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　５ 年　 ６ 月 １２ 日現在

機関番号：１７１０４
研究種目：基盤研究(B)（一般）
研究期間：2018～2022
課題番号：１８Ｈ０３３３５

研究課題名（和文）生物のde novo遺伝子探索アルゴリズムの探究

研究課題名（英文）Algorithms that organisms search genes de novo

研究代表者
矢田　哲士（Yada, Tetsushi）

九州工業大学・大学院情報工学研究院・教授

研究者番号：１０３２２７２８

交付決定額（研究期間全体）：（直接経費）　　13,200,000円

研究成果の概要（和文）：遺伝子のde novo誕生は、ゲノム中の非遺伝子領域に突然変異が蓄積することで遺伝子が誕生する過程である。これまで、このような過程はほとんど起きないと考えられてきたが、ゲノム研究の進展により、それが遥かに一般的な過程であることが明らかになった。一方、この過程は、生物が新しい配列の遺伝子を探索する過程として捉えることができる。すると、僅か90塩基から成る短い遺伝子のde novo誕生でさえ、4の90乗を越える広大な状態空間が探索されていることになる。ここでは、ゲノムデータのバイオインフォマティクス解析により、広大な状態空間から数々の遺伝子を巧みに探しだす生物のアルゴリズムの全容を明らかにする。

研究成果の学術的意義や社会的意義
出芽酵母に至る系統でのバイオインフォマティクス解析により、遺伝子のde novo誕生の典型的な過程、すなわち、GCに富む領域に中立な突然変異が蓄積することで、まず、候補遺伝子領域長が伸長し、次に、翻訳シグナル配列を獲得する、を明らかにした。そして、候補遺伝子領域長を伸長する中立な突然変異の数が翻訳シグナル配列を獲得するその数より多いことから、遺伝子のde novo誕生が機会的な過程であることを見いだした。また、自然言語処理の分野で発展した様々な技術を応用することで、遺伝子領域長に関係なく、それらのタンパク質コーディング性を推定するdeep learningモデルを初めて開発した。

研究成果の概要（英文）：De novo gene birth is the process that new genes arise from non-genic DNA sequences by accumulating mutations. Until recently, this process was thought to occur almost never, but advances in genome research have revealed that it is a far more common process. On the other hand, this process can be viewed as an organism's search for new gene sequences. Then, even the de novo birth of a short gene consisting of only 90 nucleotides would involve the exploration of a vast state space of more than 4 to the 90th power. Here, we revealed the full extent of the algorithm of the organism that efficiently searches for a number of genes from the vast state space by applying bioinformatics analysis of genome data.

研究分野：バイオインフォマティクス

キーワード：遺伝子のde novo誕生　生物の遺伝子探索アルゴリズム　バイオインフォマティクス解析

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　(共通)

## １．研究開始当初の背景

長い間、ヒトゲノムのほとんどの領域は機能をもたないジャンク DNA だと考えられてきた。しかし、ENCODE プロジェクト (`https://www.encodeproject.org`) により、その 4 分の 3 の領域が転写されていることが明らかになった (Djebali,S., *Nat.*, 2012)。もちろん、転写されているからといって、それらの全てから機能的な転写産物が生みだされている、すなわち、遺伝子であるとはいえない (Doolittle,W.F., *Proc. Adad. Natl. Sci, USA*, 2013)。それらの中には、転写のノイズと考えられるものが数多く含まれている。しかし、この発見は、環境に適応するために、ゲノム中のさまざまな座位で転写や翻訳の試行錯誤を繰り返しながら新しい遺伝子を探索する生命像をゲノム研究者に抱かせた。

