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

令和　２ 年　 ６ 月　 ８ 日現在

機関番号：１２６０８
研究種目：挑戦的研究（萌芽）
研究期間：2018～2019
課題番号：１８Ｋ１９２８６

研究課題名（和文）Hi-C法を応用した細菌叢からの全ゲノム構築を可能にするメタゲノム解析手法の開発

研究課題名（英文）Development of a metagenomic analysis method that enables whole-genome construction from metagenome using the Hi-C method

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）　　4,800,000円

研究成果の概要（和文）：本研究は、染色体高次構造を明らかにする目的で開発されているHi-C法を、メタゲノム解析に応用し、アセンブル後のビンニングに活用することを目的として実施された。メタゲノムアセンブラにてアセンブルされた配列(Scaffold)に対して、Hi-C法由来のデータをマップし、各Scaffold配列をノード、Hi-Cデータによるリンクをエッジとしたグラフを作成し、Infomap法による段階的な分割を行うことで、既存手法を上回るビンニング精度を持ったツールの開発に成功した。また、本ツールを新規に取得したウシ・ルーメンのメタゲノムデータに適用し、その実用性を確かめた。

研究成果の学術的意義や社会的意義
ある環境を構成する個々の細菌ゲノムの再構築を目指したメタゲノムアセンブルは幅広く実施されているが、その鍵となるのは情報解析手法である。一般的には、シークエンスデータをアセンブル後、得られた配列を特徴量に基づいてクラスタリングすることで分類し、個々の細菌ゲノムの再構築を目指す。様々なクラスタリング手法が開発されているが、アセンブル配列が短い場合には特徴量抽出が困難となり、精度高くクラスタリングすることは原理的に難しい。その点本研究で取り扱うHi-Cデータはアセンブル長に依存しないため、新たな情報量を付与することが可能となり、既存手法との組み合わせによりブレークスルーを与えることが期待される。

研究成果の概要（英文）：The Hi-C method is developed for analyzing the higher-order structure of chromosomes. The purpose of this study was to apply the Hi-C method to metagenomic analysis and binning after metagenome assembly. We succeeded in developing a binning tool exceeding the existing methods in terms of accuracy. This tool uses continuous sequences (scaffolds) assembled by the metagenomic assembler and Hi-C-derived sequence data as an input and provides the binning result as an output. First, the paired-end sequence data derived from the Hi-C method are mapped to the scaffolds. Then, a graph with each scaffold sequence as nodes and links by the paired Hi-C data as edges is created, and this graph is then solved using the Infomap method. We also confirmed the practical applicability of this method by applying it to newly acquired metagenomic data of the bovine rumen.

研究分野：ゲノム情報

キーワード：Hi-C法　メタゲノム

様　式　C－１９、F－１９－１、Z－１９（共通）

１．研究開始当初の背景

2004 年 J.Venter らが Sargasso 海細菌群集のメタゲノム解析を Science 誌に発表して以来、ある環境に生息する細菌群集からゲノム DNA を直接回収・解析するメタゲノム手法が注目されている。次世代シークエンサーの普及と相まって、従来の単離培養を伴う細菌研究での扱いが困難であった多くの難培養性細菌情報を網羅的に解析できるメタゲノム解析は、様々な環境へ適用され多くの研究成果を挙げている。しかし現状のメタゲノム研究の多くは、16S を用いた環境を構成する菌種組成の推定や変動を見るなどに止まっている。16S 情報のみが存在し、その菌種のゲノム情報が存在しない場合が多いため、菌種以上の情報は何も得られない。たとえゲノム情報の存在する菌種であっても、環境が異なれば保持する遺伝子レパートリは多様であり、菌種情報からのみでは、ある環境で果たしているその菌の役割は不明である。

