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　 ６ 月 ２４ 日現在

機関番号：３２６８９
研究種目：挑戦的研究（萌芽）
研究期間：2021～2023
課題番号：２１Ｋ１９８３１

研究課題名（和文）アミノ酸種が限定されていた生命の共通祖先以前のタンパク質の配列推定法の開発と評価

研究課題名（英文）Sets of limited amino acid species for reconstruction of protein sequences before the last common ancestor

研究代表者
木賀　大介（Kiga, Daisuke）

早稲田大学・理工学術院・教授

研究者番号：３０３７６５８７

交付決定額（研究期間全体）：（直接経費）　　4,900,000円

研究成果の概要（和文）：本研究では、生命の共通祖先以前のアミノ酸セットで機能するタンパク質について、祖先タンパク質の配列を情報学的手法で推定し、また、このようなタンパク質を生物実験で評価した。古典的な祖先配列推定法と、残基間相互作用を考慮した変分オートエンコーダーを組み合わせることで、より活性の高い祖先配列を特定することが可能であることが示された。さらに、16種のアミノ酸で構成されたタンパク質の進化を模倣する実験も行い、新しい活性を持つ変異体を得る可能性を示した。この研究は、タンパク質の起源を理解するための新たな情報科学的手法の導入だけでなく、産業応用可能なタンパク質を創出する基盤ともなる。

研究成果の学術的意義や社会的意義
この研究の学術的意義は、進化の初期段階でのタンパク質の機能形成を解明することにある。従来の祖先配列推定法の限界を超え、新たな情報科学的手法を導入することで、生命の起源に関する理論を進展させた。また、機械学習を生命起源研究に応用することで、共通祖先の壁を越えて、進化の過程でどのようにしてタンパク質の活性が向上していったかを理解する新たな視点を提供した。社会的には、これらの知見は医学や生物工学の分野において、より効率的で安全なタンパク質設計や新薬の開発に貢献する可能性が高い。

研究成果の概要（英文）：In this study, ancestral protein sequences were inferred by informatics methods for proteins that function on a set of amino acids before the last common ancestor of life, and such proteins were also evaluated in biological experiments. It was shown that combining classical ancestral sequence reconstruction methods and variational autoencoders that consider residue-residue interactions can identify more active ancestral sequences. In addition, experiments mimicking the evolution of a protein composed of 16 different amino acids were also performed, showing the possibility of obtaining variants with new activity. This research introduces a new informatics approach to understanding the origin of proteins and provides a basis for creating proteins with industrial applications.

研究分野：合成生物学

キーワード：遺伝暗号　合成生物学　進化　祖先配列　系統樹

様　式　C－１９、F－１９－１（共通）

１．研究開始当初の背景

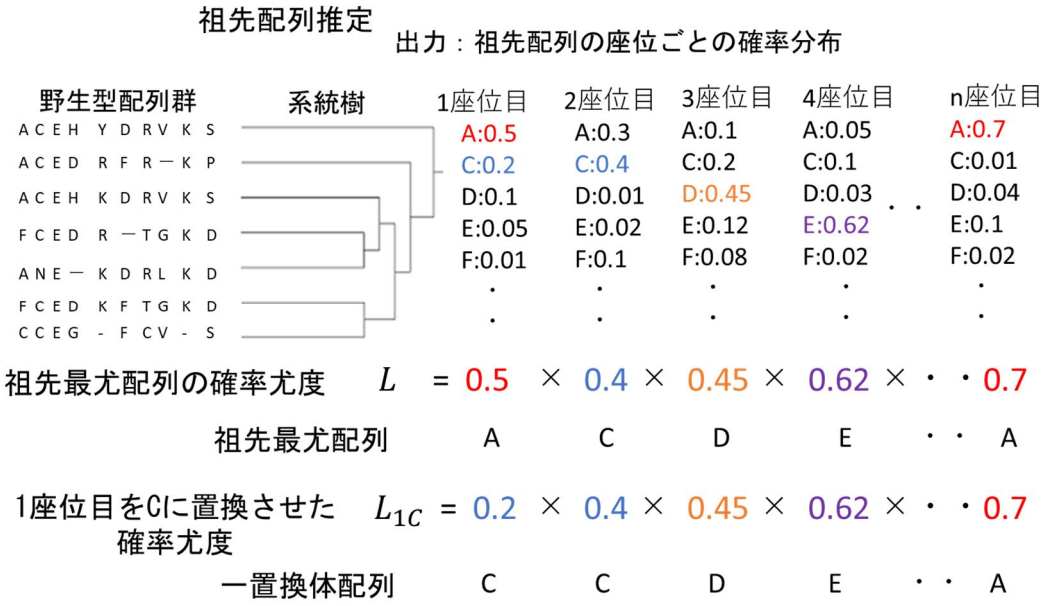
生命の共通祖先は、20 種類のアミノ酸という普遍的な「文字セット」を用いてタンパク質合成をするようになり、その子孫であるわれわれもこの共通規格を踏襲している。しかし、生命の共通祖先以前はアミノ酸の種類は 20 より少なかったと考えられている。では、そのような限定されたセットでタンパク質に機能を持たせることは可能であろうか？

