公開

研 究 終 了 報 告 書

「テンソル分解を利用した細胞間相互作用の時空間解析」

研究期間: 2019 年 10 月～2023 年 3 月
研 究 者: 露崎 弘毅

1. 研究のねらい

本研究では、生命科学データをテンソル分解という数学的な枠組みで統一的に解析する。生命は幾つもの生体分子（例: RNA、DNA）や現象（例: SNP、CNV）が複雑に関連しあったネットワークで構成される。この大規模なネットワークに関わる生体分子や現象を全て同時に計測することは現在までのところできていない。そのかわりに、生命科学研究ではオミックスと称し、特定の生体分子や現象のみに限定した上で、それらを網羅的に計測するアプローチがとられている。そのような生命科学データは、1 つの行列で全てを記述できるようなシンプルな構造にはなっておらず、互いに繋がりがある複数のデータとなる。このような複雑に繋がりあったデータ集合を、どのように解析すべきなのかは自明ではない。

本研究では、生命科学データをテンソルという数学的な枠組みで扱い、テンソル分解というアルゴリズムを適用することで、どれだけデータ構造が複雑になっても、統一的にデータに含まれるパターンを抽出する。テンソルとはベクトルや行列のある種の一般形であり、テンソル分解とは、行列データで適用する主成分分析のような行列分解アルゴリズムのテンソル版に相当する。テンソル分解の応用例として、さきがけ[多細胞]の目標である、細胞間相互作用（Cell-Cell Interaction; CCI）の検出を行う。

2. 研究成果

（1）概要

さきがけ期間に取り組んだテーマとして、以下の 5 つを挙げる。

研究テーマ 1「線虫時系列神経活動データ解析」は、1 期生豊島氏との共同研究であり、複数個体分の線虫の神経活動データから、共通する CCI パターンをテンソル分解で抽出する WormTensor という解析手法を開発したものである。

研究テーマ 2 「scTensor の性能検証」は、1 細胞 RNA-Seq データに含まれる CCI をリガンド・受容体ペアの共発現を根拠として検出する scTensor に関して、論文化のために、競合手法との大規模なベンチマークを行なったものである。

研究テーマ 3「汎用的テンソルツールの開発」は、今後益々複雑化する生命科学データにおいて、複雑なデータ構造を統一的に解析するためのツール（R パッケージ）を開発したものである。

研究テーマ 4「寄生生物時系列バルク/単核 RNA-Seq データ解析」は、1 期生吉田氏との共同研究であり、寄生生物コシオガマと、その宿主シロイヌナズナとで行った（バルク/単核）RNA-Seq データを解析し、寄生に関する寄生・宿主側の CCI 関連遺伝子群の特定を目指したものである。

研究テーマ 5「ウニ時系列バルク/1 細胞 RNA-Seq データ解析」は、1 期生谷口氏との共同研究であり、バフンウニの（バルク/1 細胞）RNA-Seq データを解析し、バフンウニの発生・分化

に関連した CCI や、光と消化管形成の関係性の調査を目的としたものである。

(2)詳細

研究テーマ 1「線虫時系列神経活動データ解析」

さきがけの同領域のメンバーである豊島有氏が計測した、線虫の 4 次元トラッキングデータに含まれる CCI を、テンソル分解ベースで検出する WormTensor (https://cran.r-project.org/web/packages/WormTensor/index.html)という解析手法を開発した。当該データは、線虫の神経細胞の発火を Ca2+イメージングの蛍光強度としてとらえたものであり、3 次元空間座標と時系列で 4 次元のデータになっている。細胞型の同定には、豊島氏が独自に開発した、アノテーションツールを利用しており、ノイズ・トレンド除去のための正規化を何種類か適用している。塩刺激を加えており、その刺激に対して ON/OFF 応答する神経細胞が含まれている。

