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 ２９ 日現在

機関番号：１４３０１
研究種目：基盤研究(C)（一般）
研究期間：2017～2019
課題番号：１７Ｋ０７２５４

研究課題名（和文）多層オミクスデータ統合解析のための情報幾何的ベイズ推定法の開発

研究課題名（英文）Development of Bayesian Estimation Method based on Information Geometry for Multi-layered Omics Data Integration

研究代表者
山田　亮（Yamada, Ryo）

京都大学・医学研究科・教授

研究者番号：５０３０１１０６

交付決定額（研究期間全体）：（直接経費）　　3,800,000円

研究成果の概要（和文）：申請者の過去の研究フィールドを中心に、複数のデータセットをベイズ手法により統合解析することに適当な例を選び、シミュレーションデータ作成とMCMCベイズ法との両方のプログラム実装を進め、学会発表に至ったが、われわれの構想とほぼ同じ枠組みの研究が海外研究者により論文発表された。これを受け、オミックスデータの定義を拡張し、特に、統計解析の枠組みに乗りにくい表現型である、形態学的情報と３次元移動・軌跡情報とを標的とし、確率微分方程式を立て、遊走細胞の３次元形態情報とその３次元空間移動情報とを、モデルに定めたパラメタの値の事後分布として特徴づけするＭＣＭＣベイズ手法の開発に成功した。

研究成果の学術的意義や社会的意義
海外他研究者による先行発表により、当初計画を変更して取り組むことを余儀なくされたが、オミックスデータの定義を拡張し、特に、統計解析の枠組みに乗りにくい表現型である、形態学的情報と３次元移動・軌跡情報とを標的として、ＭＣＭＣベイズ手法の開発に成功した。このようにして抽出した１細胞の形・動きの特徴量は、いわゆる１細胞情報（とさらに統合するのが容易な状態になっている。その基本的手法の枠組みを維持しつつ、標的に軌道修正を加えることにより、かえって、オミックス研究領域における解析の難しい表現型の解析基盤を整えることに寄与することとなり、有意義なものとなった。

研究成果の概要（英文）：Initially we surveyed the appropriate targets of Bayesian integration of multiple omics layers and developed a method and did a poster presentation. Unfortunately a study in the similar frame was published by overseas competitors. Upon this, we re-directed our study targets of MCMC Bayesian approach to the phenotypes that are difficult to handle, 3-dimensional shape and 3-dimensional movement, and successfully developed a method to extract meaningful features from them so that those phenotypes can be readily integrated with single cell omics data set. The finding was proposed in a domestic meeting.

研究分野：オミックス統計解析

キーワード：オミックス　解析手法　ベイズ　MCMC　ゲノム

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）
１．研究開始当初の背景

本応募課題は、複数のデータセットを統合して判断に資する形での出力を得る方法論の研究である。はじめに、具体例を示す。

(1)ある遺伝子のジェノタイプと発病に関連がある。

(2)その遺伝子のジェノタイプと遺伝子発現量に関連がある。

(3)遺伝子発現は臓器・組織特異的である。

(4)遺伝子発現臓器・組織は疾患にて異常が起きる。

(5)疾患の状態と特異的臓器組織における遺伝子発現量とには関係がある。

(6)この発現分子の阻害剤に治療効果があり、それは臓器・組織での発現量を変化させ、また患者を治癒する。

これら(1)-(6)に合致するデータが得られたとき、『この遺伝子の機能発現には遺伝的多様性が影響し、それが疾患の成立リスクと関連し、かつ、疾患の状態においてその発現分子の機能発現に介入することに予後改善効果がある』という一連のストーリーが成立する。

この一連のストーリーが仮説であるときに、実現可能性を度外視して研究をデザインするなら、出生から死亡までをプロスペクティブに追いかけ、ジェノタイプとフェノタイプを記録し、適切な時刻と臓器組織における発現量を測定し、発病したなら、治療薬の無作為比較試験に組み込むことになるだろう(図左上半)。このようにして得られる多彩なデータをどのように解析すればよいのかについても未定見であるが、すべてのデータは同一の対象者に紐づいており統計モデルを立てることは不可能ではないだろう。

しかしながら、これは確かに現実的ではない。プロスペクティブなコホート研究と無作為比較試験とではデータを集める時間のスケールが異なるし、実施体制をどのように組むのかという点も、スタディが巨大すぎて想像すらしにくい。

実際に行われているのは、各項目のスタディを別々にデザインしてデータを収集し、それぞれに統計学的解析を加え(図右下半)、その解析結果を突合せて、ストーリー全体の信憑性について判断をするという作業である。しかしながら、この最後の解析結果の突合せの部分は手法化・定量化されていない。この部分の手法化・定量化を目指すのが本応募課題である。実際、データ公開が進んでいる現在、異なる研究グループからの異質なスタディデザインの結果を統合して理解したいという潜在的要請は高い。

