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　３ 年　 ５ 月 １３ 日現在

機関番号：１２６０１
研究種目：基盤研究(C)（一般）
研究期間：2017～2020
課題番号：１７Ｋ００３９８

研究課題名（和文）制御工学に基づく、生命システム推定法と生命制御論の確立

研究課題名（英文）Establishing statistical inference theory for bio-systems and biological control theory using control engineering

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）　　3,600,000円

研究成果の概要（和文）：次世代シーケンシング実験の低コスト化や、顕微鏡の高性能化により、生命状態の経時的変化を細胞レベルで測定する研究が増加している。一般に時系列データは、一時刻点のみの測定データに比べ、要素間の因果関係などを高精度に推定できると期待される。しかし、現時点では、これらのデータの解析には、クラスタリング法など記述的分析法が主に使われており、測定データから、生命状態変化を引き起こすメカニカルな仕組みを推定する研究は多くない。そこで我々は、制御工学の分野で広く使われているカルマンフィルターの理論を生命データに適用するために必要な新規アルゴリズム群を整備し実装した。

研究成果の学術的意義や社会的意義
生命情報科学の分野では人工知能や機械学習といった最新のデータ科学技術を用いたデータ解析が数多く行われているが、これらの技術が既存の確立した物理・化学・生物学の知識と無矛盾な結果を出す保証はなく、自然現象とは関係ないデータの特徴を捉えているのではないかという懸念が常に残る。そこで我々は、制御工学の分野で用いられているカルマンフィルターの理論を活用して、微分方程式のパラメータを測定データから推定する手法を開発した。この手法を用いれば、既知の生命過程の知識を人工知能や機械学習のモデルと統合することが容易になるため、理論生物学の強力な道具立てになることが期待される。

研究成果の概要（英文）：Due to the low cost of next generation sequencing experiments and the high performance of microscopes, there has been an increase in research on measuring changes in the state of life over time at the cellular level. In general, time-series data are expected to provide more accurate estimates of causal relationships among elements than data measured only at a single time point. However, at present, descriptive analysis methods such as clustering are mainly used to analyze these data, and there is not much research on estimating the mechanisms that cause life state changes from measurement data. Therefore, we have developed and implemented a new set of algorithms to apply the theory of Kalman filter, which is widely used in the field of control engineering, to biological data.

研究分野：生命過程の数理モデリング

キーワード：生命情報学　カルマンフィルター　微分方程式　機械学習　１細胞シーケンシング

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

次世代シーケンシング実験の低コスト化や、顕微鏡の高性能化により、生命状態の経時的変化を細胞レベルで測定する研究が増加していた。一般に時系列データは、一時刻点のみの測定データに比べ、要素間の因果関係などを高精度に推定できると期待される。しかし、これらのデータの解析には、主成分分析やクラスタリングなど、データをサマライズする記述的分析法が主に使われており、測定データから、生命状態変化を引き起こすメカニカルな仕組みを推定する研究は多くなかった。

一方、制御工学の分野では、ノイズを含む時系列データのモデリングを長らく行ってきた。例えばカルマンフィルター(KF)理論は 1960 年前後に発表され深い数学理論が構築されている。KF はアポロ計画でロケット姿勢制御に使われたほか、飛行機の自動操縦やカーナビ・システムなど、機器状態の認識と制御に現在も広く使われている。従って KF 理論を生命時系列データに応用できれば、生命の動的性質解明のための強力な手法となり、更には生命を細胞レベルで自在に制御する先進工学技術の基礎になるだろうと考えられる。

しかしながら、膨大に存在する KF の既存アルゴリズムやソフトウェアは、工業製品への応用に特化されており、ただちに生命データに適用することが難しい。例えば、制御工学の制御対象は通常人工物のため、その基本的仕組みは明らかなのが前提だが、生物学の場合、細胞内部の仕組みがよく分かっていない。また制御工学では、固定時間間隔で、対象に影響を与えない非侵襲計測を長時間行うことを前提としている。しかし、生物実験では、計測の時間間隔が不均一で、僅か数時刻点しか測らないこともある。また測定により対象が破壊される侵襲計測のことも多い。その他、通常の制御工学の対象と異なり、生命データでは想定する状態変数の次元数が、全遺伝子数など、超高次元になることも多く、推定パラメータのオーバー・フィッティング問題はより深刻である。