ある環境からゲノム DNA を抽出し、全ゲノムショットガン法によるフルメタゲノム解析も行われているが、この場合各シークエンスがどの菌種に由来するかは不明である。そこで一般的には、シークエンスデータをアセンブル後、得られた Contig 配列を特徴量に基づいてクラスタリングすることでグループに分類し、擬似的に元の細菌毎のデータと見做すことで解析を進めることが多い。様々なクラスタリング手法が開発されているものの、菌種毎のゲノム情報を再現することは難易度が高い。この解析にはコドン使用頻度などの特徴量が用いられるが、水平伝搬の影響などから誤った分類を避けることは困難である。これらを打破するために、研究規模の大きいヒト腸内細菌のメタゲノム解析では、まず初めに、欧米中心に数 100 種レベルでの単離した個別菌ゲノム解析を行い、この参照配列へのマッピングにより解析が行われている。しかしこのような解析は研究人口の多い、ヒト腸内だからこそできるものであり、それ以外の環境では手間・コストの面から不可能である。

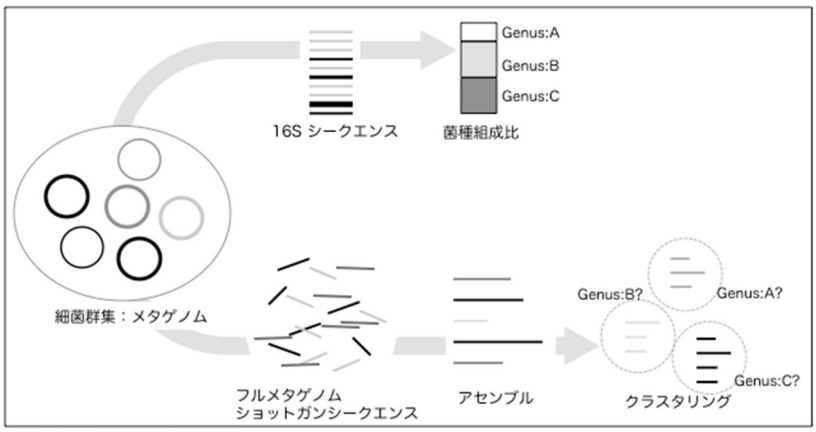


図1:主なメタゲノム解析手法

２．研究の目的

上記のような背景を踏まえ本研究提案では、「フルメタゲノムショットガンシークエンスから、その環境を構成する各種個別細菌の完全ゲノム配列を再構築する手法の開発」を目指す。本手法の適用により、未知な細菌が多くを占めているような様々な環境下においてもメタゲノム解析を用いて、各菌種が環境において果たす役割を遺伝子レベルで明らかにするとともに、産業上有用な菌の単離活用など様々な展開が期待される。背景でも述べたように同様の目的の手法はいくつか提案されているが、短い Contig が大量に生成される状況下では、精度の面で大いに問題のある結果しか得られていない場合が多いため、後述する Hi-C 法などに基づいた新しいタイプのデータを取り入れることで、この問題を打破することを本研究では目的とする。

３．研究の方法

本研究では、Hi-C 法をメタゲノムシークエンス実験に応用し、得られたデータと従来のメタゲノム手法で一般的に用いられている Illumina シークエンスデータとを合わせてアセンブルするための新規情報解析手法開発を行う。

まず、高等真核生物などを対象に、染色体がとる高次構造を明らかにする目的で開発されている Hi-C 実験手法をメタゲノム解析に応用することを試みる。Hi-C 法では、ホルムアルデヒドを用いて DNA 間を架橋固定する。次に制限酵素で DNA を切断後、切断部をビオチン付き塩基で埋めた上でライゲーションを行い、切断後ストレプトアビジンによる回収、Pair-end シークエンスという流れを取る。これにより空間上近接する二点間がライゲーションされた断片に関する情報が得られる。本提案では、ホルムアルデヒドによる架橋を細菌細胞内で行うことにより、近接する二点ではなく同一菌種由来の離れた二点間を架橋する DNA 断片を回収する目的での細菌細胞内 in situ Hi-