一方、同じように長い間、新しい遺伝子は既にある遺伝子の重複や混成によって生みだされ、*de novo* に生みだされること (突然変異の蓄積によって非遺伝子領域に新しい遺伝子が生みだされること) はほとんどないと考えられてきた。(Kaessmann,H., *Genome Res.*, 2010)。しかし、次世代シークエンサーの登場により、RNA-seq やリボソームプロファイリングのデータが蓄積されて遺伝子の転写や翻訳の全体像が明らかになると、これまで考えられてきたよりずっと多くの遺伝子が *de novo* に生みだされていて、それらの中には、重要な機能をもつものが含まれていることが明らかになった (McLysaght,A. & Guerzoni,D., *Phil. Trans. R. Soc. B*, 2015)。例えば、出芽酵母 *Saccharomyces cerevisiae* (*S. cerevisiae*) では、*Saccharomyces paradoxus* (*S. paradoxus*) との分岐後に生みだされた *de novo* 遺伝子の数は、重複や混成によって生みだされた遺伝子の数の 5 倍に達することが報告され (Carvunis,A.R. *et al.*, *Nat.*, 2012)、それらの中には、DNA 修復に関与し、合成致死性を備える遺伝子 BSC4、栄養成長を促進する遺伝子 MDF1 などが含まれていた (Cai,J. *et al.*, *Genet.*, 2008) (Li,D. *et al.*, *Cell Res.*, 2010) 。そして、出芽酵母に加え、ゲノムデータが豊富なヒトやショウジョウバエにおける *de novo* 遺伝子の体系的な探索が始まり、ORF 長やコドンの使用頻度や発現パターンなど、それらの統計的な特徴の幾つかが明らかになった (Schlötterer,C., *Trends in Genet.*, 2015)。しかし、*de novo* 遺伝子がどのようにして生まれ、どのようにして機能を獲得するかについては、ほとんど分かっていない (McLysaght,A. & Guerzoni,D., *Phil. Trans. R. Soc. B*, 2015)。

## ２．研究の目的

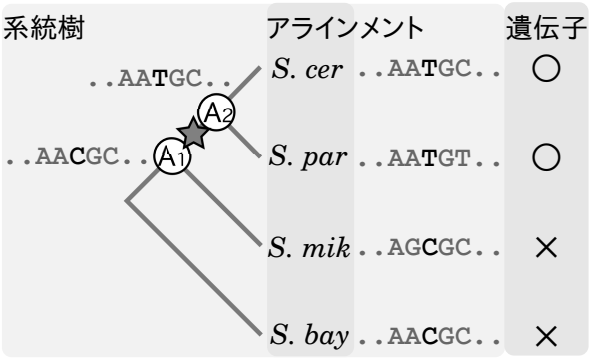
遺伝子の *de novo* 誕生は、生物が全く新しい配列の遺伝子を探索する過程として捉えることができる。*De novo* に誕生した遺伝子には配列の短いものが多いが、それでも、その配列長が 90 塩基を越えることは珍しくない。つまり、遺伝子の de novo 誕生の過程では、$4^{90}$ を越える広大な状態空間を探索していると言える。

ここでは、ゲノムデータのバイオインフォマティクス解析を通して、さまざまな生物種における遺伝子の *de novo* 誕生 (探索) の過程を塩基配列の解像度で詳らかにし、その多様性と普遍性を明らかにする。さらに、その探索過程の効率を高めるために生物が採用している戦略を明らかにし、生物が広大な状態空間から、数々の洗練された遺伝子を巧みに探しだすアルゴリズムの全容を明らかにする。具体的には、まず、(A) 遺伝子の *de novo* 誕生の過程を *S. cerevisiae* に至る系統で明らかにする。次に、(B)*de novo* 遺伝子を網羅的に発見するために、ヒト long non-coding RNA (lncRN) からタンパク質をコードする短いオープンリーディングフレーム[1]short open reading frame (sORF) を発見するコンピュータアルゴリズムを開発する。さらに、(C) 遺伝子の *de novo* 誕生の速度が遺伝子のエキソン–イントロン構造によって高められてい

[1]オープンリーディングフレームは、塩基配列上に任意のリーディングフレーム (読み枠) を設定して各コドンに対応するアミノ酸を順次当てはめる時、終止コドンが現れるまでアミノ酸コドンが続くようなオープンな (開けた) 読み枠の配列を指す。