ただしこれらは KF 理論の本質的な困難ではなく、理論を実データに適用する際の実装上の難点にすぎない。このためアルゴリズムとソフトウェアを生命データに適するよう刷新すれば、KF 理論を生命データへ適用することが可能であると考えられた。

２．研究の目的

そこで本研究では、KF 理論を生命データに適用するため、新規アルゴリズム群を整備し実装することを目標とした。これを一細胞 RNA-seq データや転写因子の過剰発現データに用い、遺伝子間相互作用の定量的推定や、制御変数を活用した生命状態の制御理論の有用性を示すことを目指した。

本申請では以下の技術開発を目的とした。

(1) 生命データに最適化された、システム同定アルゴリズム及びソフトウェアの開発:　観測データから、入力を出力に変換するシステムを推定することを、制御工学では「システム同定」と呼ぶ。生物の仕組みはよく分かっていないため、システム同定技術は制御工学の通常応用に比べ重要である。我々は隠れマルコフモデルにおけるバウム・ウェルチ法に類似の期待値最大化(EM)アルゴリズムを開発・実装する。その際、不均質な測定時間間隔、侵襲計測、状態ベクトルの超高次元性など、生命データ特有の事情を考慮してアルゴリズム設計を行う。

(2) KF 理論の制御変数を用いた生命制御理論構築: KF 理論では、制御変数と呼ばれる外部入力用のパラメータがあり、これを用いて状態変数に任意の軌道を描かせたり、外的影響要因をモデルに取り込んだりできる。本研究では、生命データにおいて人工的・外在的要素を制御変数に取り込むことの有用性を実データで検証する。また、「評価関数」となる状態変数の関数が一つ与えられた時、その期待値を最大化するような制御変数値を求める「最適制御理論」が知られており、この理論を生命データ用に一から書き換え整備する。

(3) 連続時間非線形 KF のアルゴリズム開発:　本研究では、任意実数値の計測時間間隔や、測定時刻をパラメータ推定する場合に対応するため、連続時間線形 KF に関して新規アルゴリズム群を開発する。この場合、推定されるシステムは、線形確率常微分方程式である。更に、X の非線形関数の場合もアルゴリズム開発をめざすが、この場合推定されるのは、非線形確率常微分方程式になる。これは、システムノイズの分散が小さい時、システム生物学における微分方程式の係数推定問題と同一の問題設定となる。

これらの具体的な技術的課題を解決することには以下のような研究意義があると考えられる。

(1) 時系列データから背後に潜む動的システムを推論する基礎理論が整備され、例えば一細胞 RNA-seq や、細胞レベルの動画像データから、遺伝子相互作用ネットワークや、細胞内小器官の力学パラメータなどを推定できるようになると予想される。

(2) KF 理論が含む制御変数を、環境刺激や遺伝子発現の人工改変と関連付けて、生命活動への外的摂動をモデリングできるようになる。更に最適制御理論を応用して、望ましい生命状態

を実現するように制御変数を数値最適化できる。これにより、細胞分化誘導の効率化、農作物の収量最大化、患者ごとの抗がん剤の投与量最適化などへの応用が期待できる。

３．研究の方法

本研究においては、C++言語と行列計算ライブラリ Eigen を用いてスクラッチからアルゴリズムの開発と実装を行った。

実装すべき確率微分方程式モデルのモデル構造は図１のようなものである。ここで t を時間パラメータとして Y(t)は計測された時系列的な生命データを表し、X(t)は測定の背後に存在する生命状態の時系列的変化を表している。u(t)は生命システムに加えられる人為的な摂動の効果を表している。生命状態変数 X(t)の時間変化は微小時間Δt だけ過去の生命状態 X(t-Δt)と摂動 u(t)により定められ、そこに正規分布に従うシステムノイズ v(t)が各時刻で付加されるものとする。一方測定データ変数 Y(t)は、生命状態 X(t)と摂動 u(t)の他に正規分布にしたがう測定ノイズ w(t)が加わって得られるものとする。

