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　 ５ 月　 ２ 日現在

機関番号：１２６０１
研究種目：基盤研究(C)（一般）
研究期間：2020～2023
課題番号：２０Ｋ１２０５９

研究課題名（和文）高度な生命モデリングの基盤技術となる確率偏微分方程式のパラメータ推定論の確立

研究課題名（英文）Establishing parameter estimation theory of stochastic differential equations for advanced modeling of life systems

研究代表者

木立　尚孝（Kiryu, Hisanori）

東京大学・大学院新領域創成科学研究科・准教授

研究者番号：８０４１５７７８

交付決定額（研究期間全体）：（直接経費）　　3,300,000円

研究成果の概要（和文）：DNAシーケンシング技術やカメラ性能の向上により生物過程の時空間情報が急増している。これにより遺伝子間相互作用の時間的因果関係や、細胞・組織の３次 元的配置が生物の振る舞いへ与える効果などを厳密に調べることが可能になってきた。そこで本研究では、生命過程のより高度なモデリングを可能にするための 道具として、非線形確率偏微分方程式のパラメータをデータから推定する汎用的な機械学習技術の開発・実装を行うことを目標とした。計算機代数、自動微分、拡散モデルなど様々な方法を試したが一般的に使える道具となるためにはまだ多くの課題が残っており、今後も研究を継続していく必要がある。

研究成果の学術的意義や社会的意義
近年、生物に関する細胞レベルの非常に高い解像度の時空間的情報が爆発的に蓄積しており、これらを有効に活用して、未知の生物機構を発見したり、病気の治療に役立つ情報を抽出する情報科学的手法の開発の重要性が増している。本研究では確率偏微分方程式のパラメータ推定論に取り組んだが、これに成功すれば、遺伝子発現情報などから未来の生物の変化を予測するための強力なツールとなることが期待できる。

研究成果の概要（英文）：Spatio-temporal information on biological processes is rapidly increasing due to DNA sequencing technology and camera performance improvements. This has made it possible to rigorously study the temporal causality of gene-gene interactions and the effects of the three-dimensional arrangement of cells and tissues on the behavior of organisms. Therefore, this study aimed to develop and implement a general-purpose machine learning technique to estimate parameters of nonlinear stochastic partial differential equations from data as a tool to enable more sophisticated modeling of life processes. We tried various methods, such as computer algebra, automatic differentiation, and diffusion models. However, many issues remain to be solved to make it a generally usable tool, and further research should be continued.

Translated with www.DeepL.com/Translator (free version)

研究分野：バイオインフォマティクス

キーワード：一細胞シーケンシング　機械学習　確率モデル　トランスクリプトーム　確率偏微分方程式

１．研究開始当初の背景

DNA シーケンシング技術やカメラ性能の向上により生物過程の時空間的情報が急増している。例えば、一細胞 RNA シーケンシング (scRNA-seq) や、蛍光 in situ ハイブリダイゼーション法 (FISH) により、遺伝子発現プロファイルの時系列変化や組織内空間分布の一細胞解像度かつ遺伝子網羅的な計測が可能になった。また、３次元レーザースキャン技術により胚発生時の細胞移動や植物生長の形態変化を時系列計測した網羅的データも増えている。これらの技術革新により、今後は高解像度・網羅的な時空間データから、遺伝子発現の時間的因果関係や細胞の３次元空間配置が組織の機能へ与える効果など、生物の振る舞いに変化を引き起こすメカニカルな仕組みを推定する情報技術が一層重要になると考えられる。

生命情報学は、ゲノム情報など生命データの爆発的増大により生命科学における重要性が増し、大量・高次元データを解析する情報技術を発展させてきた。近年は、深層学習やガウス過程などの強力な機械学習技術を用いて、データに潜む非線形で複雑な数理構造のモデリングが可能となり成果をあげている。一方で、これらのブラックボックス型の機械学習技術では、「研究対象について多数の学習データがなければ高精度なモデルを構築するのが難しい」、「精度の高いモデルができたとしても、それが既知の物理・化学的制約を満たし生物学的に意味のある現象を捉えているのか、あるいは生物とは関係のない実験環境や測定データのバイアスを捉えたモデルになっているのかを区別することが難しい」、「パラメータの生物学的意味が薄く、学習後のモデルから生物学的議論が展開しづらい」といった問題が現れることもよくあった。