当該データは、個体ごとに品質がまちまちであり、一部細胞データが欠損していたり、誤った細胞ラベルがアノテーションされていたり、個体によっては、細胞の種類に関わらず、全ての細胞が異常な波形を示している場合がある。このようなノイジーなデータ集合から、より信頼性の高い CCI を検出するためには、複数個体の結果を統合して解析することが重要である。WormTensor はこのような目的で開発された解析手法であり、従来手法のコンセンサスクラスタリング(個々の個体ごとのクラスタリング結果の平均をとる手法)と比較して、個体ごとの重みを自動的に推定しながら、クラスタリング結果を統合するという強みがある。細胞ごとの機能アノテーション(線虫の体の動きや、塩への応答の有無などの既存知識)をもとに、両者の結果を比較したところ、特定の機能を持つ細胞同士を同じクラスタに集めることができていることがわかった。現在は、この解析結果を論文にまとめているところである。

研究テーマ 2 「scTensor の性能検証」

1 細胞 RNA-Seq データに含まれるリガンド・受容体遺伝子の共発現を根拠として、CCI を検出する既存手法として、CellPhoneDB などのツールが採用しているラベル並び替え検定がある。scTensor の性能を検証すべく、このラベル並び替え検定やそこから派生した手法と比較を行なった結果、CCI が複数の細胞型に関連した多 vs 多の時に、より他の手法よりも正解 CCI のみを検出できることがわかったため、現在その結果を論文にまとめているところである。

研究テーマ 3「汎用的テンソルツールの開発」

生命科学のヘテロなデータをシームレスに統合するための R パッケージとして、DelayedTensor (https://www.bioconductor.org/packages/release/bioc/html/DelayedTensor.html) と mwTensor (https://cran.r-project.org/web/packages/mwTensor)を開発した。DelayedTensor は複数のテンソルを 1 つの高階テンソルにまとめるテンソル縮約を実行する einsum 関数を実装しており、高階テンソルに対してテンソル分解を適用することにより、データに含まれるパターンを抽出できる"グローバルな"分解をサポートする。一方 mwTensor は複数のテンソルを 1 つにまとめず、個々に分解し、分解後の因子行列を他のテンソルと互いに共有する、"ローカルな"分解をサポートする。これらツールは、scTensor のモデル拡張や、領域内の共同研究案件での活用を

計画している。

研究テーマ 4「寄生生物時系列バルク/単核 RNA-Seq データ解析」
1 期生吉田氏と共同で、寄生生物コシオガマと、宿主としてシロイヌナズナの時系列 RNA-Seq データをバルクレベルと単核レベルとで計測した。バルク RNA-Seq のほうは、コシオガマとシロイヌナズナとで 1,3,7 日目のデータを計測してあり、シロイヌナズナ側での wol 遺伝子変異の有無（＊）や、別々に植えた 1,3,7 日目のデータなど、実験に関するラベルデータが豊富であり、かつコシオガマとシロイヌナズナ間で共有するオーソログ遺伝子が少ないという特殊なデータ構造をしていたことから、ラベルデータ経由で、2 生物種の RNA-Seq データを共通の空間に射影する guided-PLS という解析手法を開発した。guiled-PLS により、通常の DEGs 解析ではわからなかった、1 日目と 7 日目間では発現パターンが逆相関する遺伝子群が特定できた。

単核 RNA-Seq の方は、寄生が起きている 7 日目のデータのみを VASA-Seq で計測しており、既に吉田氏の方で Seurat による解析は済んでおり、発現量行列と細胞型ラベルを受け取っている。1 細胞データで検出された細胞型が、バルクデータにどのくらいの比率で混入しているのかを推定する問題を Cell-type Deconvolution といい、この解析を今回の単核 RNA-Seq に適用することにより、寄生に関する細胞型が時系列のどの段階で増加したのかを特定する予定である。また、コシオガマとシロイヌナズナ間で共有するオーソログ遺伝子が少ないという事情は、単核 RNA-Seq でも同様であるが、Gene Ontology 経由でこれら 2 生物種の RNA-Seq データを共通の空間に射影する方法論を開発中である。
＊宿主側で wol 遺伝子に変異が入ると寄生が起きなくなる