る可能性を示す。なお、ここでは、タンパク質をコードする遺伝子の *de novo* 誕生を考える。

## 3．研究の方法




(A) *S. cerevisiae* ゲノムでは、1,916 個の *de novo* 遺伝子が近縁種における保存度とともにアノテーションされた (Carvunis,A.R. *et al.*, *Nat.*, 2012)。また、*S. cerevisiae* とこれらの近縁種では、進化系統樹とゲノム配列が明らかになっている (`https://www.yeastgenome.org/`)。今、*Saccharomyces cerevisiae* (*S.cer*)、*Saccharomyces paradoxus* (*S.par*)、*Saccharomyces mikatae* (*S.mik*)、*Saccharomyces bayanus* (*S.bay*) の系統樹とこれらのゲノムの相同領域のアラインメントが与えられ、その相同領域の *S.cer* と *S.par* には *de novo* 遺伝子が存在し、*S.mik* と *S.bay* には存在しないとする (右図)。今、この *de novo* 遺伝子の存在を、最も少ない回数の遺伝子の誕生と消失で説明しようとすると、この遺伝子は、系統樹の星印の枝で誕生したと考えられる。一方、アラインメントでは、各カラムに整列された塩基は、同祖的な塩基 (共通祖先のゲノム中の同じ塩基から分岐した塩基) だと考えることができるので、アラインメントの各カラムの塩基から、共通祖先ゲノムにおける最も尤もらしい塩基配列を推定することができる。ここでは、この相同領域について、遺伝子が *de novo* に誕生した前と後の共通祖先 $A_1$ と $A_2$ のゲノム配列を推定する。この共通祖先のゲノム配列の推定を、全ての *de novo* 遺伝子について繰り返し、$A_1$ に当たるゲノム配列と $A_2$ に当たるゲノム配列の間に生じた変化の統計的な特徴をまとめる。そして、それらが遺伝子の *de novo* 誕生にどのように関わったかを明らかにする。

(B) 配列長が 100 コドン以上の canonical ORF(以下、ORF) では、事前に学習したコーディング ORF とノンコーディング ORF における読み枠中の 6-mer(コドンの 1 文字目から始まる 6-mer) の出現頻度を用い、与えられた ORF の読み枠中の 6-mer の出現頻度と照合することで、そのコーディング性を高い精度で予測することができる。しかし、この方法を sORF に適用すると、ORF が短いことに起因する問題が生じる。まず、(B1) 与えられた sORF から取り出せる読み枠中の 6-mer は、高々数十である。このような少数のデータによるコーディング性の予測は、統計的な信頼性を大きく損う。この傾向は、短い sORF で一層深刻になる。また、(B2) 全ての可能な読み枠中の 6-mer は、$4^6$ 通りもある。統計的な信頼性を保ってこれらの出現頻度を学習するには、用意できるコーディング sORF のデータは小さ過ぎる。

これらの問題を克服するために、ここでは、sORF における読み枠中の 6-mer の出現頻度を低次の $k^{(j)}$-mer(コドンの $j$ 番目の塩基から始まる $k$-mer) の出現頻度から推定する。例えば、$4^{(j)}$-mer の出現頻度を用いると、読み枠中の 6-mer の出現頻度を

$$P(c_i^1\ c_i^2\ c_i^3\ c_{i+1}^1 c_{i+1}^2 c_{i+1}^3) \simeq P(c_i^1\ c_i^2\ c_i^3\ c_{i+1}^1) P(c_{i+1}^2 | c_i^2\ c_i^3\ c_{i+1}^1) P(c_{i+1}^3 | c_i^3\ c_{i+1}^1 c_{i+1}^2)$$

で近似的に推定する (ここで、$c_i^j$ は sORF 中の $i$ 番目のコドンの $j$ 番目の塩基)。これにより、読み枠中の 6-mer の出現頻度を用いる場合に比べ、sORF から取り出せる $4^{(j)}$-mer の数は 3 倍に増え $(3(n-1)/(n-1))$、学習する出現頻度の数はおよそ 1/5 に減る $((3\times 4^4)/4^6)$。

さらに、sORF に現われる一連の低次の $k^{(j)}$-mer を入力とし、その sORF のコーディング性を高い精度で予測する DNN(deep neural network) を開発する。ここでは、ゲノム配列と自然言語に共通の構造が観察されることに注目して、注意機構や転移学習と言った近年の自然言語処理に画期的な進展をもたらした先端的なフレームワークを導入する。注意機構では、sORF のコーディング性の予測において、注目すべき低次

の $k^{(j)}$-mer を sORF 全体の配列 (文脈) から動的に特定する。転移学習では、まず、データが豊富な ORF について、そのコーディング性を予測する DNN を用意し、それをデータが限られる sORF のコーディング性の予測に適応させる。