このため、近年の機械学習の進展を有効活用しつつ、既知の自然法則や生物知識を柔軟に取り込める数理モデリング技術が必要ではないか？また、学習されたモデルの生物学的解釈が豊かで、そこから生物メカニズムの深い考察や新発見に繋げられる技術を開発できないか？と考えた。

２．研究の目的

そこで本研究では、非線形確率偏微分方程式のパラメータをデータから推定する汎用的な機械学習技術の開発・実装を目指した。我々が対象とするモデルは図１のようなものである。ここで確率場φは隠れた生物状態を表し時空間に依存するベクトル場、 y は各時空点で観測された観測データ、 ξは、各時空点で独立に揺らぐノイズ、F と G は、場φとその空間偏微分、パラメータθに依存する任意の微分可能な非線形関数と確率分布である。我々は観測データのセットから、について周辺化したパラメータθの事後確率分布の最大化によりθを推定するソフトウェアを開発する（目的１）。また推定パラメータの信頼度を出力する機能を実装する（目的２）。さらに開発した手法を用いて細胞分化の scRNA-seq データを解析し、細胞分化の起点となる遺伝子発現変化の推定や、細胞状態変化を引き起こすポテンシャル関数（Waddington’s epigenetic landscape）の構築を行う（目的３）。

非線形確率偏微分方程式によるモデリングは、深層学習やガウス過程などブラックボックス型の数理モデリング技法に比べ、以下のような利点がある。

i) 多くの物理・化学法則は、微分方程式により記述される。また数理・理論生物学の分野では、分子反応から生態系まであらゆる生命過程を微分方程式を用いて長年記述してきた。この微分方程式の表現力を生かして既知の生物知識に適合するモデルを立てることで、生物とは無関係のデータのバイアスを学習してしまうリスクを減らせる。また学習の対象を関心ある生物パラメータに絞ることで、学習に必要な測定データ量を減らすことができる。

$$\frac{\partial \phi}{\partial t} = F\left(\phi;\theta\right) + \xi$$

$$y_{t,x}^{(i)} \sim G\left(\phi_{t,x};\theta\right)$$

図１：非線形確率偏微分方程式モデル。φ：隠れた生物状態、t：時間、x：空間、ξ：システムノイズ、y：観測データベクトル、θ：モデルパラメータ、F：生物システムモデル。G：観測データの出力確率分布。

ii) 「偏」微分方程式は近年増加している時空間的な広がりのあるデータのモデリングに適している。数値計算上では、偏微分方程式は多自由度の常差分方程式として扱われるので、一旦偏微分方程式の推定法が確立すれば、常微分方程式や差分方程式にも適用できる。

iii) mRNA の転写、生物の進化、動物行動など、生命過程にはランダム現象が見られることが多い。決定論的な方程式で生命過程を表現し、ランダム現象を全て観測ノイズや初期状態に押し込めるモデリングは無理が生じることも多い。「確率」微分方程式を用いれば生命過程のゆらぎを自然にモデル化でき、ゆらぎと制御の競争関係を調べることもできる。

iv) ミカエリス・メンテン式、ロトカ・ボルテラ方程式など、生命過程で基本的な方程式は非線形方程式でないと表現できない事が多い。このためユーザが任意に設定した「非線形」モデルを扱える汎用性をもつことがモデルの表現力のために重要である。

2. 本手法は地球科学分野などで用いられる「データ同化」

に似ている。しかしデータ同化の通常の応用に比べ、生物データの特徴に合わせて最適化されたアルゴリズムであるところに独自性がある。ここで想定する特徴としては、不均質な観測時刻点、空間的に抜け値の多いデータ、細胞・組織・種分化などの枝分かれのある時間発展、破壊計測データ、遺伝子数レベルの超高次元データ空間、などが挙げられる。既存のデータ同化の実装はこのような性質を持つデータにうまく対応できないため、スクラッチからアルゴリズムの設計・実装を行う。
3. 本研究の手法では、尤度の漸近理論を用いて推定パラメータの信頼度をフィッシャー情報行列から計算する。データから確実に決まるパラメータのみについて生物学的な議論を行うことで、学習されたモデルに対する誤った生物学的解釈を避けることができる。
4. 本手法を細胞分化の scRNA-seq データに適用する。細胞分化に分岐がある場合の細胞状態ダイナミクスでは、微分方程式の分岐理論を用いたモデリングなどがされてきた。しかし、分岐の原因となる非線形なポテンシャル関数のパラメータを推定するのは困難だった。本研究の手法を用いれば、信頼性評価つきでポテンシャル関数の形状を推定できるようになる。

