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　６ 年　 ６ 月 ２５ 日現在

機関番号：１２６０１
研究種目：基盤研究(C)（一般）
研究期間：2021～2023
課題番号：２１Ｋ１２１０４

研究課題名（和文）遺伝子変異を考慮しナノポアシーケンスから高精度にメチル化を検出する情報技術の開発

研究課題名（英文）The development of a high-performance nanopore methylation detection method with consideration of structural variation

研究代表者
張　耀中（ZHANG, Yaozhong）

東京大学・医科学研究所・准教授

研究者番号：６０８１７１３８

交付決定額（研究期間全体）：（直接経費）　　3,200,000円

研究成果の概要（和文）：ナノポアシーケンスデータでの高性能メチル化検出手法をモデルレベルとパイプラインレベルの両方で開発した。モデルレベルでは、Transformerモデルのencoderアーキテクチャを使ってmethBERTを開発した。BERTモデルで塩基配列の表現学習を検証した。同じゲノムlociにアライメントされたリードを統合利用することで新しいメチル化コーラーを開発した。パイプラインレベルでは、ハプロタイプおよびゲノム変異を考慮したメチル化予測パイプラインを構築した。このパイプラインは、正常および腫瘍細胞株を用いて検証された。その上、対照学習を通じて生物学的関係を導入することで、新しい表現学習方法を開発した。

研究成果の学術的意義や社会的意義
ゲノムシーケンシングのコストが安くなるにつれて、その利用も広がってきた。ゲノムシーケンシングデータをより迅速かつ高精度に解析することは、ヘルスケアや疾患診断において重要である。本研究では、ナノポアシーケンシングから高精度なメチル化プロファイリング解析技術を開発した。この技術により、メチル化を高速かつ高精度な検出することが可能になり、老化や疾患におけるエピジェネティックな変化を理解するために役割を果たすことが期待される。

研究成果の概要（英文）：In this project, we developed both model-level and pipeline-level high-performance methylation callers for nanopore sequencing data. We developed methBERT using the encoder architecture of the transformer model. In addition to signal analysis, we investigated the learning of nucleotide representation in the BERT model through pre-training. We analyzed representations for signals and nucleotides and developed a novel methylation caller based on the alignment of reads at target positions. At the pipeline level, we built a haplotype-aware and structural-variant-informed methylation detection pipeline, which we tested on both normal and tumor cells. Besides developing high-performance methylation callers, we extended our findings to whole-genome-level nucleotide sequence representation and single-cell representations using contrastive learning with biological constraints.

研究分野：bioinformatics

キーワード：methylation　nanopore　deep learning

様　式　Ｃ－１９、Ｆ－１９－１（共通）

１．研究開始当初の背景

DNA・RNA・ヒストンのメチル化は、DNA 配列の変化を伴わない、様々な生命現象における遺伝子発現に関わる制御機構である。様々な DNA 領域で異常なメチル化状態が生じることによって細胞の機能が損なわれ、近年では発達障害、老化、およびがんなどの疾患で研究が進んでいる（Nature Reviews Molecular Cell Biology, 2019）。メチル化の状態を正確に検出することは、疾患プロファイリングにおけるバイオマーカーの探索やエピジェネティックな治療法の臨床開発に必須の基本要件である。

本研究では、メチル化検出において優位性のあるナノポアシーケンシングによるメチル化検出に焦点を当てている。ナノポアシーケンシングには多くの利点があるが、それでもなお、そのデータ解析には解決すべき、いくつかの技術的な重要課題が存在する。第一に読み取ったゲノム情報の精度が Illumina 等に代表されるショートリードシークエンシング技術での精度に比べて相対的に低いこと（ナノポア（R9. 4）:85~95% vs. Illumina 99.9%~）。第二に、既存のモデルベースの手法はメチル化が対象領域内の 1 ヶ所のみで起こるという仮定をおいたシングルトンデータを基本として構築されており、複数箇所でメチル化が起こる現実を反映したモデルとはなっていないこと。HMM（Nature Methods, 2017）や　RNN（Nature Comm., 2019）などが多く用いられるが、特定のメチル化タイプやメチル基転移酵素に対して、完全にメチル化されているか、またはメチル化されていないかの 2 択が仮定されている。しかし、この仮定は現実的ではなく、実際のサンプルでは複雑なメチル化の組み合わせを考慮する必要がある。第三に、遺伝子型情報がメチル化検出において考慮されていないことである。