(C) 遺伝子のイントロンレス構造は、0 個のイントロンが挿入されたエキソン–イントロン構造だと考えられる。すると、遺伝子のイントロンレス構造はエキソン–イントロン構造の部分集合だと考えることができるので、ある塩基長のゲノム配列に存在するエキソン–イントロン構造の候補遺伝子の数は、イントロンレス構造の候補遺伝子の数を上回る。そして、ノイズレベルでの候補遺伝子の転写と翻訳を通し、その中の一定の割り合いが機会的に遺伝子に進化すると考えると、イントロンが存在することで、遺伝子の *de novo* 誕生が促進されると期待される。

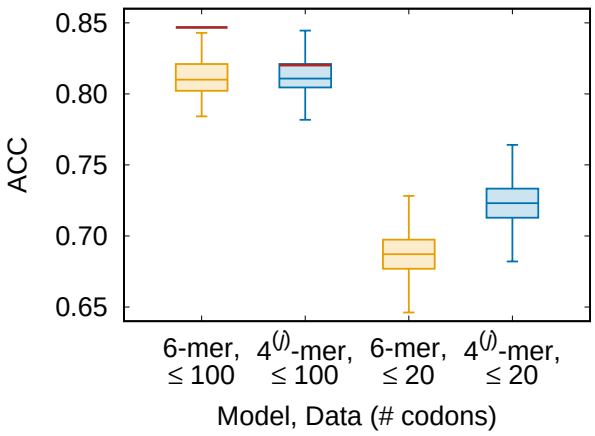
ここでは、遺伝子の *de novo* 誕生の速度が遺伝子のエキソン–イントロン構造によって高められている可能性を示すために、まず、ランダムに取りだした 10 kb のヒトゲノム配列群に含まれる候補遺伝子を数え上げ、候補遺伝子数がイントロンの数にどのように依存するかを明らかにする。さらに、遺伝子の *de novo* 誕生のシミュレータを開発し、遺伝子の *de novo* 誕生にイントロンの導入が与える影響を評価する。このシミュレータでは、まず、塩基の並びからなる染色体の集団を用意し、各染色体中の候補遺伝子を列挙する。次に、各候補遺伝子を、解として用意した遺伝子 (解遺伝子) と配列アラインメントして、そのスコアに基づいて候補遺伝子が由来する染色体に適応度を与える。そして、次世代の集団を構成する染色体を適応度に基づいて選択し、選択された染色体の塩基配列に交叉や突然変異などの操作を加える。ここでは、候補遺伝子にイントロンが含まれる場合と含まれない場合について、遺伝子の *de novo* 誕生の速度に有意差がないか、統計的に検定する。

## ４．研究成果

(A) 遺伝子が *de novo* に誕生した前と後の配列には、以下のような統計的な特徴が観察された。まず、これらの配列は、ともに GC に富む傾向があった。次に、遺伝子が *de novo* に誕生する間に蓄積した突然変異は、非遺伝子領域の配列に蓄積した突然変異とは差がなかった。また、遺伝子が *de novo* に誕生する間には、ORF の伸長は観察されたが、翻訳開始のシグナル配列 (Kozak 配列) の獲得は観察されなかった。しかし、より長い進化時間では、Kozak 配列の獲得が報告されている (Carvunis,A.R. *et al.*, *Nat.*, 2012)。

以上の観察より、以下のような遺伝子の de novo の誕生の過程を描くことができる。(A1) はじめに GC-rich なゲノム領域ありき、(A2) そこに中立な突然変異が蓄積し、(A3) まず ORF が伸長し、(A4) 次に翻訳開始のシグナル配列が獲得される。ここで興味深いのは、(A2) から (A4) に進むにつれて、各ステップに寄与する中立な突然変異の数が減っていることである。このことは、遺伝子の de novo 誕生の過程が、各ステップを引き起こす中立な突然変異の数の多さによって機会的に支配されていることを示している (遺伝子の de novo 誕生の中立説)。

