は各制限酵素サイトで固定され、3’側が回収時間に応じて削り込まれているサンプルの作成を行なった。サンプルの取得に当たっては、MiSeq による 300bp のシークエンスを行うことを念頭に、200bp ほど削り込みが進む時間をターゲットとした実験を繰り返して間隔の時間を決定した。

次に、これらの DNA に対して、Illumina の mate-pair ライブラリ作成プロトコルに従い、ライブラリ調整を実施後、Illumina MiSeq による 300bp x 2 のペアエンドシークエンスを行なった。得られた配列データからアダプタ除去などを行なった結果を *S.pombe* のゲノムへとマッピングを実施、そのインサート長を測定した。その結果は以下のグラフに示す通りである。(bin size は 500bp)

EcoRI

HindIII

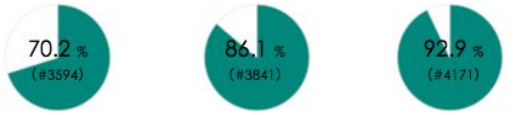


グラフからも明らかなように EcoRI, HindIII どちらで処理したサンプルに対しても、様々な長さの DNA 断片が得られていることが確認できた。しかしながら、0-500bp の分画から得られるデータが非常に少ないことも同時に明らかとなった。

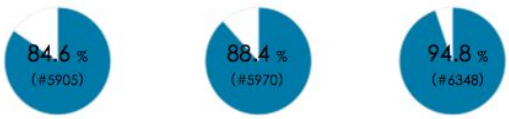
次の手順として、得られた断片配列に対して、総当たりの blastn 解析を実施し、その結果を用いたクラスタリング解析を実施した(99% identity, 50% coverage 条件)。この 5’断片配列によるクラスタリングにより、同一制限酵素区画と思われる配列を回収することに成功した。続いてクラスタ毎に回収した配列を overlap-layout-consensus アルゴリズムに基づいてアセンブルを実施した。この解析によって得られた結果をゲノム配列にアライメントを取ることで、完成度を評価した。

その結果の概要を以下のグラフで示す。グラフは左より、構築率 100%の contig、構築率 90%以上の contig、各制限酵素断片領域から抽出した構築率のもっとも高い contig の内の総塩基数が占める割合を示したものである。

EcoRI



HindIII

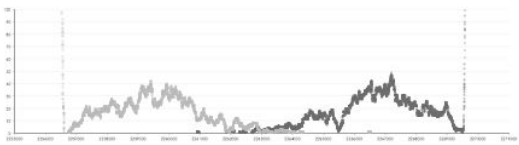


図より全体の EcoRI, HindIII どちらの場合も 80%前後のゲノム領域が再構築できていることが確認できる。

ここまでの成果から、当初の目的である DNA の 5’を修飾保護し、3’端を Exo3 ヌクレアーゼで処理することにより、5’が揃った様々な長さの DNA 断片の作成できること、両端を mate-pair 法でシークエンスし、得られた配列を配列相同性に基づいたクラスタリングにより、同一 DNA 断片由来の配列を回収することができること、クラスタ内の配列を overlap-layout-consensus 法でアセンブルすることで 80-90%程度の被覆率でゲノム配列を再構築できることが確認できた。500bp 未満の配列が取得できないことに関しては、ライブラリ調整の段階で DNA を長さにより二つに分画し、長い分画を mate-pair ライブラリ、短い分画は pair-end ライブラリ構築によりシークエンスすることで問題は解決できると考えられた。

そこでこれらの成果を踏まえ、続いて *S.pombe* の RNA から作成した cDNA を用いた解析を実施した。比較対象として通常のプロトコルによる RNA-seq 解析も実施した。RNA-seq を trinity でアセンブルした結果では 12,125 本が得られたのに対し、本手法ではまず、クラスタリングした段階で 8,200 クラスタに収束した結果が得られた。*S.pombe* の遺伝子は 5,100 程度であるため、本手法の方がより正解に近い形での遺伝子情報再構築が実現できたと考えられる。

しかし、クラスタリング単位での overlap-layout-consensus 解析結果では、1 クラスタあたり 1 アセンブル結果とはなかなか集約することができなかった。ゲノム DNA を用いた時の結果を踏まえ、pair-end, mate-pair 双方でライブラリ調整を行い、合わせてのアセンブルを実施したが、1 本の cDNA を再構成できるケースはあまり多くない結果となった。このため、得られたシークエンスデータからインサートサイズの分布を調べてみると、以下のような結果が得られた。



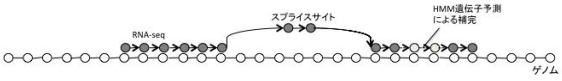
薄いグレーで示されている分布が pair-end ライブラリから得られた結果であり、濃いグレーで示されている分布が mate-pair ライブラリから得られた結果である。図からわかるように 1,000bp 付近のインサートサイズを持つシークエンス量がどちらのライブラリからも得ることが困難であり、これがクラスタ内でのアセンブル結果を 1 本にまとめることの困難さにつながっていると考えられる。

