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

平成　２９　年　　５　月　２３　日現在

機関番号：　８２４０１
研究種目：　挑戦的萌芽研究
研究期間：　2014～2016
課題番号：　２６６１００３７

研究課題名（和文）文字列解析のための統計理論とその計算生化学への応用

研究課題名（英文）Statistical theory for string data analysis and its application to computational biochemistry

研究代表者

小谷野　仁（Hitoshi, Koyano）

国立研究開発法人理化学研究所・生命システム研究センター・研究員

研究者番号：１０５７０９８９

交付決定額（研究期間全体）：（直接経費）　　2,900,000円

研究成果の概要（和文）：本研究プロジェクトでは、まず、私達の以前の研究において文字列の非可換位相半群 A* 上で展開した確率論を拡張し、いくつかの極限定理を証明した。次に、これらの定理を用いて、A* においてマージン最大化原理の下で学習する学習機械の理論を構築し、それを RNA の 2 次構造とタンパク質間相互作用の予測問題に応用して、実際のデータ解析におけるその有用性を示した。更に、A* 上で混合モデルの理論を構築して、文字列データの教師なしクラスタリング方式を導出し、上述の定理を用いて、その最適性を証明した。最後に、A* 上の分布に対して中央及び中心文字列を定義し、それらを効率的に探索するアルゴリズムを構成した。

研究成果の概要（英文）：In this research project, we first demonstrated limit theorems, extending probability theory that we constructed on a noncommutative topological monoid A* of strings in our previous studies. We next developed a theory of a learning machine that learns under the maximum margin principle in A*, using these theorems, and subsequently applied the machine to the prediction problems of RNA secondary structures and protein-protein interactions to examine its usefulness in practical data analysis. Furthermore, we derived an unsupervised procedure for string clustering by constructing a theory of a mixture model on A* and demonstrated the optimality of the procedure based on the above-mentioned theorems. Lastly, we introduced median and center strings for a distribution on A* and constructed an algorithm that searches them efficiently.

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　確率論　統計学　機械学習　生物配列　バイオインフォマティクス

１．研究開始当初の背景

（１）これまでに数学は、その長い歴史の中で、集合、数、算法、多様体、方程式、関数、確率など多くの対象について深い研究をしてきた。数学がこれまでにあまり扱ってこなかったもので、理論計算機科学が新しく取り上げた対象に文字列がある。理論計算機科学の１領域である stringology は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深い研究を蓄積してきた。しかし、理論計算機科学では、与えられたアルファベットから作られる文字列の全体を考え、それに距離や算法を定義して位相構造や代数構造を与え、そのような数学的構造を持つ文字列の集合上で、例えば算法、関数、あるいは確率などを研究するという、数学的な仕方では文字列を研究してこなかった。

（２）Koyano and Kishino, *Physical Review E*, 2010 は、文字列の集合上に Levenshtein 距離を与えて距離空間とし、そこで確率論を展開した。文字列の集合は、連接という算法により半群をなすが、文字列に対して加法やスカラー乗法は定義できず、ベクトル空間にはならない。従って、実数や実ベクトルに対してと異なり、文字列に対しては平均や分散が定義されない。私達の上の論文では、n 個の文字列が与えられた時、それらの位置の尺度として平均の代わりにコンセンサス配列を、散らばりの尺度として通常の分散の代わりにコンセンサス配列からの Levenshtein 距離の平均を用い、確率文字列の列を観測した時の、コンセンサス配列の漸近的な挙動に関する結果を証明した。この結果は、数理統計学や統計力学において必須の道具である Euclid 空間における大数強法則の、文字列の集合におけるアナロジーと見なせる。また、この結果を応用して、上の意味での分散の漸近的な挙動に関する結果も示した。これらの結果は、漸近理論の枠組みで文字列データに対する統計理論を展開する際に必須の道具になる。これまでに、小谷野は、共同研究者らと上述の結果を用いて統計理論を展開し、それを応用して生物配列の解析を行ってきた。

２．研究の目的

（１）本研究プロジェクトでは、私達がこれまでに展開し、応用してきた文字列の集合上の確率論とそれに基づいた統計理論を更に拡張し、応用することにより、文字列データを解析するための統計的機械学習の理論を構築し、構築した学習機械を用いて RNA の 2 次構造とタンパク質間相互作用の予測問題に取り組む。