２．研究の目的

本研究では、限定されたアミノ酸種しか使えなかった生命の共通祖先以前のタンパク質を検証するための、アミノ酸配列の情報学的な推定方法の検証と、得られた配列の生物実験解析を行う。また、タンパク質はアミノ酸種を限定したまま試験管内で再進化させる。これらの結果として、生命が遺伝暗号システムにアミノ酸をとりこんでいった順番に対する、新たな考察を提供することが、本研究の目的である。

３．研究の方法

祖先型タンパク質の復元は、生命の起源研究において、重要な位置を占めている。しかし、この復元は、現在の生物のアミノ酸配列群から求められる過去の最尤祖先候補配列とともに、ほぼ同等の尤度を持つ多くの祖先候補配列を示してしまう（下図）。その中で、最尤配列のみ、または数個の配列を遺伝子合成しているに過ぎない。その結果、合成した配列に活性が無いことも多々ある。

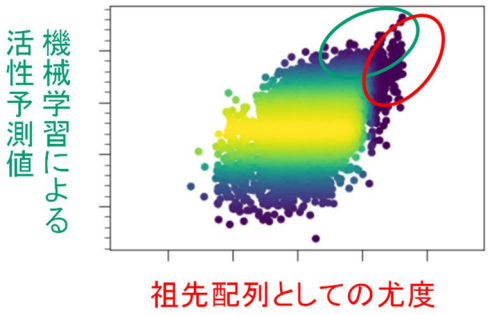


その原因は、現在主流の祖先配列推定法では、系統樹を考慮しつつも残基間相互作用を考慮しないためである。一方、タンパク質の活性は、直鎖状のポリペプチド鎖の残基間相互作用に基づいている。そこで、この相互作用を加味した予測と、一般的な祖先配列推定法と組み合わせることで、合成した祖先配列の多くが活性を持つことができるのではないか、という作業仮説に基づいて、研究を進めた。

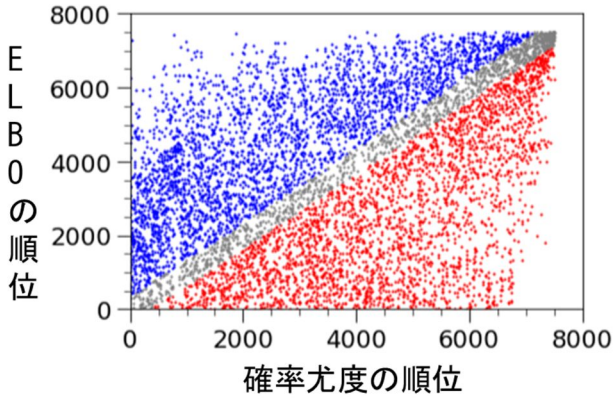
現在の生物のアミノ酸配列群を学習して、残基間相互作用を加味して任意の変異体配列の活性を予測することが、変分オートエンコーダーという深層学習によって達成されたことが報告されている。そこで、祖先配列推定に用いた同一機能を持つ各種現存生物由来のアミノ酸配列群を訓練用のデータとして学習し、アミノ酸配列に対して、機械学習による活性予測値と、祖先配列としての尤度の相関を求めた。