最後に、得られたシークエンスデータに基づいたアセンブル手法の開発について結果を示すこととする。上述した内容も含め、アセンブル手法としては、得られたシークエンスデータから 5’側配列をタグ配列と見立てて、クラスタリングする工程、クラスタ毎に 3’側配列を overlap-layout-consensus アルゴリズムに基づいてアセンブルする工程、予測対象となるゲノムデータを一旦 k-mer からなるグラフに分解し、その k-mer グラフ上に以上の工程によりある程度の長さにアセンブルされた RNA-seq contig から得られる k-mer グラフを配置して行くことで、ゲノムを「ガイド」として利用した RNA-seq アセンブルを行う工程からなっている。ここでは k-mer グラフ上での遺伝子予測を行う工程の成果について述べる。

RNA-seq アセンブル時の大きな問題として、発現量が低い遺伝子由来の k-mer とシークエンスエラー由来 k-mer の判別がつかないこと、発現量が乏しい(＝シークエンスできていない箇所)の適切な処理が必要なことが挙げられる。

前者の問題に対しては、ゲノムを「ガイド」として利用することで、エラー配列に起因した k-mer はゲノム配列に存在しないため、この情報活用によりエラー除去により、予測精度の向上に成功した。

後者に関しては、以前に開発した denovo 遺伝子予測プログラム MetaGeneAnnotator のアルゴリズムを応用し、ゲノムから構築される k-mer de bruijn グラフ上にゲノム配列情報から取得できるコーディングポテンシャルやスプライスサイトらしさを落とし込んだ遺伝子予測を実施することで遺伝子構築を実現するプロトタイプシステムの開発に成功した。また、スプライスサイト周りの k-mer はゲノム中に存在しないため、イントロン由来 k-mer とスプライスサイト周辺 k-mer の間で長さの大きく異なった bubble 構造を作ることになる。この bubble 構造を de brujin グラフ内に許容しながらアセンブルするアルゴリズムに関しては、従来開発した高ヘテロ接合性生物種向けゲノムアセンブラ Platanus 開発時に組み込んだ、ゲノムヘテロ領域に起因した bubble 構造を考慮できる k-mer グラフ作成のアルゴリズムを応用することで解決することに成功した。(下図)



しかしながら、これらの開発したプログラムはクラスタ毎に 1 本ないし数本程度にまでつながった配列を入力を前提として動作するように設計されており、その部分において研究期間内に当初見込んでいた性能を達成することができなかったため、全体を通しての、高精度な遺伝子予測システムの実現には至ることができなかった。

以上を通じ、期初に立てた、高等真核生物ゲノムを対象とし、完全長遺伝子配列の再構成を実験的手法および情報学的手法の双方を組み合わせるという目的を完全には達成することができなかったと言わざるを得ない。しかし、本研究手法で開発した実験手法を用いることで、同一遺伝子の 5'側配列と3'側配列情報の同時取得が可能となる。これらの情報をゲノム配列上にマッピングすることで、各遺伝子のゲノム上での 5’3’両境界を決定することが可能となり、これらの情報を通常の RNA-seq データからに基づいたマッピング解析手法や ab initio 法と組み合わせることによる遺伝子予測精度の向上が十分に期待できると考えられる。特に、ab initio に基づく遺伝子予測精度が低い主な原因はゲノム配列上での binning すなわち、境界領域が不明なことによる原因が極めて大きく、人為的に境界情報を与えた場合のベンチマークテストでは劇的に予測精度が向上することが知られている。そのため、本実験手法を応用することで、遺伝子の境界情報を与えられることの期待度は極めて大きい。

また、情報解析手法の開発において、de bruijn グラフに基づいたアセンブルアルゴリズムに、長さの大きく異なった bubble 構造を許容することが可能となった。このルーチンは、ゲノム配列アセンブルにおいて問題となっている相同染色体間での比較的長い挿入欠失を取り扱う際にも有用であり、これらのルーチンを組み込むことで、開発しているゲノムアセンブラの精度向上にも大きく貢献するなどの一定の成果につながったことも成果の一つとして挙げておきたい。

５．主な発表論文等

〔雑誌論文〕（計　0 件）

〔学会発表〕（計　1 件）

梶谷嶺，吉村大，奥野未来，豊田敦，伊藤武彦　Platanus2: a de novo haplotype assembler enabling comprehensive accesses to divergent heterozygous region. 第 6 回生命医薬情報学連合大会(IIBMP2017) 2017 年

〔図書〕（計　0 件）

〔産業財産権〕

○出願状況（計　0 件）

○取得状況（計　0 件）

〔その他〕
ホームページ等

６．研究組織

(1)研究代表者

伊藤　武彦（ITOH, Takehiko）
東京工業大学・生命理工学院・教授
研究者番号：９０５０１１０６