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　３ 年　 ５ 月 ３１ 日現在

機関番号：１２６０１
研究種目：若手研究
研究期間：2018～2020
課題番号：１８Ｋ１８１４５

研究課題名（和文）RNA 2次構造を利用した遺伝子制御のための点突然変異のデザイン

研究課題名（英文）Design of point mutations for gene regulation using RNA secondary structure

研究代表者
寺井　悟朗（Terai, Goro）

東京大学・大学院新領域創成科学研究科・特任准教授

研究者番号：４０７８５３７５

交付決定額（研究期間全体）：（直接経費）　　2,700,000円

研究成果の概要（和文）：大腸菌のmRNA配列とタンパク質発現量に関する大規模データを分析し、タンパク質発現量と高い相関を示すmRNAの２次構造的特徴を抽出した。この特徴は最小自由エネルギーなどの既存の２次構造的特徴よりも高い相関を持つことを示した。さらに、この特徴を改良するとともに、これを学習するアルゴリズムを開発した。また、発見した特徴に基づきタンパク質発現量を正または負に制御する突然変異をデザインするアルゴリズムを開発した。

研究成果の学術的意義や社会的意義
遺伝子発現を人為的に制御する技術の開発は、学術的にも産業的にも重要である。本研究では、近年技術発展が著しいゲノムへの点突然変異を利用して、原核生物の遺伝子発現をプラスミド非依存的に制御する新しい情報技術を開発した。最近、大腸菌などの原核生物では開始コドン付近のmRNA 2次構造と、遺伝子の翻訳効率が強く関連することが示された。そこで、開始コドン付近の2次構造変化を引き起こすことで、翻訳効率を正あるいは負に制御できる点突然変異をデザインするアルゴリズムを開発した。この点突然変異による遺伝子制御は、面倒なゲノム改変や、外来性の発現調節因子なしで内在性遺伝子を個別に制御できるというメリットがある。

研究成果の概要（英文）：We analyzed large-scale data on mRNA sequences and protein expression levels in E. coli and extracted secondary structural features of mRNA that are highly correlated with protein expression levels. This feature was shown to have a higher correlation than existing secondary structural features such as minimum free energy. In addition, we improved this feature and developed an algorithm to learn it. We have also developed an algorithm to design mutations that positively or negatively regulate protein expression based on the discovered feature.

研究分野：バイオインフォマティクス

キーワード：２次構造　RNA　点突然変異　原核生物　翻訳

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

遺伝子の人為的な発現制御と、その結果として起こる現象の観察は、遺伝子の機能推定のための基本的なアプローチである。また、微生物を利用した有用物質の生産においては、代謝バランスの調節を目的とした遺伝子発現制御が重要な役割を果たしている。したがって、原核生物における遺伝子制御技術の開発は、理学・工学の両面において重要なテーマである。

大腸菌では、開始コドン付近の mRNA 2 次構造が翻訳効率に大きな影響を与えることが示されている。そのメカニズムは mRNA 2 次構造がリボゾームの結合を阻害することで、翻訳効率が低下するというものである。従って、2 次構造を調節することで原核生物の遺伝子の発現を調節することができると考えられる。また、近年 CRISPR/Cas9 システムに代表されるような in vivo ゲノム編集技術が急速に発展している。特に、原核生物の点突然変異については確立した手法が存在する。従って、この点突然変異の技術を用いて遺伝子を制御する技術の開発は、実用的な観点から重要である。

２．研究の目的

本研究では、開始コドン付近の 2 次構造変化を引き起こすことで、翻訳効率を正あるいは負に制御できる点突然変異をデザインするアルゴリズムを開発する。公開された大規模データを用いて mRNA 開始コドン付近の RNA2 次構造を詳細に分析し、タンパク質発現量と相関の高い 2 次構造的特徴を抽出する。抽出した特徴に基づき、タンパク質発現量の正あるいは負に制御する突然変異をデザインするアルゴリズムを開発する。

３．研究の方法

下記の 2 つの研究項目を 3 年間で遂行した。

（１）原核生物の開始コドン付近の特徴抽出

公開された大規模データを利用して原核生物の mRNA 開始コドン付近の 2 次構造と翻訳効率の関係を詳細に分析した。2 次構造の安定性を評価するための指標は、広く使われる最小自由エネルギー(Minimum Free Energy; MFE)以外にも複数存在する。また、指標を計算するためのエネルギーモデルや確率モデルも複数存在する。そこで、いくつかの指標を異なるモデルで計算し、それぞれに対してタンパク質発現量との相関を求めた。

