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　４ 年　 ６ 月　 ６ 日現在

機関番号：３２６１２
研究種目：挑戦的研究（萌芽）
研究期間：2019～2021
課題番号：１９Ｋ２２８９７

研究課題名（和文）GenomeGAN: 敵対的生成ネットワークによるインシリコゲノム合成

研究課題名（英文）GenomeGAN: in silico genome design with generative adversarial networks

研究代表者
佐藤　健吾（Sato, Kengo）

慶應義塾大学・理工学部（矢上）・講師

研究者番号：２０３６５４７２

交付決定額（研究期間全体）：（直接経費）　　4,000,000円

研究成果の概要（和文）：特定の形質を持つゲノム配列生成を目指して，ある特定の二次構造を形成するRNA配列を設計するRNA配列設計問題に取り組んだ．深層強化学習を塩基配列空間の探索の最適化のための学習手法として用いることで，ターゲットの二次構造に対するより効率的な塩基配列の生成が可能とした．離散値である塩基配列をActivation Maximizationを用いて微分可能な表現に変換して最適化する手法をRNA配列設計問題へ応用した．シュードノット構造を含むRNA二次構造予測法IPknotを改良し，配列長に対して線形の計算量を実現した．

研究成果の学術的意義や社会的意義
合成生物学は，生命を再構成することによってその完全な理解を目指す究極のアプローチであると同時に，生物の工学的な応用に繋がることからその産業的な価 値も極めて高い．しかし，生命として完全に機能するゲノム配列を設計して，人工的な生命を合成することは困難を極める挑戦的な課題である．

研究成果の概要（英文）：In order to generate a genome sequence with specific traits, we tackled the RNA sequence design problem of designing an RNA sequence that forms specific secondary structures. By using deep reinforcement learning as a learning method for optimizing the search of sequence space, more efficient generation of sequences for the target secondary structure is achieved. The optimization method for converting discrete nucleotide sequences into a differentiable representation using Activation Maximization was applied to the RNA sequence design problem. We improved IPknot, a method for predicting RNA secondary structure including pseudoknot structures, to achieve linear computational time with respect to sequence length.

研究分野：バイオインフォマティクス

キーワード：バイオインフォマティクス　ゲノム合成　敵対的生成ネットワーク　深層強化学習　RNA配列設計　RNA二次構造

１．研究開始当初の背景

「生命の設計図」であるゲノムを設計し，それに基づいて合成あるいは改変して生物を作ることによって生命のシステム的理解を目指す学問分野を合成生物学と呼ぶ．従来の生命科学では，様々な生命現象の観測と実験を組み合わせて法則を見つけることによって生命の理解を目指してきた．しかし，このアプローチでは生命システムの推定はできても真の理解には辿り着けないと考え，生命を工学的に再構成することによって理解を目指す試みが 2000 年代前半から行われている．ヒトゲノム解読を主導した Craig Venter らは，バクテリアゲノムを人工的に合成する技術を開発し (Gibson et al., Science 2010)，生物が生きるための最小の遺伝子セットを持つ人工生命「ミニマル・セル」を合成することに世界で初めて成功した (Hutchison et al., Science 2016)．また，微生物やヒトなどのゲノムを人工的に合成して生命を実際に作り，生命のシステム的理解を目指す大規模な国際プロジェクト (Genome Project-Write) がスタートしている (Boeke et al., Science 2016).

合成生物学は，「生命とは何か？」という根本的な問いに対する追究のみでなく，生物を利用した工学的な応用に対する期待も極めて大きい．その対象は，バイオ燃料の開発，医薬品や化粧品の原料物質の生産，砂漠の緑地化，害虫の駆除など，非常に多岐にわたる．生物の工学的な応用へ向けて，これまでも人類は生物の人為的な改変を試みてきた．掛け合わせや選別などを通じて，家畜や農作物の品種改良を古くから行っており，近年では遺伝子組み換えの技術もこのような目的に利用されている．しかし，その複雑さゆえに生命システムの理解は進んでおらず，目標とする形質への改変は容易ではないことが，工学的な応用のボトルネックとなっている．