２．研究の目的

本研究では、高速かつ高精度に多様なメチル化の検出を、ナノポアシークエンスデータからゲノム配列の遺伝子型（SNV、インデル、SV）情報を同時に同定しつつ行うことが出来、かつ DNA、および RNA の両方に適用できるディープニューラルネットワークに基づく情報解析技術を構築することを目的とする。

３．研究の方法

本研究では、以下の 4 つのステップに分けて研究を実施する。

(1)データ収集と既存方法の網羅的な比較既存手法との比較を行うために、公開されているナノポアデータを系統的に収集する。データの収集は、様々な種類のメチル化（4mC 、 5mC、 6mA など）やメチル化酵素の種類（M.SssI など)が異なる広範囲なデータが対象となる（フローセル R9.4）。また、クロスプラットフォーム比較を目的として、ナノポアと WGBS の両方で計測されたデータの収集を行う（2023 年度には、最新のフローセル R10. 4 データも追加収集した）。

(2)パイプラインレベルでの遺伝子型情報の統合する。特定遺伝子型を考慮しながらメチル化の同時検出が可能な自動化パイプラインを構築する。ハプロタイプおよびゲノム変異を考慮するメチル化予測パイプラインを構築する。

(3)モテルレベルでの遺伝子型情報を統合する。ナノポアシーケンスデータの、信号および塩基配列の表現を設計する。信号には、k-mer 内の異なる位置で高度に差別化された信号パターンがあるため、これに対応する信号の特徴をモデルで表現する。塩基配列に対しては、異なる k-mer の類似性を考慮し、より良い塩基表現の学習方法をモデル化する。信号と塩基の表現をモデルレベルで統合し、ゲノム変異を考慮したメチル化状態を予測するモデルを開発する。

(4)がん細胞の解析に応用する。 開発した方法を腫瘍細胞株（RKO、COLO829）を用いて検証する。

４．研究成果

本プロジェクトでは、フローセル R9.4 および R10.4 によるナノポアシークンスデータでの高性能メチル化検出手法をモデルレベルとパイプラインレベルの両方で開発した。

(1)　フローセル R9.4 を用いてデータを網羅的に収集し、Transformer モデルの Encoder アーキテクチャを使って methBERT を開発した。このデータセットでの評価に基づき、長距離コンテキスト特徴はメチル化の予測に大きな影響を与えることはなく、目標 motif 周辺の特徴が重要であることが解明された（BIBM, 2021）。

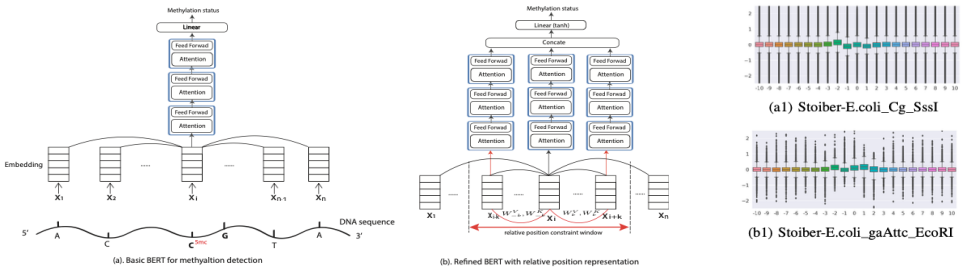

図１．(a-b)methBERT のモデル（a1-b1）motif 周辺の信号の boxplot

(2) 塩基配列における事前学習モデルの構築を行った。事前学習モデルにより塩基配列で学習したものを明らかにするため、embedding 層の k-mer 表現を解釈し、文脈に整合した k-mer 表現を学習したことを検証した(Bioinformatics, 2022, 2023)。比較研究により、先行の BERT モデルが k-mer が重複した領域において一貫した表現を学習することを示した。