（２）点突然変異をデザインするアルゴリズムの開発

上記（１）で抽出した２次構造的特徴に基づき、タンパク質発現量を正、あるいは負に制御する点突然変異をデザインするアルゴリズムを開発する。2 次構造的特徴の計算は一般に時間がかかるため、ある突然変異に対する 2 次構造的特徴の変化を高速に評価するためのアルゴリズムを開発する。そして、この高速アルゴリズムを利用して効率よく目的の突然変異を探索するアルゴリズムを開発する。

４．研究成果

（１）原核生物の開始コドン付近の特徴抽出

近年、ハイスループット DNA 合成技術の進歩により、低コストで何万もの短い（約 230nt 程度の）DNA を同時に合成できるようになった。この技術を用いてさまざまな mRNA 配列について大量のタンパク質量データを測定する研究が、これまでにいくつか報告されている。しかしながら、２次構造に関する特徴を評価するための指標については大きな変化がなく、MFE などの古く使われている指標が使われ続けている。そこで、本研究では 2 つの大規模なタンパク質発現量データを用いて有用な２次構造の特徴を明らかにした。1 つ目のデータセット(以下 D1)は Cambray ら[文献 1]により得られたもので、244,000 個のデータを含んでいる。2 つ目のデータセット(以下 D2)は Goodman ら[文献 2]により得られたもので、10,000 個以上のデータを含んでいる。本研究では、これら 2 つのデータセットを用いて、表 1 に示す 6 種類の 2 次構造的特徴を評価した。accT、mfeT、ensT は Turner のエネルギーモデルで計算した特徴、accC、mfeC、ensC は CONTRAfold モデルで計算した特徴である。mRNA の開始コドン付近の配列を用いて、これら 6 種類の 2 次構造的特徴を計算し、それらがタンパク質発現量とどの程度相関するかを調べた。

表1．2次構造的特徴のリスト

accT	Turnerモデルで計算したアクセサビリティー
accC	CONTRAfoldモデルで計算したアクセサビリティー
mfeT	Turnerモデルで計算した最小自由エネルギー
mfeC	CONTRAfoldモデルで計算したViterbiスコア
ensT	Turnerモデルで計算したアンサンブル自由エネルギー
ensC	CONTRAfoldモデルで計算した分配関数のlog

・データセット１(D1)に対する評価

D1 には 24,4000 種類の mRNA 配列と、それぞれの mRNA から翻訳されるタンパク質の発現量が含まれている。全ての mRNA に対して表１に示した 6 種類の 2 次構造的特徴を計算し、タンパク質発現量との相関係数を計算した。その結果を表 2 に示す。全体的には CONTRAfold モデルで計算した特徴が、タンパク質発現量との相関係数が高かった。CONTRAfold モデルで計算した３つの特徴（accC，mfeC，ensC）の中では accC が最も相関係数が高いことがわかった。また、Turner モデルで計算した３つの特徴（accT，mfeT，ensT）は、同程度の相関係数(0.55-0.58)を示した。図１は、6 つの 2 次構造的特徴のから 2 つを取り出し、その 2 つの特徴間の相関係数を示したものである。mfeT、mfeC、ensT、ensC は互いに相関が高かった。accT と accC の間には高い相関関係があるが、タンパク質発現量との相関は accT よりも accC の方が高くなることがわかった。