（２）コンピューターの普及により、近年ウェブ上などにテキストデータが大量に生産されている。また、生命科学の領域では、遺伝子、RNA やタンパク質のデータが大量に生み出されているが、これらは塩基とアミノ酸の配列であり、1 次データとしては文字列として表される。このため、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考えるのと同様に、ある確率法則に従ってランダムに文字列を生成する確率文字列が必要になってきている。また、統計学が確率論に基づいて厳密に構築されているのと同様に、テキストマイニングの方法やバイオインフォマティクスにおける生物配列解析の手法に対しても、文字列の集合上の確率論に基づいて様々な方法を展開する、提案した方法の理論的な根拠を示す、あるいは様々な方法を体系化するということが、今後求められてくると考えられる。本研究はこのような研究のための基礎となる数理的方法を提供するとともに、今後の方向付けを与える。

３．研究の方法

（１）本研究課題は次の 3 つの側面からなっている。

a. 確率論の側面: これまでに小谷野が共同研究者らと展開してきた文字列の集合上の確率論を更に拡張する。具体的には、次の b で統計的機械学習の理論を構築するために必要になる極限定理を揃える。数ベクトル空間、関数空間、あるいはネットワークなどこれまでに確率論や確率過程論が展開されてきた舞台とは異なる舞台で確率論を展開することにより、従来の確率論とは一味違った世界を作る。

b. 統計的機械学習の側面: a の確率論に基づいて、文字列データを教師付きの仕方で識別する統計的機械学習の理論を展開する。ハードマージンとソフトマージンの両方の場合に、マージン最大化原理の下での学習アルゴリズムを構成して、それらの計算量を評価し、その後、学習機械の汎化誤差を解析的な仕方で評価する。

c. 計算生化学の側面: b で構築した統計的学習機械を応用して計算生化学上の問題に取り組む。具体的には、塩基配列を用いた RNA の 2 次構造の予測問題とアミノ酸配列を用いたタンパク質間相互作用の予測問題に応用して、構築した学習機械の実際のデータ解析における有用性を検討する。

（２）このような多面的な研究を遂行するため、研究組織は小谷野と林田の 2 人のメンバーからなっており、2 人のこれまでの研究を踏まえて、次の分担で協力しながら、研究を遂行する。

a は小谷野が中心になり、林田が計算機科学における文字列処理の知識を生かして補佐しながら、共同で行う。

b は小谷野が中心になり、林田が c で計算生化学へ応用する際の視点から補佐しながら、共同で行う。

c は林田が中心になり、小谷野が遺伝子やタ

ンパク質など文字列として表される生物配列の統計解析の経験を生かして補佐しながら、共同で行う。

４．研究成果

本研究プロジェクトでは、まず、私達の以前の研究において、与えられたアルファベット A 上の文字列の全体が作る非可換位相半群 A*上で展開した確率論を拡張し、いくつかの極限定理を証明した。次に、これらの定理を用いて、A*においてマージン最大化原理の下で学習する統計的機械学習の理論を構築し、それを RNA の 2 次構造とタンパク質間相互作用の予測問題に応用して、実際のデータ解析におけるその有用性を示した。更に、A* 上で混合モデルの理論を構築して、文字列データの教師なしクラスタリング方式を導出し、上述の定理を用いて、その最適性を証明した。最後に、A*上の分布に対して中央及び中心文字列を定義し、それらを効率的に探索するアルゴリズムを構成した。

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 33 件）


[1] Hayashida, M. and 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. 査読有り

[2] An, A., Wang, J., Li, C., Revote, J., Zhang, Y., Naderer, T., Hayashida, M., Akutsu, T., Webb, G., Lithgow, T., and Song, J., SecretEPDB: A comprehensive web-based resource for secreted effector proteins of the bacterial types III, IV and VI secretion systems, *Scientific Reports*, **7**, 41031, 2017. 査読有り

[3] Koyano, H., Hayashida, M., and Akutsu, T., Optimal string clustering based on a statistical theory on a topological monoid of strings, *2017 13th Workshop on Stochastic Models, Statistics and Their Applications*, 207–208, 2017. 査読無し

[4] Koyano, H., Hayashida, M., and Akutsu, T., Maximum margin classifier working in a set of strings, *Proceedings of the Royal Society A*, **472**(2187), 2016. 査読有り

[5] Bao, B., Hayashida, M., and Akutsu, T., LBSizeCleav: improved support vector machine (SVM)-based prediction of Dicer cleavage sites using loop/bulge length, *BMC Bioinformatics*, **17**, 487, 2016. 査読有り