一般に、複数のデータセット解析の結果を統合する手法はメタ解析と呼ばれるが、複数の民族で同形式の関連解析を行った結果を統合するというようなごく限定的な場合にのみ手法化が進んでいる。統合を困難にしている理由の一つは、個別のスタディデザインにおいて解析手法に工夫がなされ特化していることが挙げられる。その点へ対応するためには異なる解析手法の間に共通語を導入することが適切であるが、その共通語に相当するものとして情報エントロピー・情報幾何が知られている。また、この統合が目指すところの「現在、人間が行っている結果の突合せ作業」は、事前知識を仮定しており、また、突合せ結果には曖昧さを許容しているとみなしていると言う点で、ベイズ流による事後分布推定が適している。このような多様なデータセットに対応する解析的手法の構築は一般に困難であるが、計算機乱数を用いたモンテカルロ法による事後分布推定ならば、柔軟に対応できることが知られている。

ここまでに述べた背景は、オミクスデータ等の研究データの統合解釈に関するものであったが、本応募研究はより臨床現場に近い場面への応用も念頭に置いている。たとえば、遺

伝カウンセリングの場面では、有病率に関する疫学データと遺伝子バリアント関連解析結果とバリアントの機能予測解析結果とを突合せたい、というような需要も生じうる。このような場合にも事後分布を提供して判断に活かすことは可能であると考える。

２．研究の目的

各種オミクス研究データを用いた疾患研究では、研究デザインが多彩に工夫され、そのためのデータ解析手法の開発・整備が進んでいる。しかしながら、その各種解析手法は研究デザインに特化し、データセット横断的には統合しにくいのが現状である。現在のいわゆるデータ統合手法の主流は、同じオミクスに関して同様の研究デザインで実施された複数のデータセットの結果を統合するという単純なものであるのに対し、真に求められているのは、はるかに複雑なものである。例を挙げると、同一個人からの複数のオミクスデータセットと細胞株からの別のオミクスのデータセットとを両者のゲノムバリアントの共通性を使って統合したいと言ったものである。本研究では、異質なデータセットに共通する基礎として情報エントロピーと情報幾何を活用する。また、異質なデータセットをも柔軟にモデル化し事後分布推定を行うためのモンテカルロ事後分布推定法を導入する。これらを組み合わせることにより、研究者が複数のデータセット群から読み取りたいことを、実地医療の考え方とも相性のよい事後確率としての出力する、データ解析方法を開発する。

３．研究の方法

(0)　　統計学的な研究とともに行うべきこと

複数のデータセットからの解析結果を統合する課題というのは、ありとあらゆるところに存在する。たとえば、複数の実験を行ってそのリザルトを提示している学術論文では、複数のリザルトをディスカッションして結論を述べるのがふつうであるが、このディスカッションの部分は、本応募課題で言うところの、「統合」に相当する。また、学術総説では、あるテーマに沿って多数の論文を体系的にリストアップし、それらのすべてを見渡して所見を述べている。この見渡す作業も「統合」に相当する。臨床診断において、問診・身体診察・臨床検査を経て診断を下すという作業も、複数の情報ソースからのメッセージを「統合」している。ありとあらゆるところに存在するというのは、このような意味である。このように、統合されるべきデータセットの例は数多くあることが予想されるが、どのような統合のされ方に需要があるのかについては定見があるわけではない。したがって、研究の開始にあたっては、論文調査等を通じて、統合対象をリストアップする。また、統合の仕方にもバリエーションがあるはずであるが、「人による統合」がどのような観点においてなされているのかを調査し、「統合」に関する俯瞰図を得る。

(1) 理論的課題

(0)を進めるとともに、GWAS ケースコントロール解析と eQTL 解析とはゲノム疫学・オミクスデータの個別化医療応用という視点から必須項目と考えられるので、それらの情報幾何的表現の検討を平成２９年度初めより開始する。具体的には、最も単純な枠組みとしての、２ｘ２分割表、SNP ケースコントロール関連解析の２ｘ３分割表について検討する。また、SNP ジェノタイプの遺伝子発現量影響の解析について、同様に情報幾何的表現を検討する。初年度は、この２つについて情報幾何的表現を定め、それらの統合方法を検討する。

これにより、理論的目標の２つの項目の両方についてのプロトタイプが得られるものと予想する。実施にあたっては、理論的記述と並行して、そのデータシミュレーションとその情報幾何学的計算プログラムの実装を進めることで最終的なアウトプット(単純な枠組みでの統合プログラムの完成)に研究早期から目途をつけることとする。(図の共通言語化に相当)