２．研究の目的

(1) 本研究は，人工知能・機械学習の最先端技術である敵対的生成ネットワーク (Generative Adversarial Networks; GAN) を合成生物学に応用し，計算機を用いたゲノム設計の全く新しい手法を開発する．具体的には，これまでに解読された全ての生物のゲノム配列を学習データとし，ゲノム配列の識別モデルを学習すると同時に，ゲノム配列のベクトル表現である「潜在ゲノム空間」からゲノム配列を生成するモデルを獲得する．さらに，潜在ゲノム空間における線形代数的な演算を利用して，狙った形質を持つゲノム配列の生成・デザインを実現する．

(2) 任意のシュードノット構造を含む RNA 二次構造予測の厳密解法は，その計算量が極めて複雑なクラスであることが証明されており，その実装は現実的ではないとされている．そのため，これまでの多くの RNA 二次構造予測法はシュードノット構造を無視した予測を行う．しかしながら，シュードノット構造は翻訳やスプライシングの制御，リボソームのフレームシフトなどに関与することが知られており，シュードノット構造を考慮した RNA 二次構造予測手法が求められている．そこで，我々は近似解法を用いて計算を高速化する IPknot を 2011 年に開発した [Sato et al. 2011]．しかし，配列長に対して 3 乗に比例する計算時間を必要とする上に，配列長が 500 塩基を超えたあたりから予測精度が著しく低下するため，mRNA やウイルス RNA のような長い配列に適用することは困難であった．そこで本研究では，IPknot を改良することによって，mRNA やウイルス RNA のような長い配列に対しても高速かつ高精度でシュードノット構造を含む RNA 二次構造を予測できる手法の開発を目指した．

３．研究の方法

(1) 敵対的生成ネットワーク（Generative Adversarial Networks; GAN）は，教師なし学習の枠組みで学習する深層ニューラルネットワークによる生成モデルである（Goodfellow et al., NIPS 2014）．GAN では，人工データを生成する生成器（Generator）と真のデータを識別する識別器（Discriminator）の 2 つのネットワークを同時に学習させる．生成器は真のデータとできるだけ類似しているデータを生成して識別器を騙すことを目指し，識別器は生成器が作った人工データと真のデータを見分けることを目指して，互いに敵対的に学習を進めていく．平衡状態に到達した時に，生成器は真のデータと区別することができない人工データを生成することが可能となり，識別器の識別精度も最高となる．また，生成器が人工的に生成する潜在ベクトル空間はデータの特徴をよく捉えた表現空間となり，線形代数的な演算が可能という性質を持つ．GAN は，画像・動画・音声の生成やドメイン適応，機械翻訳や質問応答など言語処理タスクにおける言語生成などに幅広く応用されている．本研究では，GAN を用いてゲノム配列を生成する．具体的には，以下の課題を実行する：(1)遺伝子配列を生成する GeneGAN を実装する (2)遺伝子セットを生成する GeneSetGAN を実装する (3)ゲノム配列を生成する GenomeGAN を実装する．

(2) IPknot の高速化のために，ベースとなる計算モデルに解析手法高速化の点で良い近似を実現する LinearPartition モデル[Zhang et al. 2020]を採用した．LinearPartition モデル自体はシュードノットを考慮した計算は行わないが，これに IPknot による近似解法を組み合わせることによって，配列長に対して線形の計算量でシュードノットを考慮した RNA 二次構造予測を実現した．さらに，あらかじめ予測精度を見積もる指標 pseudo-expected accuracy をシュード

ノット構造に適用し，これに基づいて配列ごとに最適なパラメータを自動的に選択する方法を開発し，これによって高精度化を実現した．

４．研究成果

(1) 特定の形質を持つゲノム配列生成を目指して，ある特定の二次構造を形成する RNA 配列を設計する RNA 配列設計問題に取り組んだ．深層強化学習を塩基配列空間の探索の最適化のための学習手法として用いることで，ターゲットの二次構造に対するより効率的な塩基配列の生成が可能になる．本研究では，二次構造の情報を明示的に扱うための構造化ニューラルネットワークを導入することによる，RNA 配列設計における予測性能への効果を調査した．構造化の表現方法としてグラフ CNN や木構造 LSTM を用いたところ，速度ではグラフ CNN が勝るものの，木構造 LSTM の方がターゲット二次構造に近い二次構造を形成する配列が得られる結果となった．深層強化学習手法を用いて生成配列の GC 含有量を制御する手法を実装した．ターゲット GC 含有量の情報を入力の表現と報酬の計算にそれぞれ組み込んだ．既存の学習モデルをベースモデルとして，対数確率から塩基を確率的に選択するための softmax 関数に，ターゲット GC 含有量についての疑似度数を組み込んだ．また，疑似度数の計算方法として，位置に依存しない各塩基についての疑似度数を与える方法と，ニューラルネットワークを用いて計算する方法を検討した．また，離散値である塩基配列を Activation Maximization を用いて微分可能な表現に変換して最適化する手法を RNA 配列設計問題へ応用した．二次構造予測によって得られた最小自由エネルギーとなる二次構造とターゲット二次構造との差異を微分可能な編集距離として算出した．