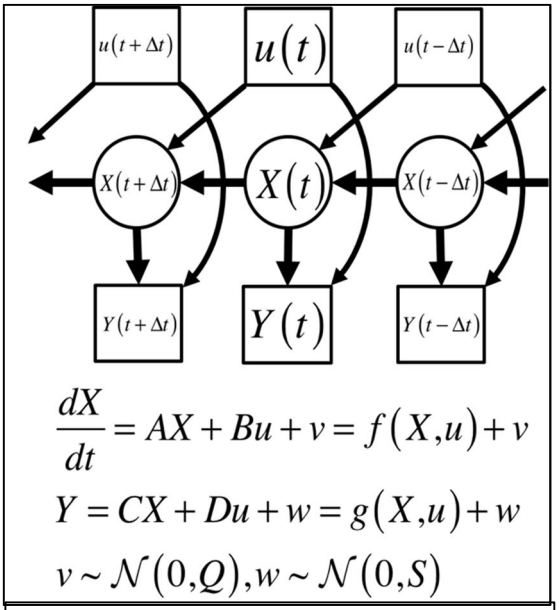

図１． KF モデル。X:隠れた生命状態。Y:測定データ。u:制御変数。v：システムノイズ。w:観測ノイズ。t:時間。*N*:正規分布。

４．研究成果

(1) 確率微分方程式が線形の場合にはアルゴリズムを完成させて実装することができた。実装に際してのデータ構造に関しては、一時刻点において X(t)が従う正規分布の平均値と分散共分散行列および正規化因子を表すオブジェクトと隣接する二時刻点において状態変数のペア(X(t),X(t-Δt))が従う正規分布の平均値、分散共分散行列、正規化因子を保持するオブジェクトが必要となる。その他に状態 X(t)から測定データ Y(t)を出力する正規分布の平均値、分散共分散行列のパラメータも保持する必要がある。この確率モデルの尤度は、初期状態 X(0)の正規分布から出発して前時刻の状態 X(t-Δt)の周辺化を繰り返し行う前向きアルゴリズムにより計算することができる。一方尤度のパラメータについての勾配を計算する際には、最終時刻の状態 X(T)に関する確率分布から出発して後ろ向きに状態変数の周辺化を繰り返し行う後ろ向きアルゴリズムを実装する必要があった。一細胞シーケンシングデータなど実際の生命データにおいては、枝分かれのある木構造状の時間発展をするシステムが多くあるため、木構造での前向き・後ろ向きアルゴリズムに相当する内向き・外向きアルゴリズムを実装した。また、生命データでは、各時刻で全ての測定値が得られるとは限らず、多くの抜け値があることも多いため、測定データ変数 Y(t)の成分で未測定のものについては、出力確率の正規分布を周辺化した上で、内向き外向きアルゴリズムを行うようにアルゴリズムを改良した。尤度の勾配を計算した後は、勾配法を用いてパラメータの最適化を行うが、最適化パラメータがどの程度の信頼性を持って特定の値に定まったのかを定量化するために、二次のアジョイント法を用いて経験的フィッシャー情報行列の逆行列を計算する手法を実装した。これよりパラメータの事後分布の広がりの程度について計算できるためパラメータ推定値の信頼性を出力できるようになった。これらのアルゴリズムは、人工的に生成したデータと、状態 X(t)の時間に順方向のサンプリングの結果と比較し計算の正しさを示すことができた。