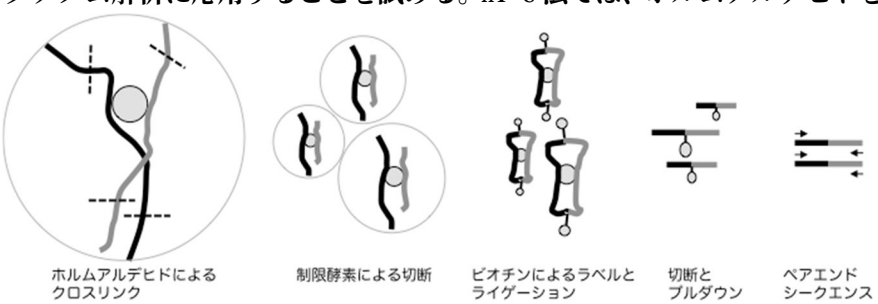


図2:細菌細胞内*insitu* Hi-C法の概略

C 法を試みる。
次に、上記実験手法で得られた同一細菌由来の「キメラ」シークエンス配列と通常のフルメタゲノムシークエンス配列とを入力とし、細菌群集を構成する個別菌ごとに完成されたゲノム配列を出力するアセンブラ開発を進める。今まで開発してきたメタゲノム用アセンブラ MetaPlatanus をベースとし、Hi-C 法によって得られる遠距離間リンク情報を活用した Scaffolding 手法をとることとする。Hi-C データのみで Contig の順序・向きの確定が困難な場合には、insert の大きな mate-pair ライブラリ調整併用も検討する。

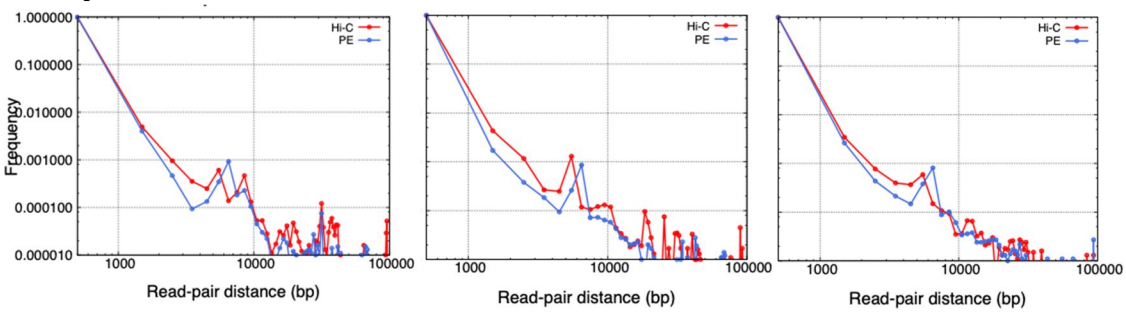
４．研究成果

（１）メタゲノム解析への Hi-C 実験法の適用

本研究では、ヒト糞便３サンプル及びウシのルーメンサンプルへの Hi-C 法の適用を試みた。まず、ヒト糞便３サンプルに対し、Illumina pair-end, mate-pair シークエンスに加えて、Phase Genomics 社の ProxiMeta 試薬キットを用いて Hi-C 法によるライブラリ調整を行い、Hiseq X Ten を用いてシークエンスを実施した。得られたシークエンスデータの概要は右表の通りであり、単位は Gb である。続いて、当研究室にて開発しているメタゲノムアセンブラにて、PE600 から MP9kb までを用いたアセンブルを実施した。アセンブル結果の概要を以下のテーブルに示す。どのサンプルでも約 300-400Mb の Total アセンブルサイズが得られており、N50 は 47kb – 267kb を示している。続いて、Hi-C ライブラリ評価の目的で、アセンブル結果に対し、得られた Hi-C データをマッピングし、pair リードのマップされた位置から推定される挿入サイズの分布を作成した。比較対象のため、挿入サイズがほぼ 600bp に作成されている PE600 ライブラリ由来のデータも同様の解析を行った。その結果を以下のグラフに示す。左から順に Sample1, 2, 3 である。

Sample ID	PE600	MP3k	MP6k	MP9k	Hi-C
1	8.36	9.96	7.78	0.89	11.09
2	7.51	1.74	1.08		10.47
3	9.23	20.72	9.34	8.39	5.49

Sample ID	Total (bp)	# scaffolds	Scaffold-N50 (bp)	Max length (bp)
1	305,468,809	37,543	89,626	2,573,347
2	316,804,572	45,086	47,866	3,323,072
3	390,795,065	33,002	266,990	3,040,465



