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　５ 年　 ６ 月　 ７ 日現在

機関番号：１４４０１
研究種目：基盤研究(B)（一般）
研究期間：2019～2021
課題番号：１９Ｈ０４２０７

研究課題名（和文）細胞動画像とオミクスデータの統合的情報解析技術の開発

研究課題名（英文）Development of integrated analysis technology for bioimaging and omics data

研究代表者
瀬尾　茂人（Seno, Shigeto）

大阪大学・大学院情報科学研究科・准教授

研究者番号：３０４３２４６２

交付決定額（研究期間全体）：（直接経費）　　13,400,000円

研究成果の概要（和文）：近年の生命科学データはマルチモーダル化が著しく、様々な様式（モード）のデータが大量に取得されるようになっている。オミクスデータはゲノムに由来する情報ゆえにその量が簡単に人間の理解を超え、また動画像は一見理解しやすいものの粒子や細胞の認識・追跡といった処理が必要となり定量的な解析を行おうとすると難しい。本研究では、細胞動画像とオミクスデータの統合的情報解析技術の開発を目的とし、動画像解析には深層学習の技法を用いた特徴量の抽出技術の開発、オミクスデータ解析としては一細胞RNA-seqのデータを中心に様々なタイプのデータを結合して、重要な基底の発見とパターンの抽出を行う方法を開発した。

研究成果の学術的意義や社会的意義
本研究では、大別して2つの方向性の要素技術の開発を行った。1つは、一細胞遺伝子発現解析のための次元削減・特徴抽出手法であり、非負値行列因子分解や変分自己符号化器による方法を開発した。もう1つは、細胞動画像から特徴量の抽出を行う方法である。相貌度画像の解析においては、十分なアノテーション情報を準備することは難しいため、自己教師つき学習や教師なし学習を用いて、重要な画像特徴量を獲得することを試みた。また、本研究課題の期間中に急速に発展普及した技術が空間トランスクリプトーム解析である。本研究課題でも実際にデータを取得し、開発した要素技術の空間トランスクリプトーム解析への応用・評価を行った。

研究成果の概要（英文）：In recent years, life science data has become increasingly multimodal, with large amounts of data in a variety of formats (modes) being collected. The amount of omics data is easily beyond human comprehension due to its genomic origin, while microscopic images are difficult to analyze quantitatively due to the need for processing to recognize and track particles and cells. In this study, we aimed to develop an integrated information analysis technology for cellular video images and omics data. For video image analysis, we developed a feature extraction technology using deep learning techniques, and for omics data analysis, we developed a method to combine various types of data, mainly single-cell RNA-seq data, to discover important features and extract patterns.

研究分野：バイオインフォマティクス

キーワード：バイオインフォマティクス　遺伝子発現解析　細胞画像処理　機械学習

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

次世代シーケンサー技術やイメージング技術、加えて自動化技術の発展により、ゲノム情報やトランスクリプトーム情報、生体分子構造情報、細胞や個体の時空間動態情報など様々な様式（マルチモーダル）の生命情報ビッグデータが日々蓄積されている。一方で、これらの複雑で膨大なデータ間の関連性を、人間が直感的に理解することはほぼ不可能である。多数の細胞の動的状態を高分解能・広視野で撮影した 4D イメージングの動画像は、そのデータ量と時空間的複雑さゆえに、細胞移動の軌跡を目で追うことすら困難であり、オミクスデータとして取得される 1 細胞ごとの遺伝子発現やエピジェネティクスの情報は並べてみようにも画面に収めることも難しい。

これらの情報の増加から導かれることは、人間の代わりに情報を把握し関連性を結び付け、理解可能な形で提示するための情報処理技術の必要性であった。本研究課題の「問い」は、これらの膨大で複雑なデータを統合し、知識発見や現象の理解を行うためにはどのようすれば良いかということである。特に、因果の上流にあると考えられる遺伝子発現を中心としたオミクスデータと、下流である細胞動画像とを、統合して解析を行うための方法論を開発することを目指した。