表2．データセット１における各特徴の評価結果

特徴	スピアマンの相関係数
accT	0.575
accC	0.709
mfeT	0.554
mfeC	0.605
ensT	0.561
ensC	0.632

[文献 3]より改変

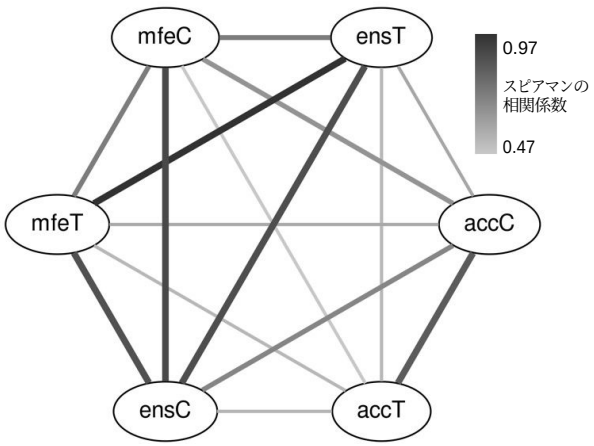

図 1．特徴間の相関 [文献 3]より改変

・データセット 2(D2)に対する評価

D2 に含まれる mRNA は、2 つの異なるプロモーターから転写されており、3 つの異なる 5'非翻訳領域（5'UTR）を持っている。プロモーターや 5'UTR のタンパク質発現量への影響を取り除くため、本研究では D2 を 6 つのグループに分けてから 2 次構造的特徴の評価に用いた（それぞれのグループに含まれる mRNA は同じプロモーターと 5'UTR を持っている）。表 3 は、各グループに対して、６つの特徴とタンパク質発現量の相関係数を計算した結果である。LW グループ以外の 5 つのグループで、accC の相関係数が最も高かった。LW グループに含まれる mRNA は、低活性プロモーターから転写されており、弱い 5'UTR を持っている。また、D2 を測定した文献[文献 2]によると、LW に含まれるタンパク質発現量の 96%は検出限界を下回っていた。したがって、LW グループのタンパク質発現量に関する信頼性は低いと考えられる。これらの結果から、D2 においても accC が最もタンパク質発現量との相関が高くなると結論した。

2 つの大規模データを用いた評価により、CONTRAfold モデルで計算した２次構造的特徴が、Terner モデルで算出した特徴よりもタンパク質発現量との相関が高くなることがわかった。CONTRAfold モデルは Turner モデルよりも RNA 2 次構造を正確に予測できるこが知られている。より正確な 2 次構造予測がタンパク質発現量との高い相関に寄与したと考えられる。また、CONTRAfold モデルで計算した 3 つの特徴量(accC，mfeC，ensC)の中では、accC がタンパク質発現量と最もよく相関することがわかった。accC は mfeC や ensC よりも広い範囲の 2 次構造を考慮することができる。このことがより正確な 2 次構造予測につながり、タンパク質発現量との高い相関に寄与したと考えられる。

また、本研究では accC をさらに改良することを試みた。具体的には、開始コドンからの相対位置ごとに異なる重みをつけた改良型のアクセサビリティーを定義し、その重みをデータから学習する方法を開発した。この改良型のアクセサビリティーを用いることにより開始コドン付近でどの位置が相対的に重要なのかを知ることができる。また、本研究ではアクセサビリティーを基準として点突然変異をデザインするアルゴリズムの開発を行なったが（下記を参照)、この改良型のアクセサビリティーは以下で開発したデザインアルゴリズムに組み込むことができる。

表3. データセット2における各特徴の評価結果

特徴	スピアマンの相関係数					
	HW	HM	HS	LW	LM	LS
accT	0.693	0.641	0.552	0.394	0.635	0.658
accC	**0.787**	**0.738**	**0.611**	0.440	**0.746**	**0.753**
mfeT	0.728	0.642	0.536	0.420	0.654	0.648
mfeC	0.632	0.573	0.492	0.353	0.558	0.579
ensT	0.747	0.658	0.550	0.428	0.668	0.663
ensC	0.754	0.665	0.560	**0.445**	0.681	0.681

HW, HM, HS, LW, LM, LSはグループ名を表す。各グループでもっと高い相関係数を太字で示す。[文献3]より改変

（2）点突然変異をデザインするアルゴリズムの開発

上記の研究により、アクセサビリティーがタンパク質発現量を予測するために有用な指標であることがわかった。そこで、アクセサビリティー値を大きく上昇あるいは低下させるような点突然変異を求めるアルゴリズムを開発した。アクセサビリティーの計算には、$O(LW^2)$の計算量が必要である(L は mRNA 配列長、W は塩基対の最大距離に関する制約)。本研究では、mRNA 配列に点突然変異を入れた時のアクセサビリティー値を高速に計算するアルゴリズムを開発した。このアルゴリズムは、Rchange アルゴリズム[文献 4]をアクセサビリティーの計算に応用したものである。このアルゴリズムを使うと $O(LW^2)$の計算量を $O(W^2)$に抑えることができるため、より多くの突然変異を評価することが可能となる。

図 2 は大腸菌の全内在性遺伝子に対して、突然変異をデザインした時の平均的な振る舞いを示している。X 軸の k は評価する突然変異の数に関するパラメータであり、値が大きいほど多くの突然変異を評価する（その分時間がかかる)。図 2(a)はアクセサビリティーを下げる（タンパク質発現量を下げる）場合、(b)はアクセサビリティーを上げる（タンパク質発現量を上げる）場合の、アクセサビリティー値の平均変化量を示している。大腸菌の内在性遺伝子はアクセシビリティーを正方向よりも負方向に変化させやすい傾向が見られた。これは大腸菌の内在性遺伝子はもともとアクセサビリティーが高い傾向があるため、さらにアクセサビリティーを上げるのが難しかったからだと考えられる。また、この結果から、パラメータ k は 100 程度までは効果が大きいが、それ以上大きくしても効果が限られることがわかった。