(2)　　実用的課題

ベイズ流事後分布推定を柔軟に行う方法としてモンテカルロ法があり、その実用アプリケーションとして BUGS 法と STAN 法を採用する。初年度には、(1)の進行とは切り離して、ごく単純な事後分布推定課題を設定し、相互に影響し合う複数のデータセットからリサーチクエスチョンを体現する生成量の事後分布生成ルーチンの研究室内構築を達成する。これにより、第２年度以降には、研究課題の学術的内容に精通していないアルバイト等にもプログラム作成支援がしやすい状況となることが見込まれ、研究全体の効率化を図る。(図のデータ統合・柔軟なモデルで事後分布に相当)

４．研究成果

申請者の過去の研究フィールドを中心に、複数のデータセットをベイズ手法により統合解析することに適当な例を選び、シミュレーションデータ作成と MCMC ベイズ法との両方のプログラム実装を進めるにあたり、いくつかの技術的課題が確認されたので、それに関する情報収集を進め、適宜、改良を行い、学会発表に至ったが、われわれの構想とほぼ同じ枠組みの研究が海外研究者により論文発表された。これを受け、オミックスデータの定義を拡張子、特に、統計解析の枠組みに乗りにくい表現型である、形態学的情報と３次元移動・軌跡情報とを標的として、確率微分方程式を立て、それに対して、生物学的事前知識をある程度用いた事前分布を定めた上で、離散時刻観測データにより事後確率分布を更新する手法の開発に取り組んだ。これにより、遊走細胞の３次元形態情報とその３次元空間移動情報とを、モデルに定めたパラメタの値の事後分布として特徴づけするＭＣＭＣベイズ手法の開発に成功した。この成果は、学会発表した。また、このようにして抽出した１細胞の形・動きの特徴量は、いわゆる１細胞情報（発現情報等）とさらに統合するのが容易な状態になっている。本研究では、当初計画が海外他研究グループに先を越されたがゆえに、軌道修正することが適当となったが、その基本的手法の枠組みを維持しつつ、標的に軌道修正を加えることにより、かえって、オミックス研究領域における解析の難しい表現型の解析基盤を整えることに寄与することとなり、有意義なものとなった。

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<br>Okada, D., Nakamura,N., Wada,T., Iwasaki,A.,Yamada,R | ４．巻<br>34(5) |
| ２．論文標題<br>Extension of Sinkhorn Method: Optimal Movement Estimation of Agents Moving at Constant Velocity. | ５．発行年<br>2019年 |
| ３．雑誌名<br>Transactions of the Japanese Society for Artificial Intelligence | ６．最初と最後の頁<br>D-J13_1-7 |
| 掲載論文のDOI（デジタルオブジェクト識別子）<br>https://doi.org/10.1527/tjsai.D-J13 | 査読の有無<br>有 |
| オープンアクセス<br>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<br>－ |

| | |
|---|---|
| １．著者名<br>Basak, T., Nagashima, K., Kajimoto, S., Kawaguchi, T., Tabara, Y., Matsuda, F., Yamada, R. | ４．巻<br>12 |
| ２．論文標題<br>A Geometry-Based Multiple Testing Correction for Contingency Tables by Truncated Normal Distribution | ５．発行年<br>2020年 |
| ３．雑誌名<br>Statisitcs in Biosciences | ６．最初と最後の頁<br>63-77 |
| 掲載論文のDOI（デジタルオブジェクト識別子）<br>https://doi.org/10.1007/s12561-020-09271-6 | 査読の有無<br>有 |
| オープンアクセス<br>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<br>－ |

〔学会発表〕　計3件（うち招待講演　0件／うち国際学会　1件）

| |
|---|
| １．発表者名<br>Yusri Dwi Heryanto, Ryo Yamada |
| ２．発表標題<br>Cell Shape and Movement Analysis in Frenet-Serret Moving Frame |
| ３．学会等名<br>日本分子生物学会 |
| ４．発表年<br>2019年 |

| |
|---|
| １．発表者名<br>Yusri Dwi Heryanto, Ryo Yamada |
| ２．発表標題<br>Omics Data Integration Using Bayesian Non-Negative Matrix Factorization |
| ３．学会等名<br>The 63rd Annual Meeting of the Japan Society of Human Genetics（国際学会） |
| ４．発表年<br>2018年 |

| |
|---|
| １．発表者名<br>Satoshi Koyama, Ryo Yamada |
| ２．発表標題<br>Bayesian integration of multilayered omics data |
| ３．学会等名<br>日本人類遺伝学会 |
| ４．発表年<br>2017年 |

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

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