２．研究の目的

本研究では、細胞動画像とオミクスデータの統合的情報解析技術の開発を行う。動画像解析には深層学習の技法を用いた特徴量の抽出技術の開発を行い、オミクスデータ解析としては一細胞 RNA-seq のデータを中心に様々なタイプのデータを結合して、重要な基底の発見とパターンの抽出を行う方法を開発する。深層学習のメリットの一つとして、End-to-End で学習を行うことができるというものがあるが、単に動画像を入力、オミクスデータを目的として学習を行うには目的変数が多すぎる。またその逆もしかりである。すなわち、データを統合して解析するためには、それぞれを高次元のまま扱うのではなく、各データの持つ特有の構造・パターンを抽出・分類していったん抽象化し、その後にパターン間の類似性同士を比較するのが良いと考えた。このアイデアに基づき、細胞動画像とオミクスデータをそれぞれに低次元の特徴量やパターンとして縮約する。そして、動画像の自動定量化と複雑なオミクスデータの統合とを実現し、お互いの情報をリンクすることで知識発見を実現することを目的とした。また生物学を専門とする研究分担者とともに実際の生物学的データを用いたケーススタディを行い、データ駆動的に新たな知識の発見を行う方法論の構築を目指した。

３．研究の方法

大別して 2 つの方向性の要素技術の開発を行った。1 つは、一細胞遺伝子発現解析のための次元削減・特徴抽出手法であり、非負値行列因子分解や変分自己符号化器による方法を開発した。もう 1 つは、細胞動画像から特徴量の抽出を行う方法である。細胞動画像の解析においては、十分なアノテーション情報を準備することは難しいため、自己教師つき学習や教師なし学習を用いて、重要な画像特徴量を獲得することを試みた。

また、本研究課題の期間中に急速に発展普及した技術が空間トランスクリプトーム解析である。空間トランスクリプトーム解析は、組織や臓器の切片等における位置情報を保存しつつ、遺伝子発現解析を行うことができる技術であり、細胞画像や組織染色画像とあわせてデータを取得可能である。本研究課題でも実際に空間トランスクリプトーム解析のデータを取得し、開発した要素技術の応用や評価を行った。

４．研究成果

(1)　非負値行列因子分解に基づく一細胞発現データのクラスタリング手法

一細胞遺伝子発現解析では、教師なし学習による細胞のクラスタリングが細胞種の同定、細胞の多様性や亜集団の発見に重要な役割を果たしている。識別された細胞クラスタは、その後の解析にも利用され、例えば、発現遺伝子の違いの発見や、細胞分化の系譜を推測すること等が可能である。しかしながら、遺伝子発現プロファイルは、発現量の定量化手法によって異なる結果が得られることが知られている。本研究では、同一の一細胞 RNA-Seq データから、異なる手法で定量化された複数の遺伝子発現プロファイルを利用して、非負値行列因子分解（non-negative matrix factorization; NMF）によるロバストで高精度なクラスタリング手法を提案した。行列分解はデータの次元削減や特徴抽出に優れた手法であり、特に NMF は全ての値が非負である 2 つの行列の積としてデータ行列を近似する。本研究で開発した joint-NMF は、分解された行列の 1 つが複数の NMF の間で共有されているという制約の下で、各 NMF を複数の遺伝子発現プロファイルに適用することで、それらに共通する因子を抽出することが可能である。提案手法では、単一の遺伝子発現プロファイルのみを用いた従来のクラスタリング手法と比較して、複数の定量化手法を活用することで、よりロバストかつ正確な細胞のクラスタリング結果を得ることができた。また、抽出した特徴量を用いたマーカー遺伝子発見への有用性も示した。

(2)　変分自己符号化器による次元削減とクラスタリングへの応用

一細胞遺伝子発現データは、高次元かつスパースで多くのノイズを含む。よって、一細胞遺伝子発現解析を行う前に前処理として、ノイズを減らし、データを低次元に落とし込む次元削減を行うことが重要である。本研究では、深層学習に基づく次元削減手法の１つである変分自己符号化器を用いることで、一細胞遺伝子発現プロファイルを対象とした次元削減を行なった。従来の一般的な変分自己符号化器では、潜在変数の事前確率分布にガウス分布を仮定するのに対し、本研究では潜在変数の事前確率分布として混合ガウス分布を仮定した。これにより柔軟かつ非線形な表現を保持した次元削減が可能になった。図１は変分自己符号化器によって獲得された特徴量を t-SNE によって可視化したものである。図左は 3 層、図右は 1 層のときの特徴量であり、3 層の方が細胞の種類を分類するのに適切な情報を獲得していると考えられる。また、次元削減が正しく行えているか検証するため、潜在変数に対してクラスタリングを行なった。そして、そのクラスタリング精度から、本研究をクラスタリングへの応用に用いられることを示した。さらに、変分自己符号化器の潜在変数とデコーダーを用いることで、各細胞種のマーカー遺伝子の発見と擬似遺伝子発現プロファイルデータ生成への応用の可能性を示唆した。