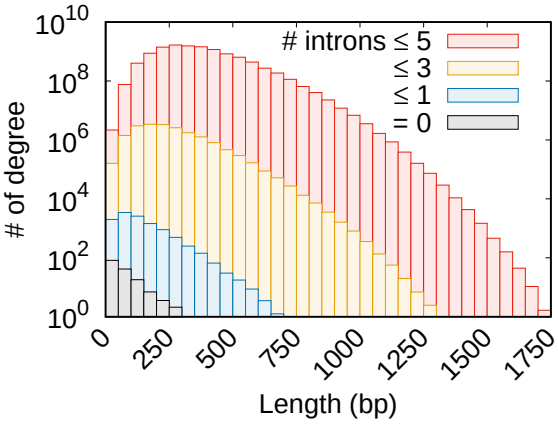
(B) 低次の $k^{(j)}$-mer を導入する効果を確かめるために、sORF のコーディング性の予測精度を評価した (右図)。sORF 全体について、交差検定による読み枠中の 6-mer の予測精度 (箱ひげ) は、クローズドテストのもの (赤線) から大きく低下した。これは、問題 (B2) による過学習が起きていることを示す。一方、$4^{(j)}$-mer では、過学習が起きることなく、また、低次の統計量であるにも関わらず、読み枠中の 6-mer と同等の予測精度を示した。また、短い sORF(≤20 コドン) について、交差検定による読み枠中の 6-mer の予測精度は、sORF 全体に



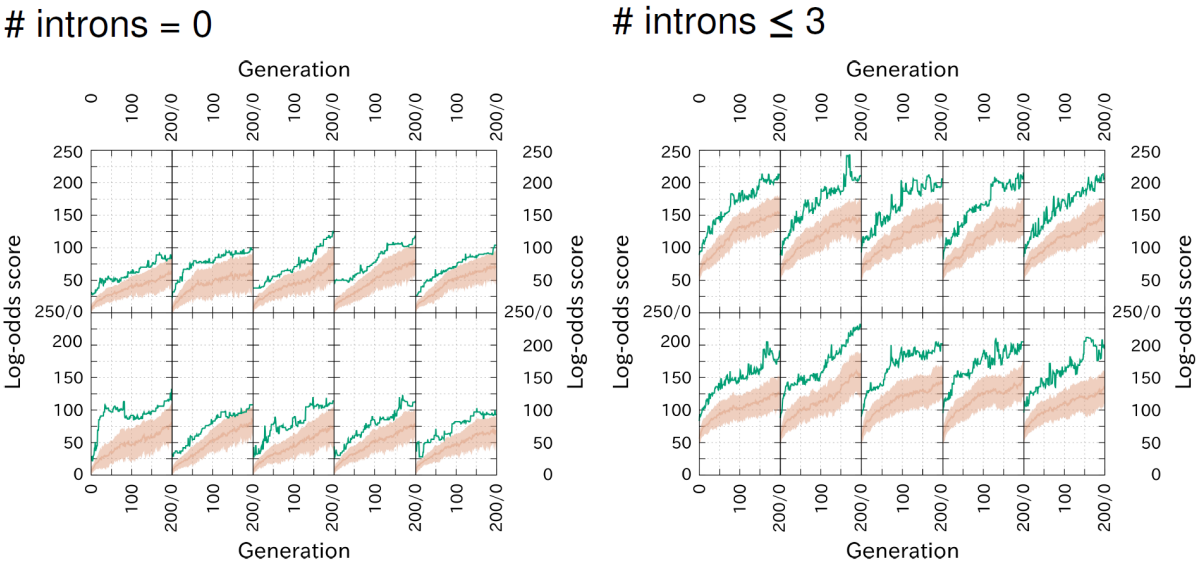
ついてのものから大きく低下した。これは、問題 (B1) が深刻化したことを示す。一方、$4^{(j)}$-mer では、予測精度の低下が抑えられ、読み枠中の 6-mer より高い予測精度を示した。

さらに、DNN による sORF のコーディング性の予測精度を評価したところ、sORF に対する予測精度は 0.922、canonical ORF に対する予測精度は 0.956 であった。このことは、ここでの試みにより、sORF だけでなく、canonical ORF のコーディング性も高い精度で推定できる汎用モデルの開発に初めて成功したことを示している。

