R-STEINER: mRNA 高翻訳化のための 5'UTR 生成手法

田中　宏昌† 鈴木　優† 山崎将太朗†† 吉野幸一郎† 加藤　晃††
中村　哲†

† 奈良先端科学技術大学院大学 情報科学研究科　〒 630-0101 奈良県生駒市高山町 8916 番地の 5
†† 奈良先端科学技術大学院大学 バイオサイエンス研究科　〒 630-0101 奈良県生駒市高山町 8916 番地の 5
E-mail: †{tanaka.hiroaki.sy2,ysuzuki,koichiro,s-nakamura}@is.naist.jp, ††{s-yamasaki,kou}@bs.naist.jp

あらまし　mRNA から翻訳されるタンパク質の量—翻訳量—を増加させる事は，バイオサイエンスの主要な課題の一つである．特定遺伝子の翻訳量を増加させる方法の一つとして，5'UTR(mRNA の一部) を改変するという手法がある (特定遺伝子の翻訳量増加を誘発するような 5'UTR 配列を翻訳エンハンサーと呼ぶ)．しかし，翻訳エンハンサーを発見するための実験には多大な費用・時間・労力が必要である．この課題を解決するためには，現実の実験を計算機上で擬似的に行うことで翻訳エンハンサーを発見出来れば良いと考えた．本稿では R-STEINER (generate nucleotide sequence Randomly and Select a TrEmendous 5'-untranslated region which INcrEase the amount of tRaslated proteins of a certain gene) を提案する．R-STEINER では mRNA の翻訳量予測モデルを作成し，そのモデルを用いて特定遺伝子の翻訳エンハンサーを発見する．本手法により，現実の実験を行うことなく翻訳エンハンサーを発見することが可能となり，費用・時間・労力の削減が可能となる．本稿ではイネの mRNA を用いて R-STEINER の評価を行った．その結果，R-STEINER によって生成された 5'UTR の翻訳量の予測値は実測値と高い相関 (相関係数 0.89) を示しており，R-STEINER によって実際に翻訳エンハンサーを生成出来ることを示した．

キーワード　機械学習，ランダムフォレスト，勾配ブースティング，XGBoost，mRNA，翻訳量，タンパク質合成，mRNA，遺伝子発現，バイオインフォマティクス

1. はじめに

植物や植物の培養細胞を用いたタンパク質生産が近年注目を集めている [26], [27]．タンパク質を生産する際の宿主にはいくつかの選択肢があり，宿主によって一長一短がある．ヒトへの使用を前提としたワクチンを開発する場合を考えると，その安全性が重要である．ヒトの培養細胞でワクチンを開発すると，ヒト遺伝子に対して有害な操作を行う物質がワクチンに混入する可能性があるのに対し，植物でワクチンを開発すれば，そのような可能性は限りなく低くなる [23]．その他にも植物を用いたタンパク質生産には様々な利点があり，Jian ら [23] にまとめられている．

しかし，植物を宿主とした組換えタンパク質の発現量は，他の宿主よりも低いことが課題となっている [24], [27]．この課題を解決する方法の一つとして，天然の 5'UTR (mRNA の一部) 配列の代わりに，翻訳量を増加させるような 5'UTR 配列—これを翻訳エンハンサーと呼ぶ—を用いる方法がある．つまり，与えられた遺伝子に対して，「結合させると発現量が向上するような 5'UTR」を使う方法である．ただし，翻訳エンハンサーを発見することは容易ではない．5'UTR 内のどのような部分列が翻訳に影響を与えるかは分かっておらず，翻訳エンハンサーを演繹的に作成することは不可能である．そのため，翻訳エンハンサー発見の為には 1) 翻訳エンハンサー候補を複数用意して 2) それらを生命体に注入して実際に翻訳量を測定する という探索的な実験が必要になる．このような探索的実験を何度も繰り返すことは，多大な費用・時間・労力が必要である．

そこで本稿では，合成実験の費用・時間・労力の削減を可能にする R-STEINER を提案する．R-STEINER では，計算機上で多数の 5'UTR を生成し，その翻訳量を予測する．これは，5'UTR の合成実験を仮想的に行っていることに相当する．したがって本手法を用いることで，実際に合成実験を行うことなく，特定遺伝子の翻訳エンハンサーを生成することが可能となる．すなわち，翻訳量を向上させる 5'UTR を発見するための合成実験にかかる費用・時間・労力を削減することが可能となる．本稿ではイネの遺伝子を用いて，R-STEINER によって 5'UTR の生成を行った．