研究テーマ 5「ウニ時系列バルク/1 細胞 RNA-Seq データ解析」
1 期生谷口氏と共同で、バフンウニの時系列 RNA-Seq データをバルクレベルと 1 細胞レベルとで計測中である。バルク RNA-Seq のほうは既に計測済みであり、2 条件（2 細胞期の細胞の分離の有無）×5 ステージ（2,11,14,18,24h）で、2 細胞期に分離した細胞が各々ウニの成体になる過程で、遺伝子発現的にはどのような変化が起きているのかを調べている。1 細胞 RNA-Seq のほうは、2 条件（Delta-Notch シグナルの遮断＊＊の有無）×4 ステージ（1,2,3,4day）、1 条件につき 3000 細胞（2 個体分）程度計測する予定である。

これらバルクデータと 1 細胞データの統合にテンソル分解が役立つと考えている。Cell-type Deconvolution により、バルクデータで特定された 2 細胞期に分離した細胞が分離前の状態に戻るステージで、どの細胞型のどの遺伝子が変化したのか、変化したのは、細胞全体の数・サイズか、または特定の遺伝子だけが発現変動したのか、といった情報の検出が期待できる。

また、過去に谷口氏が計測したバフンウニの 1 細胞データも解析しているが、RNA Velocity 解析（＊＊＊）の向きが、実時間と逆行しており、明らかに誤った推定となっているため、現在トラブルシューティング中である。横軸をスプライシング後の RNA のカウント値（s）、縦軸をスプライシング前の RNA のカウント（u）値とした相図を見る限りでは、ツールが想定しているような、「細胞の分化に伴い u/s 比が増加して、そのうち定常状態となり、その後は u/s 比が減少して、u、s 共に 0 に近づく」という構図にはそもそもなっておらず、ランダムな散布図にしか見えないことから、今回のウニのデータのクオリティに問題があるか、ウニという生物種のデータで起きる本

質的な問題である可能性がある。理由が前者であった場合は、次に得られる 1 細胞データで同様の解析をして、結果を見比べることで判断ができると思われる。理由が後者であった場合は、従来の RNA Velocity 解析が、あらゆる生物種において利用できるほど汎用的なものでは無く、ウニに特化した RNA Velocity 解析の必要性があるということになるため、新しい方法論開発に繋がるため、現在注目しているところである。

＊＊オプシン発現細胞を増やすための操作

＊＊＊mRNA のスプライシング情報から、細胞の分化の向きを推定する解析法、次元圧縮の図に矢印を描くベクトル場を出力する

3. 今後の展開

本さきがけ領域において、行列・テンソル分解を様々なデータに適用することで、その有効性を示せたと考えている。このような各研究分野の諸問題の解決を引き続き続けていくことで、テンソル分解の利用者数を増やしていくつもりである。また、今のところ、領域内のデータ解析については、ほとんどは共同研究という形で、自分が手を動かして進めてきたが、本研究テーマの副産物である DelayedTensor や mwTensor は、オープンソースなソフトウェアとして、誰でも利用可能な形態で公開していることから、今後の利用者数の増加に貢献できるのではと考えている。これらは、複雑なデータ構造での解析に特化した汎用的なソフトウェアであることから、生命科学に限らず様々なデータサイエンスに応用可能であり、特に、お互いに関連しているものの、既存のツールの入力データにしづらい、雑多なデータ集合に関して、本ツールのみが適用可能な事例が出てくると思われる。

4. 自己評価

【研究目的の達成状況】scTensor や WormTensor に関しては、さきがけ期間中に論文としてまとめられると考えている。また scTensor のバックエンドで実行される非負値テンソル分解の実装は nnTensor という別のパッケージとして開発されているが、こちらも OSS 論文に投稿済みである。それ以外にも、DelayedTensor や mwTensor など、汎用的なテンソル分解ツールを開発しており、これらも順次 OSS 論文として投稿する予定でいる。