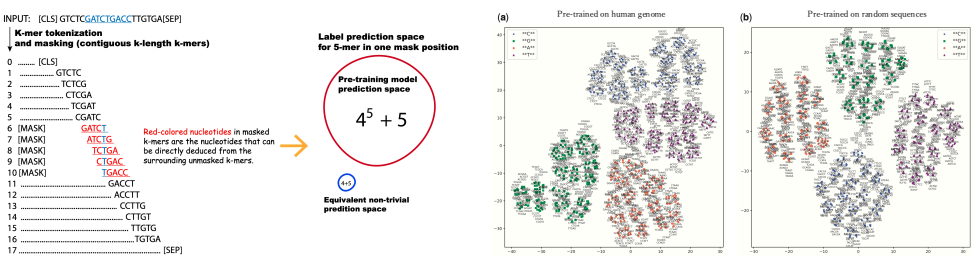

図 2. BERT モデルでは k-mer が重複した領域において一貫した表現を学習すること

(3) K-mer の信号および塩基データを分析し、最新のフローセル R10.4 によるシーケンスデータに基づき、リード信号と塩基配列のアライメントを利用して新しいメチル化コーラーを開発した。開発したアルゴリズムは、同じゲノム loci にアライメントされたリードを統合利用することで、モデルからゲノムレベルのメチル化の検測が可能となった。このモデルは優れた性能を達成しつつ、小さなモデルサイズを維持している。（投稿予定）

(4) パイプラインレベルの開発では、最新のフローセル R10.4 によるシーケンスデータに特化し、ハプロタイプおよびゲノム変異を考慮するメチル化予測パイプラインを構築した。このパイプラインは、正常細胞株（HG002、COLO829BL）および腫瘍細胞株（RKO、COLO829）を用いて検証された。特に、HG002 がタンデムリピート領域におけるリードの特性とメチル化予測の結果について詳細な検証を行った。

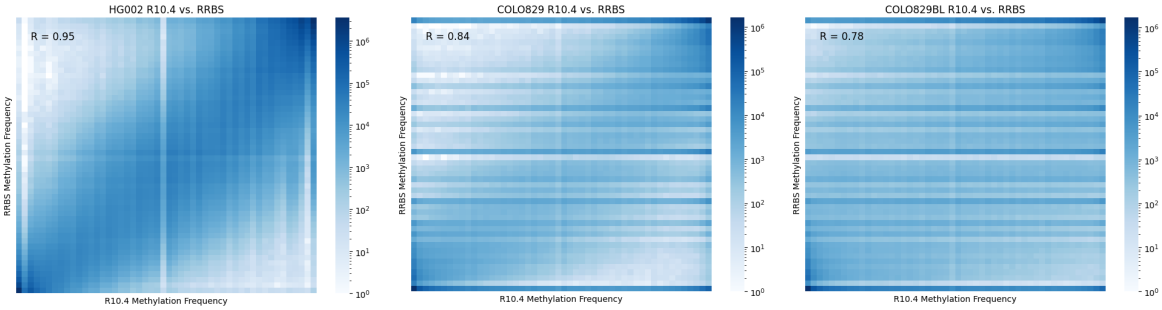

図３．メチル化検出結果の RRBS(Reduced Representation Bisulfite Sequencing)との比較

(5) 高性能メチル化コーラーの開発に加えて、対照学習を通じて生物学的関係を導入することで、新しい表現学習方法を開発した。全ゲノムレベルの塩基配列表現（Briefings in Bioinformatics,2023）およびシングルセル表現の学習に応用した（Scientific Reports,2024）。