３．研究の方法
1．（パラメータ推定アルゴリズムの実装）本手法では、勾配法を用いた事後確率最大化によりパラメータ推定を行う。この方法はグリッド・サーチやモンテカルロ・サンプリングに比べ、パラメータ数が多いときに計算効率がよい利点がある。先述したように、我々は、図１のようなシステムを対象とする。ここで、システムノイズは観測できないので、周辺化を行うと事後分布はは場φの全ての可能な時空間分布についての積分（ファインマン汎関数積分）となる。一般にはこの積分は実行できないので、場を中心座標とそこからの確率的揺らぎに分解し、エネルギー関数を揺らぎの２次項までで打ち切るラプラス近似を用いる。これにより、中心座標の部分で非線形モデル最適化、ゆらぎの部分で確率的推論を取り込めるモデルとなる。この事後分布に拡張ラグランジアン法を適用しパラメータを最適化する。
2．（パラメータ推定値の信頼性出力アルゴリズムの実装）パラメータの推定値の不確定性を定量評価するために最尤推定量の漸近理論を用いる。これによると推定パラメータの分散はフィッシャー情報行列の逆行列から得られる。我々は、二次アジョイント法を用いて、少ないメモリ量で経験的フィッシャー情報行列の逆を計算する手法（Ito,Phys.Rev.E,2016）を実装する。これによりパラメータの事後分布の広がりの程度についての情報が得られる。
3．（精度評価と効率的実装の探索）我々の手法を生物学で使われるさまざまな微分方程式に適用する。具体的にはミカエリス・メンテンの化学反応式、ロトカ・ボルテラの競争方程式、反応拡散方程式、等を考えている。まずシミュレーションデータを用いて、学習効率が方程式の種類やデータサイズにどのように依存しているかを調べる。その後これらの方程式でモデリングされた既存実験データに本手法を適用し、推定精度やモデルが実験データにどの程度よくフィットしているか、などを比較評価する。我々の手法は事後分布の勾配の計算の度に、偏微分方程式の数値積分と、確率的前向き・後ろ向き計算を行う必要があり、計算量は非常に大きい。このため効率的な実装とともにポスト京計算機（富岳）の活用も検討する。
4．（細胞分化過程のデータ解析）scRNA-seq データからは様々な分化進行度の細胞の遺伝子発現量が得られる。また昨年、遺伝子発現の変化速度（RNA velocity)を推定する解析手法が現れ、より精度の高いダイナミカルモデルの構築が可能になってきた。我々は、発現量と発現変化速度の両方に最もよくフィットするモデル（図１の関数 F,G）をデータ駆動的に決定し、細胞分化の起点となる遺伝子発現変化や、分岐を引き起こすポテンシャル関数の推定を行う。

４．研究成果
令和２年度には３．研究の方法の項目１．に対応するプログラムの作成を行った。その過程で場の任意関数の偏微分を計算する計算代数的アルゴリズムを開発した。C++言語のテンプレートメタプログラミング機能を活用して、場の任意回の微分係数を計算する関数を生成することに成功した。このソフトウェアでラグランジアンの任意回の微分係数が求まるようになったが、計算効率があまり高くない結果がえられたため、微分計算については JAX の自動微分の方法を採用することにした。
次に生成 AI の一種である拡散過程を用いた確率モデルの推定法の開発を行った。変分オートエンコーダーの枠組みを用いて変分下限値を最大化 することによりライト・フィッシャーモデルのパラメータの最適化を行うソフトウェアを作成した。次に真のライト・フィッシャーモデルを用いて観測データを シミュレーションにより生成し、本ソフトウェアによりパラメータを推定することを試みた。その結果、推定されたパラメータ値は、多くの場合データを生成したパラメータ値とはかなり異なったものになることがわかった。この理由は、多層のニューラルネットワークを用いても変分関数は真のモデルの条件付き確率を表現するには十分な柔軟性を持たないため、真のモデルとの変分関数とのずれが、パラメータ推定値のずれにつながると考えられた。このため変分オートエンコーダー ではなく、近似的にでも真の尤度を直接計算するアルゴリズムを開発することが今後の課題だと考えて現在研究をさらに進めている。