(C) ランダムに取りだした 10 kb のヒトゲノム配列群に含まれる候補遺伝子を数え上げたところ、候補遺伝子に含まれるイントロンの数が増えると、ゲノム中の候補遺伝子の数が指数的に増大し、候補遺伝子の配列が長くなった (右図)。このことは、遺伝子の配列にイントロンが存在すると、遺伝子の *de novo* 誕生が促進される可能性があること、また、長い遺伝子を生みだす下地が自然に整うことを示唆している。さらに、ゲノム中には膨大な数の候補遺伝子が存在することから、それらは互いに重複し、ひとつの突然変異が数多くの候補遺伝子の配列に変化をもたらすことを示している。




さらに、遺伝子の *de novo* 誕生シミュレータを開発し、遺伝子の *de novo* 誕生にイントロンの導入が与える影響を評価した (下図)。すると、全ての試行において、候補遺伝子にイントロンが含まれる場合 (下右図) では、含まれない場合 (下左図) より解遺伝子に配列が似た遺伝子が生みだされた。また、Brunner-Munzel 検定により、初期世代集団における適応度の平均値 (橙線) と最大値 (緑線)、最終世代集団における適応度の平均値と最大値、初期世代集団と最終世代集団の適応度の平均値と最大値の差のいずれにおいても、候補遺伝子にイントロンが含まれる場合と含まれない場合には統計的な有意差が検出された ($p$ 値 $\leq 4.303 \times 10^{-5}$)。このことは、遺伝子の *de novo* 誕生の速度が遺伝子のエキソン–イントロン構造によって高められている可能性があることを示唆している。



５．主な発表論文等

〔雑誌論文〕　計3件（うち査読付論文　3件／うち国際共著　0件／うちオープンアクセス　1件）

| | |
|---|---|
| １．著者名<br>Yada T, Taniguchi T | ４．巻<br>in press |
| ２．論文標題<br>A putative scenario of how de novo protein-coding genes originate in the Saccharomyces cerevisiae lineage | ５．発行年<br>2023年 |
| ３．雑誌名<br>BMC Bioinformatics | ６．最初と最後の頁<br>- |
| 掲載論文のDOI（デジタルオブジェクト識別子）<br>なし | 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>Ichinose N, Kawashima T, Yada T, Wada H | ４．巻<br>526 |
| ２．論文標題<br>Dynamical robustness and its structural dependence in biological networks | ５．発行年<br>2021年 |
| ３．雑誌名<br>J Theor Biol | ６．最初と最後の頁<br>110808 |
| 掲載論文のDOI（デジタルオブジェクト識別子）<br>10.1016/j.jtbi.2021.110808 | 査読の有無<br>有 |
| オープンアクセス<br>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<br>－ |

| | |
|---|---|
| １．著者名<br>Yada T | ４．巻<br>2 |
| ２．論文標題<br>Genome sequence alignment | ５．発行年<br>2019年 |
| ３．雑誌名<br>Encyclopedia of Bioinformatics and Computational Biology (Gaeta B, Nakai K, ed.) | ６．最初と最後の頁<br>268-283 |
| 掲載論文のDOI（デジタルオブジェクト識別子）<br>なし | 査読の有無<br>有 |
| オープンアクセス<br>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<br>－ |

〔学会発表〕　計4件（うち招待講演　0件／うち国際学会　1件）

| |
|---|
| １．発表者名<br>Yada T, Taniguchi T |
| ２．発表標題<br>A putative scenario of how de novo protein-coding genes originate in the Saccharomyces cerevisiae lineage |
| ３．学会等名<br>GIW/ISCB Asia 2022（国際学会） |
| ４．発表年<br>2022年 |

１．発表者名
矢田哲士，佐藤巽

２．発表標題
低次のk-merの出現頻度を用いてRNA配列中のコーディングsmORFを発見する

３．学会等名
第44回日本分子生物学会年会

４．発表年
2021年

１．発表者名
Sato T, Yada T

２．発表標題
Prediction of human protein-coding smORFs using k-mer based machine learning

３．学会等名
2021年日本バイオインフォマティクス学会年会

４．発表年
2021年

１．発表者名
Keisuke Ando, Tetsushi Yada

２．発表標題
Does existance of intron increase birth rate of de novo gene?

３．学会等名
2019年日本バイオインフォマティクス学会年会・第8回生命医薬情報学連合大会（IIBMP2019）

４．発表年
2019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 |
|---|---|---|---|

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 |
| --- | --- |