(2) 本研究で開発した IPknot++は，網羅的なベンチマークにおいて幅広い条件でシュードノットを含まない RNA 二次構造予測と同等の計算速度でありながら（図 1），シュードノットを含む配列に対しても良好な予測精度であることを示した（表 1）．

表 1　配列長ごとの他手法との F 値の比較

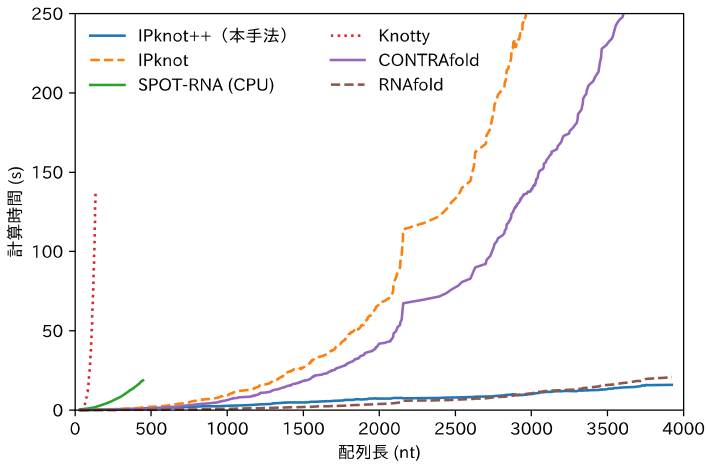

図 1　他手法との計算時間の比較

配列長	12-150 塩基		151-500 塩基		501-4381 塩基	
	PK なし	PK あり	PK なし	PK あり	PK なし	PK あり
IPknot++（本手法）	0.681	0.552	0.492	**0.482**	**0.433**	**0.428**
IPknot	0.669	0.500	0.480	0.461	0.212	0.317
Knotty	0.641	0.550	–	–	–	–
SPOT-RNA	0.658	**0.621**	0.462	0.479	–	–
CONTRAfold	**0.682**	0.519	**0.500**	0.479	0.425	0.415
RNAfold	0.668	0.472	0.474	0.442	0.361	0.347

F 値：予測した塩基対が正解構造に含まれる割合（正解率; PPV）と正解構造に含まれる塩基対を予測できた割合（網羅率; SEN）の調和平均
PK なし：シュードノットを含まない配列の F 値
PK あり：シュードノットを含む配列の F 値

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　5件）

１．著者名 Sato Kengo、Akiyama Manato、Sakakibara Yasubumi	４．巻 12
２．論文標題 RNA secondary structure prediction using deep learning with thermodynamic integration	５．発行年 2021年
３．雑誌名 Nature Communications	６．最初と最後の頁 941
掲載論文のDOI（デジタルオブジェクト識別子） 10.1038/s41467-021-21194-4	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 Taneda Akito、Sato Kengo	４．巻 22
２．論文標題 A Web Server for Designing Molecular Switches Composed of Two Interacting RNAs	５．発行年 2021年
３．雑誌名 International Journal of Molecular Sciences	６．最初と最後の頁 2720～2720
掲載論文のDOI（デジタルオブジェクト識別子） 10.3390/ijms22052720	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 Jayakumar Vasanthan、Ishii Hiromi、Seki Misato、Kumita Wakako、Inoue Takashi、Hase Sumitaka、Sato Kengo、Okano Hideyuki、Sasaki Erika、Sakakibara Yasubumi	４．巻 21
２．論文標題 An improved de novo genome assembly of the common marmoset genome yields improved contiguity and increased mapping rates of sequence data	５．発行年 2020年
３．雑誌名 BMC Genomics	６．最初と最後の頁 243
掲載論文のDOI（デジタルオブジェクト識別子） 10.1186/s12864-020-6657-2	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 Sato Kengo、Kato Yuki	４．巻 23
２．論文標題 Prediction of RNA secondary structure including pseudoknots for long sequences	５．発行年 2021年
３．雑誌名 Briefings in Bioinformatics	６．最初と最後の頁 bbab395
掲載論文のDOI（デジタルオブジェクト識別子） 10.1093/bib/bbab395	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 Kashiwagi Shunya、Sato Kengo、Sakakibara Yasubumi	4．巻 11
2．論文標題 A Max-Margin Model for Predicting Residue-Base Contacts in Protein-RNA Interactions	5．発行年 2021年
3．雑誌名 Life	6．最初と最後の頁 1135～1135
掲載論文のDOI（デジタルオブジェクト識別子） 10.3390/life11111135	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 －