図は横軸に read-pair 間の距離、縦軸にはその頻度を示したものであり、赤線が Hi-C ライブラリ由来データ、青線が Pair-end ライブラリ由来データである。確かに pair の距離が 1,000bp を過ぎたあたりから赤線の方が上、すなわち pair 間が遠距離のサンプルの割合が増えていることが確認できるが、その差はわずかであり、Hi-C 法によるライブラリ作成では遠距離作用をうまく捉えることができておらず、アセンブルへの寄与率は低いことが想定された。この原因としては、糞便サンプルからのライブラリ調整時に分解が進み、DNA 断片を短くしてしまっていることなどが想定された。

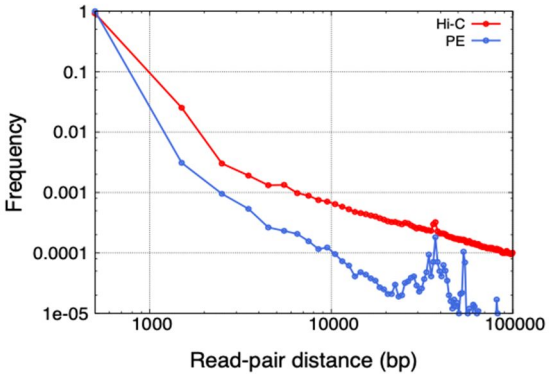
ヒト糞便サンプルからの Hi-C ライブラリ作成の結果を踏まえ、次のウシルーメンサンプルからの Hi-C ライブラリ作成では、DNA 抽出時にビーズ破砕時間を短くするなど、よりマイルドな条件で実験を実施した。ヒト糞便サンプル同様に、Illumina pair-end, mate-pair ライブラリに加え、Phase Genomics 社の ProxiMeta 試薬キットを用いて Hi-C 法によるライブラリ調整を行い、Hiseq X Ten を用いてシークエンスを実施、以降の Hi-C ライブラリ評価解析についても同様の方法を採用した。結果をシークエンス量、アセンブル概要を以下に示す。

total DNA amount (ug)	シークエンシング量 (Gbp)				
	PE	MP			Hi-C
		2k	5k	10k	
19.6	8.67	3.87	4.49	4.09	28.37

Total (bp)	N50 (bp)	Longest (bp)
2,408,688,782	24,895	4,476,563

総シークエンス量はヒト糞便サンプルと大差がないにも関わらず、総アセンブル長は約 2.4Gb とヒト糞便サンプルの 6-8 倍となっており、菌種組成がヒト腸内よりも多様であることが想定される結果となっている。このアセンブル結果に、Hi-C ライブラリ由来、比較対象の PE ライブラリ由来 pair リードをマップし、推定挿入サイズを求めた結果が以下のグラフの通りである。

ヒト糞便サンプルの結果とは明らかに異なり赤線で示された Hi-C ライブラリ由来のデータの線が青線より上に分布しており、遠位のコンタクトを捉えられている事が示唆される結果を得ることに成功した。リードペアが異なる scaffold にマップされる割合も、Pair-end ライブラリでは 9.2%なのに対し、Hi-C ライブラリでは 26.0%となっており、Scaffold のさらなるアセンブリやビンニングに Hi-C ライブラリ由来のデータが活用され得ることを示唆する実験データの産出に成功した。



（2）Hi-C 法由来データを活用した、メタゲノムアセンブリ／ビンニングツールの開発・解析

（1）で得られたデータを入力としたメタゲノムアセンブリ／ビンニングツールの開発を実施した。期初は、我々が開発しているメタゲノム用アセンブラ MetaPlatanus への Hi-C データに基づいたビンニング機能の組込みを想定していたが、他のアセンブラなど様々な入力に対応した方が望ましい点、研究開始後 Hi-C データをビンニングに用いる手法が他研究グループより報告されたため、その手法とのアルゴリズム・精度比較が可能となった方が望ましい点などにより、メタゲノムアセンブラによるアセンブル結果を入力とした Hi-C 法データを活用したビンニングツールを完成させ、既存手法を上回る精度を達成することを目指した。また、MetaPlatanus についても改良を加え、Hi-C 法データを入力とはしないが、更なる性能の向上を試みた。

