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　元 年　 ６ 月 １３ 日現在

機関番号：１２６０８
研究種目：基盤研究(B)（特設分野研究）
研究期間：2016～2018
課題番号：１６ＫＴ００２０

研究課題名（和文）文字列の非可換位相半群上の確率・統計理論とそれらの環境再生生物工学への応用

研究課題名（英文）Probability theory and statistics on a noncommutative topological monoid of strings and their application to bioremediation

研究代表者
小谷野　仁（Koyano, Hitoshi）

東京工業大学・生命理工学院・特任助教

研究者番号：１０５７０９８９

交付決定額（研究期間全体）：（直接経費）　　14,000,000円

研究成果の概要（和文）：アルファベット A = {a, c, g, t} 上の文字列がなす非可換位相半群 A* 上で、DNA 配列の集団の進化を記述する偏微分方程式を導出し、モデルの数理解析を行った。その混合モデルを用いて、ある環境中の DNA 配列の集団にその環境から掛かる淘汰圧の分布を表すよう設計された Laplace 様分布という分布を A* 上に導入し、その混合モデルのパラメーターを推定する方法を開発し、それに対して数理的基礎付けを与え、数値実験を行ってその有効性を確かめた。1 つの環境中の DNA 配列の集団の動態解析の方法を開発し、それを植物の周辺環境中の微生物群集に応用して、有効性を確かめた。

研究成果の学術的意義や社会的意義
1990 年代の前半に微生物を利用する環境浄化技術の研究が始まり、2000 年代半ばに、微生物によるバイオレメディエーション利用指針が制定された。このガイドラインでは、1 つの環境中の生物群集の時間発展を長期先まで予測することが求められているが、現在の数理・情報科学的技術では、このことを忠実に実行することは出来ない。本研究により、1 つの環境中の微生物群集の時間発展をコンピューターで計算することが出来るようになってきた。また、微生物群集のダイナミクスの特徴を抽出することも可能になった。今後は、本研究の結果を基礎にして、環境再生微生物工学の立案と制御のための体系的な技術の開発を目指す。

研究成果の概要（英文）：In this research project, we constructed a theory of a partial differential equation that describes the evolution of DNA sequences in an environment on the noncommutative topological monoid A* of strings on an alphabet A = {a, c, g, t}. Moreover, we introduced on A* a probability distribution named the Laplace-like distribution, which was designed to represent a population distribution of gene sequences in an environment using the mixture model of the distributions, and developed a theory for estimating the parameters of the mixture model. Furthermore, we developed a statistical method to analyze the dynamics of a population of DNA sequences in an environment and applied it to environmental samples of microbial sequences collected from surrounding environments of plants to demonstrate its usefulness.

研究分野：応用数学、数理統計学、バイオインフォマティクス

キーワード：文字列の位相半群　DNA 配列の集団　ダイナミクス　統計的推定　制御　バイオレメディエーション

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９、ＣＫ－１９（共通）

１．研究開始当初の背景

a. 研究の背景

データと言えば、数や数ベクトルなど数を用いて表されるものが大部分を占めていたが、近年、生物配列やテキストなどの文字列データの量が飛躍的に増加し、文字列データの統計的な解析方法が、様々な領域で求められるようになってきている。数を用いて表されるデータに対する統計学は、母集団に関して、ある確率法則に従って観測されたその一部である標本から推測するために、実数の集合 R や実ベクトル空間 R^p 上の確率論に基づいて厳密に構築されている。同様に、文字列データに対する統計学は、文字列の集合上の確率論に基づいて作られるのが自然であると考えられる。しかし、数学はその長い歴史の中で多くの対象について深い研究をしてきたが、文字列はあまり研究してこなかった。文字列は数学の対象というよりは、計算機科学の対象であり、その 1 領域である stringology は、これまでに文字列処理のためのアルゴリズムとデータ構造について徹底的に研究してきた。しかし、計算機科学は、ある対象の集合を考え、それに距離や算法を定義することにより位相構造や代数構造を与え、そのような数学的構造が付与された集合の上で関数や作用素などを研究するという数学的な方法では、文字列を研究してこなかった。

b. 私達のこれまでの研究

あるアルファベット A = { a_1, …, a_z } から作られる文字列の全体 A* は、数学の立場から見ると、編集距離によって距離空間に、連接によって半群になり、非可換な位相半群をなす。そこで、私達は、以前に、ある確率法則に従ってランダムに数や関数を生成する確率変数や確率過程を扱うための R 上や 2 乗可積分関数の全体がなす Hilbert 空間 L^2 上の確率論や確率過程論をモチーフにして、上述の数学的構造を持つ空間 A* 上で確率論を作り、ある確率法則に従ってランダムに生成された文字列を扱うための理論を作れないだろうかという考えを打ち出した。そうして、実際に、A* 上で確率論を展開し、それに基づいて統計理論を作り、それを応用して生物配列の解析を行なってきた。