さきがけ採択時に設定していた研究テーマとしては、scTensor を時空間 scRNA-Seq データに拡張させることを計画していたが、後発の手法 Tensor-cell2cell（Erick Armingol et al., 2022）が、そのような解析を行ったことや、テンソルを高階化した後はこれまでと同様テンソル分解をするだけで、あまり困難な問題に挑戦したことにはならないことや、領域内でメンバー同士が共同研究をする機運が高まっていたため、scRNA-Seq データに限らず、領域内の様々な生命科学データに対してテンソル分解を適用する方針に変更した。

【研究の進め方（研究実施体制及び研究費執行状況）】コロナ禍のため、当初予定していた海外出張は行かずに、研究費のほとんどはテクニカルスタッフの人件費として活用した。1 期生豊島氏との共同研究の実際の作業は、テクニカルスタッフが行なっており、新規解析手法の開発や、論文化にまで繋がりそうであることから、有効に活用できたと考えている。また、1 期生谷口氏に依頼し、より多条件での scRNA-Seq データを計測中（年内には納品予定）であり、これまでに無い高次元データになることから、これも提案するテンソル分解のテストデータとして活用できると思われる。

【研究成果の科学技術及び社会・経済への波及効果】 本研究テーマの副産物であるDelayedTensorやmwTensorは、複雑なデータ構造での解析に特化した汎用的なソフトウェアであることから、生命科学に限らず様々なデータサイエンスに応用可能だと思われる。特に、お互いに関連しているものの、既存のツールの入力データにしづらい、雑多なデータ集合に関して、本ツールのみが適用可能な事例が出てくると思われるので、そのような分野を中心に展開していく予定である。

【領域独自の評価項目】 さきがけ[多細胞]領域独自の方針としては、高橋総括が何度も強調していたように、領域を一つの仮想的なラボと見立て、異分野のメンバー間でシナジー効果をもたらすということだと思われる。理論やプログラミングを中心として研究を進めていた自分に対しては、特にそれが求められており、当初から実験研究者との共同研究が期待されていた。そのため、以下で示すように、意識的に他のメンバーと交流を図り、データ解析の受託や、実験デザインやデータ解析の情報提供を行なってきており、他のメンバーとの連携は十分に行ったと自己評価している。

まず実際に生データを提供してもらい、一般的なデータ解析や、本研究テーマであるテンソル分解によるモデリングまで行った共同研究案件としては、以下のものである。

・1 期生豊島氏: 線虫神経活動データ → WormTensor 開発

・1 期生吉田氏: 寄生生物(バルク/1 細胞)RNA-Seq データ → DEGs 解析、共通次元埋め込み(バルク/1 細胞)

・1 期生谷口氏: ウニ(バルク/1 細胞)RNA-Seq データ → Seurat/RNA Velocity 解析、多次元テンソル分解(予定)

また、単に私が開発した scTensor の技術提供(データの用意の仕方他、scTensor の使い方に関する情報提供)を依頼された案件としては、以下のものがある。

・1 期生橋本氏: Osteoblast 内に含む細胞間相互作用検出

・1 期生木戸屋氏: 血球細胞間での細胞間相互作用検出

コロナ禍により、交流が少なかったものの、同期以外のメンバーとも今後共同研究が期待できる。例えば、2 期生村瀬氏の自家不和合性に関して、リガンド・受容体の配列データから、相互作用予測を行う解析や、2 期生秋山氏の、縞パターンに関係した遺伝子発現解析についても、テンソル分解によるモデリングができると考えており、Zoom ベースのやり取りで、具体的なデータ解析の計画を進めている。さきがけの期間終了後も、これら共同研究は並行して進めていくつもりである。