４．研究成果

研究の主な成果

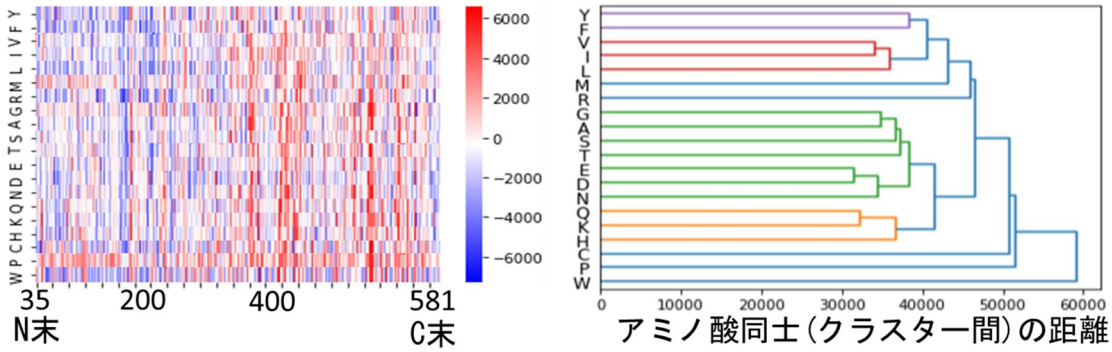


祖先配列推定として推定されたアミノ酸配列群に対して、機械学習（変分オートエンコーダー）による活性予測値と、祖先配列としての尤度の相関を求めたところ、２つの値はおおむね正の相関を示しつつ、祖先配列としての尤度が高い一方で機械学習の活性予測値が低い配列、その逆の配列、尤度も活性予測値も共に高い配列、が存在することがわかった。このことは、機械学習を古典的な祖先配列推定と組み合わせることで、祖先配列再構築について、二重のふるいをかけられることを示している。



また、変分オートエンコーダーについて、別の実装方法を試した場合でも、祖先配列としての尤度が高い一方で機械学習の活性予測値が低い配列（右図赤）、その逆の配列（右図）青、尤度も活性予測値も共に高い配列、が存在していた。

どの座位やどのアミノ酸への置換がどちらのツールに影響が出やすいのかをヒートマップグラフで可視化したところ、置換後のアミノ酸種ごとの違いが示唆された（下図左）。そこで、クラスタ解析を行った。まず、化学的性質が類似したグループ（例えばロイシン、セリン、バリン）が同一のクラスタに位置することから、アライメント時の置換行列の性質を継いでいるとはいえ、アライメントファイルを入力とする双方の予測手法の妥当性が示された。一方、20 種類のアミノ酸の中で他とは性質が大きく異なるプロリン（イミノ酸）とトリプトファン（大きな芳香環を側鎖にもつ）が、独立したグループとして判別された。これらについて、プロリンへの置換は祖先配列推定において過剰に高く評価され、トリプトファンへの置換は祖先配列推定で過剰に低く評価されることが示されている。別途の解析から、タンパク質中のアミノ酸配列の文脈を学習しつつも出現頻度を含め各種の正規分布化が行われる変分オートエンコーダーでは、プロリンやトリプトファンなど、出現頻度の低いアミノ酸への置換が高く評価されることを見出した。この結果と合わせると、タンパク質中のアミノ酸配列の文脈を考慮しない祖先配列推定では、プロリンへの置換を高く評価しすぎている可能性がある。これらのアミノ酸について、今後の予測での使用に注意を要することがわかった。



機械学習も祖先配列推定も、それぞれのアルゴリズムによる配列の評価だけでなく、そのアルゴリズムによる配列生成も行うことができる。祖先配列については、その配列を持つタンパク質を実際に合成して活性を生物実験により確認した事例が、多々報告されている。しかし、機械学習により生成された配列を持つタンパク質については、このような生物実験による活性確認の事例はまだ少ない。そこで、機械学習として上記で用いた変分オートエンコーダーによってタンパク質配列を生成し、そのタンパク質について、生物実験で活性確認を試みたところ、実際に活性を持っていた。このように、機械学習の結果の生物実験による確認事例を追加できたことは、今後の両評価方法の比較の基盤となる。

また、タンパク質について、アミノ酸種を限定したまま試験管内で再進化させることが可能であるかを確認するために、16 アミノ酸種のみによって構成されるタンパク質をコードする DNA についてランダムな変異を施し、16 アミノ酸のみを使用する単純化遺伝暗号表を用いた翻訳、

および、タンパク質の活性評価を行った。350 変異体以上の活性を調べた結果、３割以上の変異体について活性向上が見られた。これは、アミノ酸種を限定してもタンパク質の進化を進めることができる可能性が高いことを示している。