(2) (1)で完成させたソフトウェアは線形方程式でのモデリングが適切な領域においては、生命データの解析に有用なツールであると考えられる。その一方で、生命過程を表現する微分方程式の多くは非線形微分方程式で表される。また(1)のモデルの実装は状態ベクトルの次元の二乗の行列のメモリ確保と解放を頻繁に繰り返すため、次元をあまり大きくすることができなかった。そこで、非線形関数による時間発展も許し、より計算効率の高いソフトウェアの実装を行った。まず状態変数 X(t)を中心座標 X0(t)と確率的揺らぎΔX(t)の和として表現し(X(t)=X0(t)+ΔX(t))、対数尤度を揺らぎ成分ΔX(t)でテイラー展開し二次項までで打ち切る近似を行った。中心座標 X0(t)の時間発展はニュートン・ラフソン法を繰り返し適用することで計算し、揺らぎ成分ΔX(t)については(1)で行ったものと同様の内向き・外向きアルゴリズムにより周辺化を行う方法をとった。アルゴリズムの導出の過程では、尤度関数の最大４回までの X(t)に関する偏微分が必要となり、各モデルについて人間が微分の表式を与えるのは煩雑であるため、自動微分を含むテンソル代数演算の機能も実装した。ソフトウェアの実装には成功したが、まだ十分なテストが行われていないため、今後は人口データによるテストと実データを用いた応用へとつなげていきたい。

５．主な発表論文等

〔雑誌論文〕　計5件（うち査読付論文　5件／うち国際共著　0件／うちオープンアクセス　3件）

１．著者名 Kuwabara Y、Tsuji S、Nishiga M、Izuhara M、Ito S、Nagao K、Horie T、Watanabe S、Koyama S、Kiryu H、Nakashima Y、Baba O、Nakao T、Nishino T、Sowa N、Miyasaka Y、Hatani T、Ide Y、Nakazeki F、Kimura M、Yoshida Y、Inada T、Kimura T、Ono K	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 434
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 オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Kiryu Hisanori、Ichikawa Yuto、Kojima Yasuhiro	４．巻 14
２．論文標題 TMRS: an algorithm for computing the time to the most recent substitution event from a multiple alignment column	５．発行年 2019年
３．雑誌名 Algorithms for Molecular Biology	６．最初と最後の頁 23
掲載論文のDOI（デジタルオブジェクト識別子） 10.1186/s13015-019-0158-3	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 Kojima Yasuhiro、Matsumoto Hirotaka、Kiryu Hisanori	４．巻 36
２．論文標題 Estimation of population genetic parameters using an EM algorithm and sequence data from experimental evolution populations	５．発行年 2019年
３．雑誌名 Bioinformatics	６．最初と最後の頁 221～231
掲載論文のDOI（デジタルオブジェクト識別子） 10.1093/bioinformatics/btz498	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、Kiryu Hisanori、Iwakiri Junichi、Sese Jun	４．巻 20
２．論文標題 reactIDR: evaluation of the statistical reproducibility of high-throughput structural analyses towards a robust RNA structure prediction	５．発行年 2019年
３．雑誌名 BMC Bioinformatics	６．最初と最後の頁 130
掲載論文のDOI（デジタルオブジェクト識別子） 10.1186/s12859-019-2645-4	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 Hirotaka Matsumoto, Hisanori Kiryu, Chikara Furusawa, Minoru S. H. Ko, Shigeru B. H. Ko, Norio Gouda, Tetsutaro Hayashi, Itoshi Nikaido	4．巻 33
2．論文標題 SCODE: An efficient regulatory network inference algorithm from single-cell RNA-Seq during differentiation.	5．発行年 2017年
3．雑誌名 Bioinformatics	6．最初と最後の頁 2314-2321
掲載論文のDOI（デジタルオブジェクト識別子） 10.1093/bioinformatics/btx194	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0件

〔産業財産権〕

〔その他〕

SCODE Source
https://github.com/hmatsu1226/SCODE
reactIDR Source
https://github.com/carushi/reactIDR
EMWER Source
https://github.com/kojikoji/EMWER

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