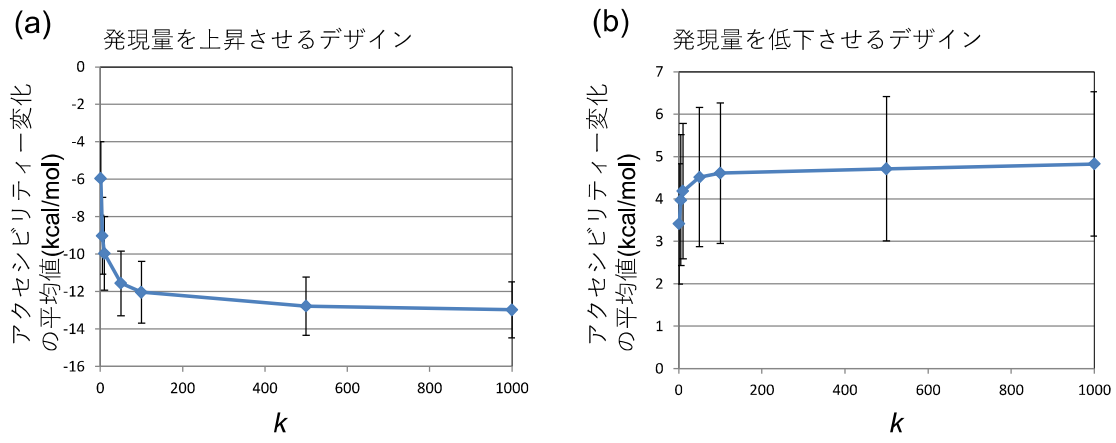

図 2. デザインした突然変異によるアクセシビリティーの変化（k は設計のパラメータ）

・まとめ

本研究では、公開された大規模データを用いて、6 種類の 2 次構造的特徴とタンパク質発現量の相関を調べた。その結果、CONTRAfold モデルで計算したアクセサビリティーが最も高い相関を示すことを見出した。そこで、アクセサビリティーに基づき、タンパク質発現量を正あるいは負に制御する点突然変異をデザインするアルゴリズムを開発した。mRNA の 2 次構造形成は物理化学的現象である事から、本研究で開発した技法は大腸菌以外の原核生物にも幅広く適用可能であると期待される。

[参考文献]

1) G. Cambray, J. C. Guimaraes, A. P. Arkin, Evaluation of 244,000 synthetic sequences reveals design principles to optimize translation in Escherichia coli. *Nat. Biotechnol.* **36**, 1005-1015 (2018).
2) D. B. Goodman, G. M. Church, S. Kosuri, Causes and Effects of N-Terminal Codon Bias in Bacterial Genes. *Science* **342**, 475-479 (2013).
3) G. Terai, K. Asai, Improving the prediction accuracy of protein abundance in Escherichia coli using mRNA accessibility. *Nucleic Acids Res.* **48**, e81-e81 (2020).
4) H. Kiryu, K. Asai, Rchange: algorithms for computing energy changes of RNA secondary structures in response to base mutations. *Bioinformatics* **28**, 1093-101 (2012).

5．主な発表論文等

〔雑誌論文〕　計1件（うち査読付論文　1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 Goro Terai, Kiyoshi Asai	4．巻 48
2．論文標題 Improving the prediction accuracy of protein abundance in Escherichia coli using mRNA accessibility	5．発行年 2020年
3．雑誌名 Nucleic Acids Research	6．最初と最後の頁 e81
掲載論文のDOI（デジタルオブジェクト識別子） 10.1093/nar/gkaa481	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2件（うち招待講演　0件／うち国際学会　0件）

1．発表者名 寺井悟朗、浅井潔
2．発表標題 大腸菌のタンパク質発現量予測のためのRNA2次構造特徴の抽出
3．学会等名 第43回日本分子生物学会
4．発表年 2020年

1．発表者名 Goro Terai, Kiyoshi Asai
2．発表標題 Discovery of mRNA secondary structural features for the accurate prediction of protein abundance in Escherichia coli
3．学会等名 2020年日本バイオインフォマティクス学会年会
4．発表年 2020年

〔図書〕　計0件

〔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

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