２．研究の目的

a. A* 上の確率差分方程式論の構築

ある生物群集が持つ DNA やある遺伝子配列の全体とその分布は、それぞれアルファベット A = { a, c, g, t } 上の文字列の非可換位相半群 A* の部分集合と A* 上の分布として表される。私達は、まず、これまでに作ってきた A* 上の確率論を拡張して、各配列に確率的に起こる突然変異と環境からかかってくる淘汰圧の下で、生物配列の集団がどのように時間発展をしていくのか、すなわち生物配列の集団の進化を記述する A* 上の確率差分方程式の理論を構築する。

b. 文字列データに対するデータ同化法の開発

地球科学におけるこれまでのシミュレーション研究が示しているように、遠い過去や未来をコンピューターの中で再現したり、予測したりしようとすると、モデルと現実のずれが計算の過程で蓄積していき、最終的には大きな違いとなってしまうことがよく起こる。気象予測におけるデータ同化法は、この問題に対処して予測の精度を向上するために生まれた。高等生物と異なり、微生物の群集は環境の中で物質代謝のシステムを形成して、長い時間の中で環境へ適応しようとして進化するだけでなく、環境の方を改変する。私達は、データベース中の配列データをシミュレーションに取り込むことにより、上の 1 で構築した A* 上の確率差分方程式に基づいて、微生物配列の集団が環境と相互作用しながら、長い時間の中でゆっくりと、しかしダイナミックに変化していく様子をコンピューターの中で高精度に再現、予測するために文字列データの同化法を開発する。

c. 微生物を利用した環境再生工学への応用

古来、微生物は、人間の活動によって放出された様々な物質を分解し、自然環境を維持する役割を生態系の中で担ってきた。現在、環境の維持は、私達にとって最も重要な問題のうちの 1 つとなっているが、近年、微生物の分解能力を借りて汚染環境を浄化する技術の研究が進み、2000 年代半ばにその利用指針が制定された。その中で、利用する微生物の浄化終了後の増殖の可能性や、汚染現場の他の微生物への影響などの事前の評価の実施が求められ、1 つの環境下の微生物配列の集団の時間変化を長期先まで予測する技術が必要とされている。そこで、私達は、上の a と b において構築、開発した理論と方法と私達のこれまでの研究の結果を応用して、汚染現場から抽出された微生物配列の標本からその全体の分布を推定しながら、配列の集団の時間発展を高精度に予測し、汚染物質を分解する能力を持つ微生物の集団の導入規模や栄養源となる物質の供給量を変化させた時の、その時間発展の様々なシナリオをコンピューターの中で求めることにより、環境再生微生物工学の立案と制御のための技術を提供する。

３．研究の方法

数学、統計学、計算機科学、及び微生物生態学を専門とする 4 人の研究者が連携して,(1) A* 上の確率差分方程式の理論の構築（確率論)、(2) 文字列データに対するデータ同化法の開発(統計学)、(3) (1) と (2) の理論と方法に基づき、微生物配列のデータと環境データを同化さ

せるシミューレションシステムの開発（バイオインフォマティクス)、及び（4）（3）を応用した､微生物を利用する環境再生の立案と制御の技術の開発（微生物生態学）という 4 つの側面を持つ本研究課題に取り組む。

４．研究成果

アルファベット A = {a，c，g，t} 上の文字列がなす非可換位相半群 A* 上で、DNA 配列の集団の進化を記述する偏微分方程式を導出し、モデルの数理解析を行った。その混合モデルを用いて､ある環境中の DNA 配列の集団にその環境から掛かる淘汰圧の分布を表すよう設計された Laplace 様分布という分布を A* 上に導入し､その混合モデルのパラメーターを推定する方法を開発し、それに対して数理的基礎付けを与え、数値実験を行ってその有効性を確かめた。1 つの環境中の DNA 配列の集団の動態解析の方法を開発し､それを植物の周辺環境中の微生物群集に応用して、有効性を確かめた。

５．主な発表論文等

〔雑誌論文〕（計 76 件）

1. Koyano, H., Hayashida, M., and Akutsu, T., Maximum margin classifier working in a set of strings, Proceedings of the Royal Society A, 472(2187), 20150551, 2016. **(査読あり)**

2. Hayashida, M., Koyano, H., Finding median and center strings for a probability distribution on a set of strings under Levenshtein distance based on integer linear programming. Communications in Computer and Information Science, 690, 108-121, 2017. **(査読あり)**

3. Yano, K., Yano, Y., and Yen, J.-Y., Weak convergence of h-transforms for one-dimensional diffusions. Statistics and Probability Letters, 122, 152-156, 2017. **(査読あり)**