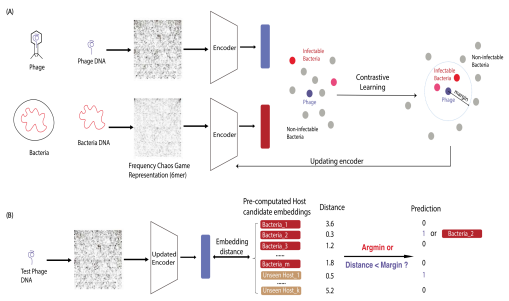

図４．全ゲノムの塩基配列の表現学習の応用

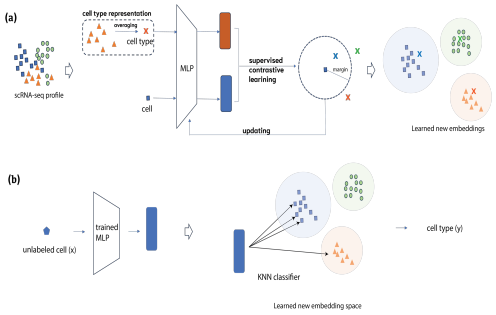

図５．シングルセルの表現学習の応用

５．主な発表論文等

〔雑誌論文〕　計11件（うち査読付論文　9件／うち国際共著　0件／うちオープンアクセス　7件）

１．著者名 Zhang Yao-zhong、Liu Yunjie、Bai Zeheng、Fujimoto Kosuke、Uematsu Satoshi、Imoto Seiya	４．巻 24
２．論文標題 Zero-shot-capable identification of phage?host relationships with whole-genome sequence representation by contrastive learning	５．発行年 2023年
３．雑誌名 Briefings in Bioinformatics	６．最初と最後の頁 1-10
掲載論文のDOI（デジタルオブジェクト識別子） 10.1093/bib/bbad239	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Zhang Yao-zhong、Bai Zeheng、Imoto Seiya	４．巻 39
２．論文標題 Investigation of the BERT model on nucleotide sequences with non-standard pre-training and evaluation of different k-mer embeddings	５．発行年 2023年
３．雑誌名 Bioinformatics	６．最初と最後の頁 1-10
掲載論文のDOI（デジタルオブジェクト識別子） 10.1093/bioinformatics/btad617	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Heryanto Yusri Dwi、Zhang Yao-zhong、Imoto Seiya	４．巻 14
２．論文標題 Predicting cell types with supervised contrastive learning on cells and their types	５．発行年 2024年
３．雑誌名 Scientific Reports	６．最初と最後の頁 1-16
掲載論文のDOI（デジタルオブジェクト識別子） 10.1038/s41598-023-50185-2	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Bai Zeheng、Zhang Yao-zhong、Miyano Satoru、Yamaguchi Rui、Fujimoto Kosuke、Uematsu Satoshi、Imoto Seiya	４．巻 38
２．論文標題 Identification of bacteriophage genome sequences with representation learning	５．発行年 2022年
３．雑誌名 Bioinformatics	６．最初と最後の頁 4264～4270
掲載論文のDOI（デジタルオブジェクト識別子） 10.1093/bioinformatics/btac509	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Zhang Yao-zhong、Bai Zeheng、Imoto Seiya	４．巻 preprint
２．論文標題 Dysfunctional analysis of the pre-training model on nucleotide sequences and the evaluation of different k-mer embeddings	５．発行年 2022年
３．雑誌名 bioAxiv	６．最初と最後の頁 1~7
掲載論文のDOI（デジタルオブジェクト識別子） 10.1101/2022.12.05.518770	査読の有無 無
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Zhang Yao-Zhong、Yamaguchi Kiyoshi、Hatakeyama Sera、Furukawa Yoichi、Miyano Satoru、Yamaguchi Rui、Imoto Seiya	４．巻 IEEE
２．論文標題 On the application of BERT models for nanopore methylation detection	５．発行年 2021年
３．雑誌名 Proceedings of 2021 IEEE International Conference on Bioinformatics and Biomedicine	６．最初と最後の頁 320~327
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/BIBM52615.2021.9669841	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Bai Zeheng、Zhang Yao-zhong、Miyano Satoru、Yamaguchi Rui、Uematsu Satoshi、Imoto Seiya	４．巻 preprint
２．論文標題 Identification of Bacteriophages Using Deep Representation Model with Pre-training	５．発行年 2021年
３．雑誌名 BioAxiv	６．最初と最後の頁 1~7
掲載論文のDOI（デジタルオブジェクト識別子） 10.1101/2021.09.25.461359	査読の有無 無
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

〔学会発表〕　計3件（うち招待講演　0件／うち国際学会　3件）

１．発表者名 Yao-zhong Zhang
２．発表標題 Dysfunctional analysis of the pre-training model on nucleotide sequences and the evaluation of different k-mer embeddings
３．学会等名 27th Annual International Conference on Research in Computational Molecular Biology（国際学会）
４．発表年 2023年

１．発表者名
Yao-zhong Zhang

２．発表標題
On the application of BERT models for nanopore methylation detection

３．学会等名
IEEE International Conference on Bioinformatics and Biomedicine（国際学会）

４．発表年
2021年

〔図書〕　計0件

〔産業財産権〕

〔その他〕
bert investigation
https://github.com/yaozhong/bert_investigation

methBERT open source software
https://methbert.readthedocs.io/en/latest/index.html

CL4PHI open source software
https://github.com/yaozhong/CL4PHI

SCLSC open source software
https://github.com/yaozhong/SCLSC

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

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関