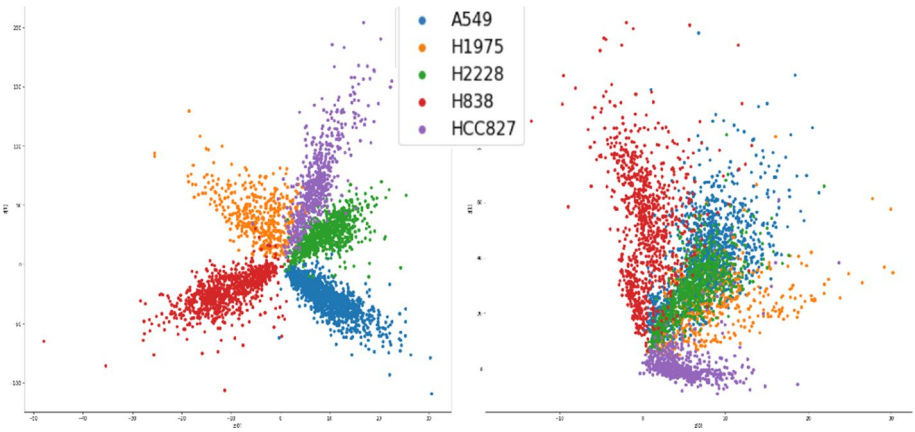

図１．変分自己符号化器による特徴量の可視化

(3)　畳み込みニューラルネットワークを用いた細胞画像における繰り返しパターンの検出手法

本研究では、1 枚の顕微鏡画像には同種の細胞が複数写っていることが頻繁にあることを利用して、アノテーションを行わずとも 1 枚の画像から必要な特徴を学習する方法を開発した。具体的には、Deep Feature Factorization (DFF)を用いて画像中に含まれる特徴量を抽出し、その特徴量をグラフ化した後にグラフマイニングを行うことで繰り返しパターンを検出する方法を提案し、顕微鏡画像へ応用し有効性を評価した（図 2)。その結果として、1 種類の細胞や微生物のみを含む顕微鏡画像では VPRRS に比べて物体検出の検出漏れが少なくなり、IoU も高くなることを示した。また、テンプレートマッチングに比べて細胞や微生物で個体間に大きさ、色、向きなどで目に見える差異があれば高い検出をできることを示した。2 種類の細胞や微生物を含む顕微鏡画像では 1 種類の時に比べて検出精度は下がるもののある程度検出できることを示した。さらに、大きさの異なる別々の繰り返しパターンも検出できることを示した。DFF において畳み込みニューラルネットワークから特徴量を得る際に、その層を変えることで検出精度に影響を及ぼすことがわかった。