4. Guidi, L., Chaffron, S., Bittner, L., Eveillard, D., Larhlimi, A., Roux, S., Darzi, Y., Audic, S., Berline, L., Brum, J., Coelho, L.P., Espinoza, J.C., Malviya, S., Sunagawa, S., Dimier, C., Kandels-Lewis, S., Picheral, M., Poulain, J., Searson, S., Tara Oceans Consortium Coordinators, Stemmann, L., Not, F., Hingamp, P., Speich, S., Follows, M., Karp-Boss, L., Boss, E., Ogata, H., Pesant, S., Weissenbach, J., Wincker, P., Acinas, S.G., Bork, P., de Vargas, C., Iudicone, D., Sullivan, M.B., Raes, J., Karsenti, E., Bowler, C., and Gorsky, G., Plankton networks driving carbon export in the oligotrophic ocean, Nature, 532, 465-470, 2016. **(査読あり)**

5. Roux, S., Adriaenssens, E. M., Dutilh, B. E., Koonin, E. V., Kropinski, A. M., Krupovic, M., Kuhn, J. H., Lavigne, R., Brister, J. R., Varsani, A., Amid, C., Aziz, R. K., Bordenstein, S. R., Bork, P., Breitbart, M., Cochrane, G. R., Daly, R. A., Desnues, C., Duhaime, M. B., Emerson, J. B., Enault, F., Fuhrman, J. A., Hingamp, P., Hugenholtz, P., Hurwitz, B. L., Ivanova, N. N., Labonte, J. M., Lee, K.-B., Malmstrom, R. R., Martinez-Garcia, M., Mizrachi, I., Ogata, H., Paez-Espino, D., Petit, M.-A., Putonti, C., Rattei, T., Reyes, A., Rodriguez-Valera, F., Rosario, K., Schriml, L., Schulz, F., Steward, G. F., Sullivan, M. B., Sunagawa, S., Suttle, C. A., Temperton, B., Tringe, S. G., Vega, T. R., Webster, N. S., Whiteson, K. L., Wilhelm, S. W., Wommack, K. E., Woyke, T., Wrighton, K., Yilmaz, P., Yoshida, T., Young, M. J., Yutin, N., Allen, L. Z., Kyrpides, N. C., and Eloe-Fadrosh, E. A., Minimum Information about Uncultivated Virus Genomes (MIUViG): a community consensus on standards and best practices for describing genome sequences from uncultivated viruses. Nature Biotechnology, 37, 29-37, 2019. **(査読あり)**

など

〔学会発表〕（計 53 件）

1. Koyano, H., Hayashida, M., and Akutsu, T., Optimal string clustering based on a statistical theory on a topological monoid of strings, The 13th Workshop on Stochastic Models, Statistics and Their Applications, Berlin, Germany, February, 2017.

2. Hayashida, M., Kamada, M., and Koyano, H., Predicting strengths of protein-protein interactions through online regression algorithms, 2017 International Conference on Parallel and Distributed Processing Techniques and Applications, Las Vegas, USA, July, 2017.

3. Hayashida, M., Koyano, H., and Akutsu, T., Grammar-based compression for directed and undirected generalized series-parallel graphs using integer linear programming, The 9th International Conference on Bioinformatics Models, Methods and Algorithms, Funchal, Portugal, January, 2018.

4. Hayashida, M., Ishibashi1, K., and Koyano, H., Analyzing order of domains in grammar-based compression of a proteome, The 2018 International Conference on Parallel and Distributed Processing Techniques and Applications, Las Vegas, USA, July, 2018.

5. Hayashida, M. and Koyano, H., Artificial neural network approach to prediction of protein–RNA residue-base contacts, The 10th International Conference on Bioinformatics Models, Methods and Algorithms, Prague, Czech Republic, February, 2019.

など

〔図書〕（計　0　件）

〔産業財産権〕
○出願状況（計　0　件）

名称：
発明者：
権利者：
種類：
番号：
出願年：
国内外の別：

○取得状況（計　0　件）

名称：
発明者：
権利者：
種類：
番号：
取得年：
国内外の別：

〔その他〕
ホームページ等

６．研究組織

(1) 研究分担者
研究分担者氏名：緒方 博之
ローマ字氏名：Ogata Hiroyuki
所属研究機関名：京都大学
部局名：化学研究所
職名：教授
研究者番号（8 桁）：70291432

研究分担者氏名：矢野 孝次
ローマ字氏名：Yano Koji
所属研究機関名：京都大学
部局名：大学院理学研究科
職名：准教授
研究者番号（8 桁）：80467646

研究分担者氏名：林田 守広

ローマ字氏名：Hayashida Morihiro

所属研究機関名：松江工業高等専門学校

部局名：電気情報工学科

職名：准教授

研究者番号（8 桁）：40402929

(2)研究協力者

研究協力者氏名：

ローマ字氏名：