5. 主な研究成果リスト

(1)代表的な論文(原著論文)発表

研究期間累積件数:11件

1. WormTensor: a clustering method for time-series whole-brain activity data from C. elegans, Koki Tsuyuzaki, Kentaro Yamamoto, Yu Toyoshima, Hirofumi Sato, Manami Kanamori, Takayuki Teramoto, Takeshi Ishihara, Yuichi Iino, Itoshi Nikaido, BMC Bioinformatics, 2023 (Accepted)
   さきがけ同期の豊島有氏との共同研究成果。線虫時系列神経活動データから、細胞間

相互作用を検出する問題をテンソル分解の一種 MC-MI-HOOI で解いたもの。

2. Kei Ikeda, Taka-Aki Nakada, Takahiro Kageyama, Shigeru Tanaka, Naoki Yoshida, Tetsuo Ishikawa, Yuki Goshima, Natsuko Otaki, Shingo Iwami, Teppei Shimamura, Toshibumi Taniguchi, Hidetoshi Igari, Hideki Hanaoka, Koutaro Yokote, Koki Tsuyuzaki, Hiroshi Nakajima, Eiryo Kawakami. Detecting time-evolving phenotypic components of adverse reactions against BNT162b2 SARS-CoV-2 vaccine via non-negative tensor factorization. iScience. 2022, 25(10), 105237-105237
   本研究テーマにも関係する非負値テンソル分解を活用して、新型コロナワクチンの副反応データ（アンケート）を次元圧縮し、どのような副反応のパターンがあったのか調査した。
3. Koki Tsuyuzaki, Naoki Yoshida, Tetsuo Ishikawa, Yuki Goshima, and Eiryo Kawakami, Non-negative tensor factorization workflow for time-series biomedical data, STAR Protocol, 2023 (Accepted)
   2. の論文の解析ワークフローの解説。

（2）特許出願

特になし

（3）その他の成果（主要な学会発表、受賞、著作物、プレスリリース等）

1. BioC Asia 2021（https://biocasia2021.bioconductor.org）での大会長、運営、口頭発表、ワークショップ開催

BioC は生命科学で利用される R 言語のパッケージを集めたレポジトリである Bioconductor のカンファレンスであり、BioC Asia はそのアジア分科会に相当する。例年アジアの地域で開催されるものであり、2021 年度は日本に誘致することに成功した。大会長として会の運営に携わり、例年通りのオミックス解析に関する発表だけでなく、言語の障壁問題に配慮したオープニングトークや、ワークショップの開催を行ない、国際的なプレゼンスを発揮できたと思われる。

2. 露崎弘毅．行列・テンソル分解によるヘテロバイオデータ統合解析の数理．JSBi Bioinformatics Review

2-1. 露崎弘毅. 行列・テンソル分解によるヘテロバイオデータ統合解析の数理 —第 4 回 質的データ、距離、グラフ—. JSBi Bioinformatics Review. 2022, 3(2), 33-46

2-2. 露崎弘毅. 行列・テンソル分解によるヘテロバイオデータ統合解析の数理 —第3回 テンソル分解—. JSBi Bioinformatics Review. 2022, 3(1), 20-33

2-3. 露崎弘毅. 行列・テンソル分解によるヘテロバイオデータ統合解析の数理 —第 2 回 行列同時分解—. JSBi Bioinformatics Review. 2021, 2(1), 15-29

2-4. 露崎弘毅. 行列・テンソル分解によるヘテロバイオデータ統合解析の数理 —第 1 回 行列分解—. JSBi Bioinformatics Review. 2021, 1(2), 18-25

日本バイオインフォマティクス学会が取りまとめている査読付き論文である、JSBi Bioinformatics Review にて連載を行なっている。内容は、本研究テーマにも関係する行列・テンソル分解を、初学者に対して解説する総説論文。

3. バイオ DB とウェブツール ： ラボで使える最新 70 選 ： 知る・学ぶ・使う、バイオ DX 時代の

羅針盤

小野, 浩雅 (担当:分担執筆), 羊土社 2022

オープンソースで利用可能な Web ツールの紹介雑誌に、Code Ocean という論文のデータ解析の再現性に関わるサービスを紹介した。