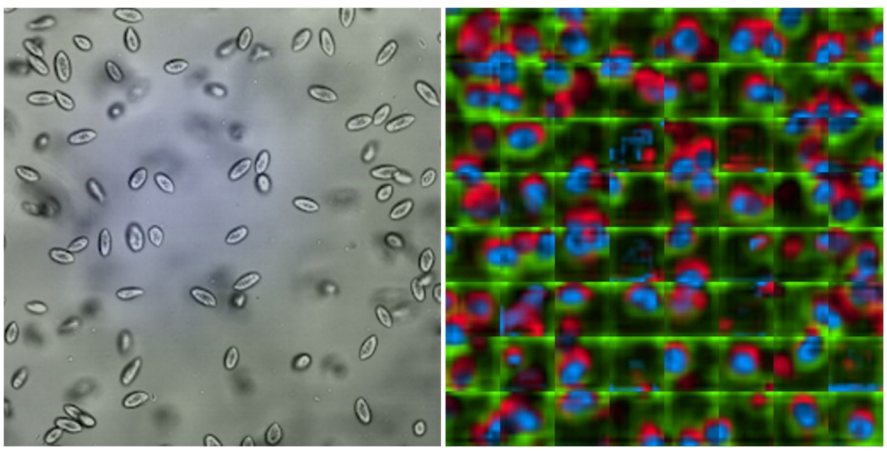

図 2．繰り返しパターンの検出

(4)　クラスタリング結果を教師データとした深層学習による細胞画像の分類

近年では、畳み込みニューラルネットワークを用いて、画像のクラスタリングを行う手法が開発されている。クラスタリングを用いることで、画像データセットを意味のあるいくつかの集合に分割することが可能であり、この集合をサブクラスや、スーパークラスとして学習に用いることで、画像分類の精度が改善される可能性があると考えた。本研究では、深層学習で画像分類を行

う際に、クラスタリング結果を教師データに追加することで、分類精度を向上させることを目的とした実験を行った。組織分類の実験結果から、クラスタリング結果を教師データとして学習に利用することで、深層学習による画像の分類精度が上がることを確認した。

５．主な発表論文等

〔雑誌論文〕　計2件（うち査読付論文　2件／うち国際共著　0件／うちオープンアクセス　2件）

１．著者名 Shiga Mikio、Seno Shigeto、Onizuka Makoto、Matsuda Hideo	４．巻 9
２．論文標題 SC-JNMF: single-cell clustering integrating multiple quantification methods based on joint non-negative matrix factorization	５．発行年 2021年
３．雑誌名 PeerJ	６．最初と最後の頁 e12087～e12087
掲載論文のDOI（デジタルオブジェクト識別子） 10.7717/peerj.12087	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 Fujimoto Kenji、Graduate School of Information Science and Technology, Osaka University, Suita, Osaka, Japan、Seno Shigeto、Shigeta Hironori、Mashita Tomohiro、Ishii Masaru、Matsuda Hideo	４．巻 10
２．論文標題 Tracking and Analysis of FUCCI-Labeled Cells Based on Particle Filters and Time-to-Event Analysis	５．発行年 2020年
３．雑誌名 International Journal of Bioscience, Biochemistry and Bioinformatics	６．最初と最後の頁 94～109
掲載論文のDOI（デジタルオブジェクト識別子） 10.17706/ijbbb.2020.10.2.94-109	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 －

〔学会発表〕　計7件（うち招待講演　0件／うち国際学会　0件）

１．発表者名 志賀幹夫，瀬尾茂人，鬼塚真，松田秀雄
２．発表標題 一細胞遺伝子発現解析のためのJoint-NMFを用いたクラスタリング手法
３．学会等名 第42回日本分子生物学会年会
４．発表年 2019年

１．発表者名 瀬尾茂人
２．発表標題 バイオインフォマティクスツールのエコシステム
３．学会等名 NGS EXPO 2022
４．発表年 2022年

１．発表者名
瀬尾茂人

２．発表標題
生命科学データの情報処理について

３．学会等名
化学工学会第52回秋季大会，バイオ部会

４．発表年
2021年

１．発表者名
瀬尾茂人

２．発表標題
シングルセル発現解析ツールの動向について

３．学会等名
Single-Cell 2021 Osaka セミナー

４．発表年
2021年

１．発表者名
Toshiya Tanaka, Shigeto Seno, Hideo Matsuda

２．発表標題
Feature selection with VAE for scRNA-seq analysis

３．学会等名
29th Conference on Intelligent Systems for Molecular Biology (ISMB)

４．発表年
2021年

１．発表者名
椙村渉，瀬尾茂人，深田宗一朗，松田秀雄

２．発表標題
時系列ラベルを用いた弱教師あり学習による筋組織再生過程の定量化手法の提案

３．学会等名
第137回数理モデル化と問題解決研究発表会

４．発表年
2022年

１．発表者名 新田恭晟，瀬尾茂人，細田一史，松田秀雄
２．発表標題 畳み込みニューラルネットワークを用いた繰り返しパターンの検出手法と顕微鏡画像への応用
３．学会等名 第137回数理モデル化と問題解決研究発表会
４．発表年 2022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
研究分担者	山下　英里華 (Yamashita Erika) (10880106)	大阪大学・医学系研究科・特任研究員（常勤） (14401)	
研究分担者	水野　紘樹 (Mizuno Hiroki) (90707655)	大阪大学・生命機能研究科・助教 (14401)	削除：2019年11月30日

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