〔学会発表〕　計11件（うち招待講演　1件／うち国際学会　6件）

1．発表者名 Yuki Hotta, Yasubumi Sakakibara and Kengo Sato
2．発表標題 深層強化学習を用いた二次構造に基づくRNA配列の設計
3．学会等名 第9回生命医薬情報学連合大会(IIBMP2020)
4．発表年 2020年

1．発表者名 Akiyama, M., Sato, K., Sakakibara, Y.
2．発表標題 A max-margin training of RNA secondary structure prediction integrated with the thermodynamic model
3．学会等名 Noncoding RNAs: Mechanism,Function and Therapies, Keystone Symposia（国際学会）
4．発表年 2020年

1．発表者名 Akiyama, M., Sato, K., Sakakibara, Y.
2．発表標題 A max-margin training of RNA secondary structure prediction integrated with the thermodynamic model
3．学会等名 RNA Informatics（国際学会）
4．発表年 2019年

１．発表者名
Sato, K., Akiyama, M., Sakakibara, Y.

２．発表標題
RNA secondary structure prediction using deep learning with thermodynamic integration

３．学会等名
Noncoding RNAs: Biology and Applications, Keystone Symposia（国際学会）

４．発表年
2021年

１．発表者名
Sato, K., Akiyama, M., Sakakibara, Y.

２．発表標題
RNA secondary structure prediction using deep learning with thermodynamic integration

３．学会等名
RNA meeting 2021（国際学会）

４．発表年
2021年

１．発表者名
Kato, Y., Sato, K., Havgaard, JH., Kawahara, Y.

２．発表標題
Deep learning-based prediction of potential RNA G-quadruplexes with D-Quartet

３．学会等名
The 29th Intelligent Systems for Molecular Biology and the 20th European Conference on Computational Biology (ISMB/ECCB 2021)（国際学会）

４．発表年
2021年

１．発表者名
Sato, K., Akiyama, M., Sakakibara, Y.

２．発表標題
RNA secondary structure prediction using deep learning with thermodynamic integration

３．学会等名
The 29th Intelligent Systems for Molecular Biology and the 20th European Conference on Computational Biology (ISMB/ECCB 2021)（国際学会）

４．発表年
2021年

１．発表者名
Sato, K., Akiyama, M., Sakakibara, Y.

２．発表標題
RNA secondary structure prediction using deep learning with thermodynamic integration

３．学会等名
第10回生命医薬情報学連合大会，日本バイオインフォマティクス学会2021年年会

４．発表年
2021年

１．発表者名
Kawaguchi, K., Sakakibara, Y., Sato, K.

２．発表標題
プライバシー保護技術を用いた遺伝子発現差異解析

３．学会等名
第10回生命医薬情報学連合大会，日本バイオインフォマティクス学会2021年年会

４．発表年
2021年

１．発表者名
Kengo Sato, Yuki Kato

２．発表標題
Prediction of RNA secondary structure including pseudoknots for long sequences

３．学会等名
情報処理学会第68回バイオ研究発表会

４．発表年
2021年

１．発表者名
佐藤健吾，秋山真那斗，榊原康文

２．発表標題
MXfold2: 深層学習を用いたRNA二次構造予測

３．学会等名
第44回日本分子生物学会年会（招待講演）

４．発表年
2021年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

MXfold2 Server
http://www.dna.bio.keio.ac.jp/mxfold2/

RNA二次構造予測で世界最高精度を達成
https://www.keio.ac.jp/ja/press-releases/2021/2/12/28-78076/

IPknot Server
http://rtips.dna.bio.keio.ac.jp/ipknot++/

より複雑で長いRNAの二次構造を高速に予測可能に
https://www.keio.ac.jp/ja/press-releases/2021/10/6/28-83032/

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