５．主な発表論文等

〔雑誌論文〕　計8件（うち査読付論文　6件／うち国際共著　0件／うちオープンアクセス　0件）

１．著者名 Zhang Yanting、Kiryu Hisanori	４．巻 23
２．論文標題 MODEC: an unsupervised clustering method integrating omics data for identifying cancer subtypes	５．発行年 2022年
３．雑誌名 Briefings in Bioinformatics	６．最初と最後の頁 -
掲載論文のDOI（デジタルオブジェクト識別子） 10.1093/bib/bbac372	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 Ishikawa Masato、Sugino Seiichi、Masuda Yoshie、Tarumoto Yusuke、Seto Yusuke、Taniyama Nobuko、Wagai Fumi、Yamauchi Yuhei、Kojima Yasuhiro、Kiryu Hisanori、Yusa Kosuke、Eiraku Mototsugu、Mochizuki Atsushi	４．巻 6
２．論文標題 RENGE infers gene regulatory networks using time-series single-cell RNA-seq data with CRISPR perturbations	５．発行年 2023年
３．雑誌名 Communications Biology	６．最初と最後の頁 -
掲載論文のDOI（デジタルオブジェクト識別子） 10.1038/s42003-023-05594-4	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 Uehata Takuya、et al	４．巻 143
２．論文標題 Regulation of lymphoid-myeloid lineage bias through regnase-1/3-mediated control of Nfkbiz	５．発行年 2024年
３．雑誌名 Blood	６．最初と最後の頁 243～257
掲載論文のDOI（デジタルオブジェクト識別子） 10.1182/blood.2023020903	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 Genuth Miriam A.、Kojima Yasuhiro、Julich Dorthe、Kiryu Hisanori、Holley Scott A.	４．巻 9
２．論文標題 Automated time-lapse data segmentation reveals in vivo cell state dynamics	５．発行年 2023年
３．雑誌名 Science Advances	６．最初と最後の頁 -
掲載論文のDOI（デジタルオブジェクト識別子） 10.1126/sciadv.adf1814	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 Zhang Yanting、Kiryu Hisanori	4．巻 13
2．論文標題 Identification of oxidative stress-related genes differentially expressed in Alzheimer’s disease and construction of a hub gene-based diagnostic model	5．発行年 2023年
3．雑誌名 Scientific Reports	6．最初と最後の頁 -
掲載論文のDOI（デジタルオブジェクト識別子） 10.1038/s41598-023-34021-1	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 Kawaguchi Risa Karakida、Kiryu Hisanori	4．巻 -
2．論文標題 RNA Secondary Structure Alteration Caused by Single Nucleotide Variants	5．発行年 2023年
3．雑誌名 Methods in Molecular Biology	6．最初と最後の頁 107～120
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/978-1-0716-2768-6_7	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 －

1．著者名 Kawaguchi Risa Karakida、Kiryu Hisanori	4．巻 -
2．論文標題 Genome-Wide RNA Secondary Structure Prediction	5．発行年 2023年
3．雑誌名 Methods in Molecular Biology	6．最初と最後の頁 35～48
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/978-1-0716-2768-6_3	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 －

1．著者名 Kuwabara Yasuhide、et al	4．巻 3
2．論文標題 Lionheart LincRNA alleviates cardiac systolic dysfunction under pressure overload	5．発行年 2020年
3．雑誌名 Communications Biology	6．最初と最後の頁 -
掲載論文のDOI（デジタルオブジェクト識別子） 10.1038/s42003-020-01164-0	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 －

〔学会発表〕　計0件

〔図書〕　計1件

１．著者名 浜田 道昭、木立 尚孝	４．発行年 2022年
２．出版社 コロナ社	５．総ページ数 268
３．書名 生物統計	

〔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

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