[6] Liu, L., Mori, T., Zhao, Y., Hayashida, M., Akutsu, T., Euler string-based compression of tree-structured data and its application to analysis of RNAs, *Current Bioinformatics*, **11**, 2016. 査読有り

[7] Jindalertudomdee, J., Hayashida, M., and Akutsu, T., Enumeration method for structural isomers containing user-defined structures based on breadth-first search approach, *Journal of Computational Biology*, 23, 625–640, 2016. 査読有り

[8] Jindalertudomdee, J., Hayashida, M., Zhao, Y., and Akutsu, T., Enumeration method for tree-like chemical compounds with benzene rings and naphthalene rings by breadth-first search order, *BMC Bioinformatics*, **17**, 113, 2016. 査読有り

[9] Li, C., Chang, C. C. H., Porebski, B. T., Hayashida, M., Akutsu, T., Song, J., and Buckle, A. M., Critical evaluation of in silico methods for prediction of coiled-coil regions in proteins, *Briefings in Bioinformatics*, **17**(2), 270–282, 2016. 査読有り

[10] Hayashida, M. and Akutsu, T., Complex network-based approaches to biomarker discovery, *Biomarkers in Medicine*, **10**(6), 621–632, 2016. 査読有り

[11] Koyano, H., Hayashida, M., and Akutsu, T., Optimal string clustering based on a Laplace-like mixture and EM algorithm on a topological monoid of strings, *Abstracts Book of the 1st IMA Conference on Theoretical and Computational Discrete Mathematics*, 10–11, 2016. 査読無し

[12] Hayashida, M. and Koyano, H., Integer linear programming approach to median and center strings for a probability distribution on a set of strings, *Proceedings of the 9th International Joint Conference on Biomedical Engineering Systems and Technologies*, **3**, 35–41, 2016. 査読無し

[13] Hayashida, M., Cao, Y., Zhao, Y., Grammar-based compression for unrooted trees and subclass of plane

graphs using integer linear programming, *IPSJ SIG Technical Report*, volume 2016-MPS-108, 2016. 査読無し

[14] Ruan, P., Hayashida, M., Akutsu, T., and Vert, J.-P., Improving accuracy for predicting heterodimeric protein complexes using combination kernels. *IPSJ SIG Technical Report*, volume 2016-MPS-108, 2016. 査読無し

[15] 林田 守広、小谷野 仁、文字列の集合上の確率分布における中央文字列および中心文字列に対する整数計画問題、*IPSJ SIG Technical Report*、volume 2016-MPS-108、number 7、2016 年 査読無し

[16] Hayashida, M., Jindalertudomdee, J., Zhao, Y., and Akutsu, T., Parallelization of enumerating tree-like chemical compounds by breadth-first search order, *BMC Medical Genomics*, **8**(Supplement 2), S15, 2015. 査読有り

[17] Zhao, Y., Hayashida, M., Cao, Y., Hwang, J., and Akutsu, T., Grammar-based compression approach to extraction of common rules among multiple trees of glycans and RNAs, *BMC Bioinformatics*, **16**, 128, 2015. 査読有り

[18] 小谷野 仁、林田 守広、文字列クラスタリングのための Laplace 様混合モデルに対する EM アルゴリズム、情報処理学会第 77 回全国大会講演論文集、 3B-04、1-247-1-248 ページ、2015 年 査読無し

[19] 小谷野 仁、林田 守広、阿久津 達也、文字列の集合上の Laplace 様混合モデルと EM アルゴリズムに基づく文字列クラスタリング、*IPSJ SIG Technical Report*、volume 2015-MPS-103、number 31、2015 年 査読無し

[20] 鎌田 真由美、林田 守広、条件付き確率場を用いたタンパク質残基間コンタクト予測、*IPSJ SIG Technical Report*, volume 2015-MPS-103、2015 年 査読なし

[21] 劉 立偉、森 智弥、趙 楊、林田 守広、阿久津 達也、根付き順序木の圧縮における分割型文法とオイラー文字列との比較、*IPSJ SIG Technical Report* 、 volume 2015-MPS-103、2015 年 査読無し

[22] Koyano, H., Tsubouchi, T., Kishino, H., and Akutsu, T., Archaeal beta diversity patterns under the seafloor along geochemical gradients, *Journal of Geophysical Research G: Biogeosciences*, **119**(9), 1770-1788, 2014. 査読有り

[23] Hayashida, M., Ruan, P., and T. Akutsu., Proteome compression via protein domain compositions, *Methods*, **67**, 380-385, 2014. 査読有り