① MetaPlatanus の改良

開発中のメタゲノム用アセンブラに対し、Contig アセンブル時により低 coverage のものを積極的に採用するために MEGAHIT の結果を de Bruijn グラフを経由して統合する機能、異種間ミスアセンブルを防ぐ機能（塩基使用頻度及び coverage depth 情報の活用）、配列をより伸長するための反復機能(Scaffolding, Gap-closing, Contig-merge を繰り返す)、MetaBat2 を用いてビンニングする機能、各 bin で再度 scaffolding する機能などを追加することにより、既存のアセンブラよりもより長く、精度高くアセンブルすることに成功した。

② Hi-C 法に基づいたビンニングツールの開発

ビンニングとは、メタゲノムアセンブル結果である Scaffold を、その環境を構成する菌種ごとのゲノムに分類することを指し、一般には同一菌種由来の Scaffold 同士は同様の振る舞いをすることが想定されるため、Scaffold から各種統計量を取得し、その統計量に基づいた分類を行う。統計量にはコドンや k-mer の頻度情報、GC 含有量、Sequence Coverage 情報などが用いられる事が一般的である。しかし入力となる Scaffold が十分な長さを持たない場合には計算される統計量に誤差が大きくなり、ビンニングの精度を出す事が困難となる。これは根本的な問題であり、別の情報量を入力としない限り解決は難しい。その際の活用できるデータとして、本研究では Hi-C 法データを採用しようと試みた。

以下、本研究において開発したビンニングツールの概略を説明する。前提条件とし、同一菌種由来の Scaffold は空間的に近接し、近い Coverage の値を取るものとしている。また、画一的な条件によるビニングでは、複数菌種由来のゲノムが混じって生じた大きなビンを避ける事ができなかったため、段階的にフィルタリング・ビンニングを行うことでその欠点の解消を試みた。開発した手法の具体的手順を以下に示す。

まず、Hi-C 法により得られたデータを各 Scaffold にマッピングし、その結果を集計、Scaffold をノード、リンクの本数をエッジの重みとするグラフを構築する。その際に重みの小さいエッジは信頼度が低いとして除去する事で、偽陽性を減らす工夫を採っている。続いて、得られたグラフを Infomap 法により分割する。Infomap 法はデータ圧縮技術を応用したグラフ分割アルゴリズムの一つであり、Ravasz 法や Girvan-Newman 法などと比べて計算量が $O(NlogN)$と少なく済むメリットがある。また、同様の計算量となる Louvan 法と比べても解像度限界に達しにくいとの特徴があり、実際に両者を試しても Infomap 法によるものの方が高い精度の結果を得られたため、Infomap 法を採用した。分割により得られた各ビンについて、一般的なバクテリアのゲノムサイズより大きいものについて、ビン内のエッジをフィルタリングし、再度グラフ作成・分割を実施する機能及び、カバレッジ情報を利用することで各ビンに少量混在している別菌種由来と思われるゲノムを分離する機能を実装することにより、高精度なビンニングの実現に成功した。

③ ベンチマーク結果

開発した手法は、正解がわかっているシミュレーションデータに対してベンチマークを行う

ことで精度評価を実施した。用いたデータは Mathew *et al.*2019 論文で用いられていたもので、63 種のゲノム既知バクテリア（全菌種合計 263Mb）から計算機上で仮想的に作成されたメタゲノム Illumina データおよび、sim3C により作成された Hi-C シミュレーションデータである。ベンチマークは先行研究(bin3C)と同じく、上記リードをアセンブルして得られている Scaffold に対して実施された。入力した Scaffold データは合計 240Mb であり、N50 は 30.4kb である。この入力データに対して、本研究において開発したビンニングツールおよび先行研究である bin3C を適用し、得られたビン毎に CheckM というシングルコピーオーソログ遺伝子セットを用いた完全性(Completeness)、重複率(Contamination)を指標とした評価ツールで完成度評価を実施した。その結果を以下の表に示す。なお、Nearly Complete とは Completeness 90%以上かつ Contamination 5%以下、Substantially とは同 70%以上かつ 10%以下、Moderately とは同 50%以上かつ 15%以下、High Contamination とは Contamination 15%以上を指し、この指標も先行研究と同じものを用いている。結果を右の表に示す。bin3C と比べてわずかではあるが nearly complete のビン数が多くなっており、40/63=63.5%のビンについて、ほぼ完成されている。また、high contamination の bin 数が 0 と bin3C と比べて大きく減少していることも本開発ツールの大きな特徴である。ビンニング結果に複数菌種由来のデータが混じっている「Contamination」は、さらなる下流において大きな障害となる事が想定されるため、この bin 数が 0 であることは大きな利点となり得る。