得られた成果の国内外における位置づけとインパクト

タンパク質の起源を探る研究において現在主流になっている祖先配列推定法について、別の情報科学的手法を併用することで、この分野の研究がより効率的に進展できることを示した本研究のインパクトは大きい。

今後の展望

近年進展の著しい機械学習を、タンパク質の起源を探る研究において伝統的に用いられてきた祖先配列推定を併用することで、生命の共通祖先以前のタンパク質についての研究を進められることが可能になると期待できる。そして、普遍遺伝暗号中のどのアミノ酸が遺伝暗号に最後に加わったのか、というシナリオ自体の確からしさを見積もることも可能になると期待できる。さらに、本研究で開発された、アミノ酸の種類を限定されつつも活性を保持したタンパク質を創出する方法は、工学的には、特定の使用方法に際して弱点を持たないタンパク質の創出、という意義がある。

５．主な発表論文等

〔雑誌論文〕　計2件（うち査読付論文　2件／うち国際共著　1件／うちオープンアクセス　2件）

１．著者名 Senda Naoko、Enomoto Toshihiko、Kihara Kenta、Yamashiro Naoki、Takagi Naosato、Kiga Daisuke、Nishida Hirokazu	４．巻 7
２．論文標題 Development of an expression-tunable multiple protein synthesis system in cell-free reactions using T7-promoter-variant series	５．発行年 2022年
３．雑誌名 Synthetic 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） 10.1093/synbio/ysac029	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 Beal Jacob、Telmer Cheryl A、Vignoni Alejandro、Boada Yadira、Baldwin Geoff S、Hallett Liam、Lee Taeyang、Selvarajah Vinoo、Billerbeck Sonja、Brown Bradley、Cai Guo-nan、Cai Liang、Eisenstein Edward、Kiga Daisuke、Ross David、Alperovich Nina、Sprent Noah、Thompson Jaclyn、Young Eric M、Endy Drew、Haddock-Angelli Traci	４．巻 7
２．論文標題 Multicolor plate reader fluorescence calibration	５．発行年 2022年
３．雑誌名 Synthetic 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） 10.1093/synbio/ysac010	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 該当する

〔学会発表〕　計8件（うち招待講演　4件／うち国際学会　0件）

１．発表者名 満冨健太,木賀大介
２．発表標題 配列からタンパク質の活性を推定する二つの計算手法が示す予測値の差異の検証
３．学会等名 第47回生命の起原および進化学会学術講演会
４．発表年 2023年

１．発表者名 木賀大介
２．発表標題 遺伝子組換え技術、合成生物学的手法、ゲノム合成のリスク・ベネフィット感と社会受容
３．学会等名 第96回日本細菌学会総会（招待講演）
４．発表年 2023年

１．発表者名
木賀大介

２．発表標題
合成生物学の国際学生コンテストiGEM等における教育事例

３．学会等名
科学技術社会論学会 第21回年次研究大会

４．発表年
2022年

１．発表者名
高木有隣，木賀大介

２．発表標題
Trpを含まない酵素群によって構成される解糖系に依存して生育する大腸菌の作成に向けた活性測定

３．学会等名
第60回日本生物物理学会年会（函館）

４．発表年
2022年

１．発表者名
高木有隣 橋本真奈 西田暁史 柘植謙爾 木賀大介

２．発表標題
Trpを持たない酵素群によって構成される解糖系の構築に向けた活性測定

３．学会等名
生物物理学会関東支部会

４．発表年
2022年

１．発表者名
木賀大介 宮崎和光 山村雅幸

２．発表標題
ありえた生命のかたちの設計と実装をDXする

３．学会等名
ラボオートメーション研究会（招待講演）

４．発表年
2022年

１．発表者名 木賀大介
２．発表標題 合成生物学とバイオセキュリティ
３．学会等名 バイオエコノミーの現状 セミナー（シリーズ） 応用編（招待講演）
４．発表年 2022年

１．発表者名 木賀大介
２．発表標題 合成生物学に期待される役割
３．学会等名 生命倫理学会（招待講演）
４．発表年 2021年

〔図書〕　計1件

１．著者名 四ノ宮 成祥，木賀 大介，須田 桃子，原山 優子，島薗 進	４．発行年 2022年
２．出版社 専修大学出版局	５．総ページ数 182
３．書名 合成生物学は社会に何をもたらすか	

〔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
研究協力者	満冨　健太 (Mitsutomi Kenta)		

６．研究組織（つづき）

	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
研究協力者	黄　潤一 (Hwang Yunil)		

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