[24] Hayashida, M. and Akutsu, T., Domain-based approaches to prediction and analysis of protein-protein interactions, in *International Journal of Knowledge Discovery in Bioinformatics*, volume 4, chapter 3, pages 24-41, 2014. 査読有り

[25] Kamada, M., Sakuma, Y., Hayashida, M., and Akutsu, T., Prediction of protein-protein interaction strength using domain features with supervised regression, *The Scientific World Journal*, 2014:240673, 2014. 査読有り

[26] Ruan, P., Hayashida, M., Maruyama, O., and Akutsu, T., Prediction of heterotrimeric protein complexes by two-phase learning using neighboring kernels, *BMC Bioinformatics*, **15**(Supplement 2), S6, 2014. 査読有り

[27] Koyano, H. and Hayashida, M., Probability theory on a topological monoid of strings and its application to statistical machine learning, *Abstract Book of the International Conference on Recent Advances in Pure and Applied Mathematics 2014*, 166, 2014. 査読無し

[28] Hayashida, M., Koyano, H., and Akutsu, T., Measuring the similarity of protein structures using image local feature descriptors SIFT and SURF, *2014 8th International Conference on Systems Biology*, 167-171, 2014. 査読無し

[29] Hayashida, M., Jindalertudomdee, J., Zhao, Y., and Akutsu, T., Parallelization of enumerating tree-like chemical compounds by breadth-first search order, *IPSJ SIG Technical Report*, volume, volume 2014-MPS-97, 2014. 査読無し

[30] Ruan, P., Hayashida, M., Maruyama, O., Akutsu, T., Prediction of heterotrimeric protein complexes by two-phase learning using neighboring kernels, *IPSJ SIG Technical Report*, volume 2014-MPS-98, 2014. 査読無し

[31] Zhao, Y., Hayashida, M., Cao, Y., Hwang, J., and Akutsu, T., Grammar-based compression for multiple trees using integer programming. *IPSJ SIG Technical Report*, volume 2014-MPS-98, 2014. 査読無し

[32] Ruan, P., Hayashida, M., and Akutsu, T., Study on weight of protein-protein interaction network for prediction of heterodimers, 電子情報通信学会総合大会通信講演論文集 2, pages S-189-S-190, 2014. 査読無し

[33] 小谷野 仁、林田 守広、阿久津 達也、文字列の距離空間上の最大マージン識別器とそのタンパク質科学への応用、*IPSJ SIG Technical Report*、volume 2014-MPS-98、number 13、2014 年 査読無し

〔学会発表〕（計 21 件）

[1] Koyano, H., Hayashida, M., and Akutsu, T., Optimal string clustering based on a statistical theory on a topological monoid of strings, 13th Workshop on Stochastic Models, Statistics and Their Applications, Berlin, Germany, February 24, 2017.

[2] Jindalertudomdee, J., Hayashida, M., Song, J., and Akutsu, T., Host-pathogen protein interaction prediction based on local topology structures of a protein interaction network, IEEE 16th International Conference on Bioinformatics and BioEngineering, Taichung, Taiwan, October 31, 2016.

[3] Hayashida, M. and Koyano, K., Integer linear programming approach to median and center strings for a probability distribution on a set of strings, 15th European Conference on Computational Biology, The Hague, Netherlands, September 3, 2016.

[4] 林田 守広、小谷野 仁、文字列の集合上の確率分布における中央文字列および中心文字列に対する整数計画問題、日本情報処理学会「数理モデル化と問題解決研究会」、「バイオ情報学研究会」及び日本電子情報通信学会「ニューロコンピューティング研究会」、「情報論的学習理論と機械学習研究会」合同研究会、沖縄科学技術大学院大学、2016 年 7 月 4 日

[5] Koyano, H., Hayashida, M., and Akutsu, T., Optimal string clustering based on a Laplace-like mixture and EM algorithm on a topological monoid of strings, 1st IMA Conference on Theoretical and Computational Discrete Mathematics, Derby, United Kingdom, March 23, 2016.

[6] Hayashida, M. and Koyano, H., Integer linear programming approach to the calculation of median and center strings for a probability distribution on a set of strings, 7th International Conference on Bioinformatics Models, Methods and Algorithms, Rome, Italy, February 22, 2016.

[7] Mihara, T., Koyano, H., Hingamp, P., Goto, S., and Ogata, H., Assessing the diversity of the Megaviridae giant viruses based on RNA polymerase beta genes, 2015 年日本バイオインフォマティクス学会年会, 京都大学, 2015 年 10 月 30 日

[8] Ngouv, H., Hayashida, M., Nacher, J., and Akutsu, T., Prediction of ncrna-disease association based on sequence expression and tripartite network. 2015 Annual Conference of the Japanese Society for Bioinformatics, Kyoto, Japan, October 29, 2015.