	# nearly	# substantially	# moderately	# high contamination
開発ツール	40	4	5	0
bin3C	39	4	6	4

続いて、ヒト糞便サンプル由来の実公開データを用いたベンチマークも実施した。これは Press *et al.* 2017 論文によるもので、MluC1, Sau3A1 制限酵素を用いた Hi-C データも公開されている。ビンニングに用いた Scaffold データは、合計 724Mb、N50= 5.2kb と繋がり具合は今までのデータと比べて短くなっている。上記シミュレーションデータと同様に CheckM を用いた指標でビンニング結果を比較した。なお比較対象には bin3C に加えて、Hi-C 法を用いず塩基の特徴量などに基づいてビンニングを実施する Metabat2 も用いた。結果を以下の表に示す。

	# nearly	# substantially	# moderately	# high contamination
開発ツール	53	28	8	0
bin3C	57	22	13	2
Metabat2	18	23	8	15

結果より、本開発ツールおよび bin3C によるビンニングの結果が Metabat2 よりも良好な成績を収めていることが明らかであり、入力 Scaffold が短い場合には Hi-C 法を用いたビンニングが有効であることが示唆される。また、本開発ツールは nearly complete の数では bin3C に少し及ばないものの、nearly + substantially の数では少し上回っており、またシミュレーションデータで示した最大の特徴である high contamination のビン数 0 を実データでも達成できている。

④ ウシ・ルーメン実データ解析結果

最後に、（１）で取得したウシ・ルーメンの実サンプルデータに対して本開発ツールを適用した結果を示す。Metabat2 よりも多くのビン（moderately 以上で約 2.5 倍）の構築に成功することができた。また、Metabat2 よりも contamination のビンも少なく、本開発手法による Hi-C 法を用いたビンニングの有用性を示すことに成功した。

	# nearly	# substantially	# moderately	# High contami
開発ツール	42	80	74	28
Metabat2	27	34	19	34

以上示したように本研究では、期初に立てた Hi-C 法を応用した実験データの取得およびビンニングツールの開発に成功した。また開発したツールは、既存手法と比べて同等以上の完成度のビンを作成しつつ、Contamination の高いビン作成の抑制に成功した。しかしながら、ヒト糞便サンプルなどで示せた性能と比べてウシ・ルーメンサンプルでは、完成度・重複度双方の点から改善の余地が大きいことを示唆する結果が得られているなど、開発したツールの堅牢性に改良すべき点が認められた。今後は、より多くの実データを取得、本開発ツールを適用し、その結果を踏まえて改良を加えていくことで、幅広いデータに対応した精度の高いビンニングツールの完成を目指す。さらには、ホストゲノムと plasmid, virus ゲノムとの対応関係をつけるなどの可能性も探っていきたい。現在は独立した Hi-C 法を用いたツールとなっているが、最終的には、塩基使用頻度情報などとの統合も図り、アセンブルツールとの再帰的な利用により、メタゲノムアセンブル・ビンニングツールとして完成させ論文化を図る予定である。

５．主な発表論文等

〔雑誌論文〕　計0件

〔学会発表〕　計1件（うち招待講演　0件／うち国際学会　0件）

１．発表者名 Jun Hattori, Takehiko Itoh and Yoshimura Dai
２．発表標題 Comprehensive detection of insertion sequences in bacterial genomes
３．学会等名 第8回生命医薬情報学連合大会IIBMP2019
４．発表年 2019年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考