[9] 小谷野 仁、林田 守広、阿久津 達也、文字列データの統計的クラスタリングのための Laplace 様混合モデルと EM アルゴリズムの理論、日本応用数理学会 2015 年度年会、金沢大学、2015 年 9 月 9 日

[10] Hayashida, M., Kamada, M., and Koyano, H., Online learning approach to prediction of protein-protein interaction strengths, 9th International Conference on Systems Biology, Luoyang, China, August 22, 2015.

[11] Mihara, T., Koyano, H., Goto, S., and Ogata, H., Diversity of marine giant DNA viruses, International Workshop on Bioinformatics and System Biology, Boston, USA, July 20, 2015.

[12] 小谷野 仁、林田 守広、阿久津 達也、文字列の集合上の Laplace 様混合モデルと EM アルゴリズムに基づく文字列クラスタリグ、日本情報処理学会「数理モデル化と問題解決研究会」、「バイオ情報学研究会」及び日本電子情報通信学会「ニューロコンピューティング研究会」、「情報論的学習理論と機械学習研究会」合同研究会、沖縄科学技術大学院大学、2015 年 6 月 24 日

[13] 小谷野 仁、林田 守広、文字列クラスタリングのための Laplace 様混合モデルに対す

る EM アルゴリズム、日本情報処理学会第 77 回全国大会、京都大学、2015 年 3 月 18 日

[14] Ruan, P., Hayashida, M., Akutsu, T., Vert, J.-P., Improving accuracy for predicting heterodimeric protein complexes using combination kernels, 13th Asia-Pacific Bioinformatics Conference (Best Poster Presentation Award), HsinChu, Taiwan, January 21, 2015.

[15] Koyano, H. and Hayashida, M., Probability theory on a topological monoid of strings and its application to statistical machine learning, International Conference on Recent Advances in Pure and Applied Mathematics, Antalya, Turkey, November 7, 2014.

[16] Hayashida, M., Koyano, H., and Akutsu, T., Measuring the similarity of protein structures using image local feature descriptors SIFT and SURF, 8th International Conference on Systems Biology, Qingdao, China, October 25, 2014.

[17] Hayashida, M., Jindalertudomdee, J., Zhao, Y., and Akutsu, T., Parallelization of enumerating tree-like chemical compounds by breadth-first search order. 8th International Conference on Systems Biology, Qingdao, China, October 25, 2014.

[18] Koyano, H., Probability theory on a topological monoid of strings and its application to machine learning, Sweden–Kyoto Symposium (co-organized by Uppsala University, Stockholm University, Royal Institute of Technology, Karolinska Institute, and Kyoto University), Stockholm, Sweden, September 12, 2014.

[19] 小谷野 仁、林田 守広、阿久津 達也、文字列の距離空間上の確率論とその機械学習への応用、日本応用数理学会 2014 年度年会、政策研究大学院大学、2014 年 9 月 5 日

[20] 小谷野 仁、林田 守広、阿久津 達也、文字列の距離空間上の最大マージン識別器とそのタンパク質科学への応用、日本情報処理学会「数理モデル化と問題解決研究会」、「バイオ情報学研究会」及び日本電子情報通信学会「ニューロコンピューティング研究会」、「情報論的学習理論と機械学習研究会」合同研究会、沖縄科学技術大学院大学、2014 年 6 月 26 日

[21] Ruan, P., Hayashida, M., Maruyama, O., and Akutsu, T., Prediction of heterotrimeric protein complexes by two-phase learning using neighboring kernels, 12th Asia Pacific Bioinformatics Conference, Fudan University, Shanghai, China, January 17, 2014.

〔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：
出願年月日：
国内外の別：

○取得状況（計０件）

名称：
発明者：
権利者：
種類：
番号：
出願年月日：
取得年月日：
国内外の別：

〔その他〕
ホームページ等

６．研究組織
(1)研究代表者
小谷野　仁（KOYANO, Hitoshi）
国立研究開発法人理化学研究所・生命システム研究センター・研究員
研究者番号：10570989

(2)研究分担者
林田　守広（HAYASHIDA, Morihiro）
松江工業高等専門学校・電気情報工学科・准教授
研究者番号：40402929

(3)連携研究者
（　　）

研究者番号：