R-STEINER は B-step (building step) と G-step (generating step) の 2 つの段階に分けられる．B-step では mRNA の翻訳量予測モデルを作成する．ここで作成した予測モデルを用いて，G-step で 5'UTR の生成・選択を行う．G-step では B-step で作成した予測モデルを評価関数として，特定遺伝子の翻訳エンハンサーを選択する．R-STEINER の評価に関して，我々は 2 つのことについて議論しなければならない．一つ目は，予測モデルの作り方である．これについては予備実験を行って決定している (XXX 節)．二つ目は，予測モデルが人工 mRNA に対しても高い精度で翻訳量を予測できるかという点である．予測モデルを構築する際には天然のイネのデータを学習データとして使用している．そのため，人工 mRNA のデータに対して高い精度で翻訳量を予測できるとは限らない．この点に関する

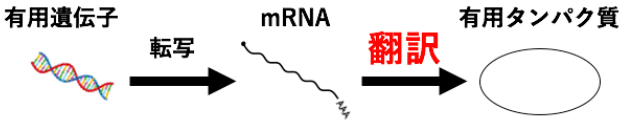

図 1　DNA からタンパク質が作られるまで

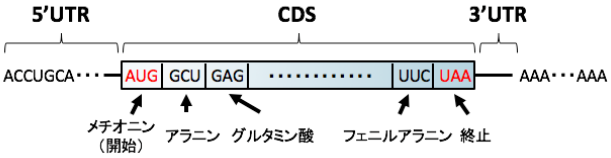

図 2　mRNA の構造

評価のため，R-STEINER と同様の手順で生成した 5'UTR を実際に合成して翻訳量を実測し，予測モデルが人工 mRNA に対しても翻訳量を高い精度で予測できていることを確認した．

2.　基本的事項

本節では，本稿を理解するために必要な基本的事項と関連研究を紹介する．mRNA からタンパク質が作られる過程を翻訳とよぶ（図 1）．mRNA は四種類のヌクレオチド A, U, G, C が連なって構成されており，5'UTR (5'-untranslated region), CDS (cording DNA sequence), 3'UTR (3'-untranslated region) と呼ばれる 3 つの領域に分けられる（図 2）．リボソームが mRNA を 5'UTR 側から読み取り，タンパク質を合成する．合成されるタンパク質がどのようなものになるかは CDS によって決まる．CDS では 3 つの連続したヌクレオチドに対応して一つのアミノ酸が生成される（図 2）．これらのアミノ酸がリボソーム内で重合され，タンパク質が合成される．5'UTR および 3'UTR は実際にはタンパク質に翻訳されないが，特に 5'UTR が遺伝子の翻訳量に影響を及ぼしていることが報告されている [18], [8], [10]．翻訳に関するより詳細な事項は，文献 [1] を参照されたい．

mRNA の翻訳量を測定する指標には PR (Polysome Ratio) 値 [25] を用いた．活発に翻訳が行われている mRNA には，多数のリボソームが重合してポリソームが形成される．PR 値は，特定の mRNA がどの程度のポリソームを形成しているかの比率で定義される．

mRNA はその一部が結合し合うことで，2 次構造と呼ばれる複雑な構造をとっている（図 3）．ヌクレオチドの配列から，とり得る立体構造と，その時の自由エネルギーを予測することが可能である．自由エネルギーは，立体構造が強固であるほど値が小さくなる評価値であり，本稿では ViennaRNA Package [9] を用いて自由エネルギーを推定した．

3.　関 連 研 究

Kawaguchi [8] らは ribosome loading と 5'UTR, CDS, 3'UTR の長さ，および A, U, G, C, AU, GC, CU, AG, GU の含有量との関係を解析した．ribosome loading は mRNA の

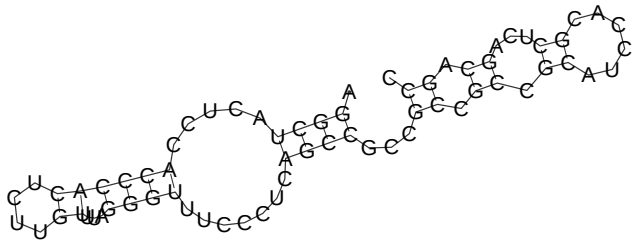

図 3　mRNA の 2 次構造

翻訳量を図る指標の一つである．しかし，Kawaguchi ら [8] が行った解析は ribosome loading と単変量との解析であるから，ribosome loading が複数の変数に影響を与えられような関係を捉えきることはできない．

複数の特徴量と翻訳の関係を捉えようとした研究として，Matsuura ら [10] の研究が挙げられる．Matsuura ら [10] は PLS モデルを用いて relative F-luc activity と呼ばれる指標の予測モデルを構築した．ただし，relative F-lucactivity は翻訳量を表す指標ではなく，熱ストレスが翻訳に与える影響を評価するための指標である．すなわち，本研究とは予測モデルの目的変数が異なっている．しかし，mRNA 配列から計算される特徴量と翻訳に関する指標の予測モデルを構築した例としてはおそらく唯一の関連研究であり，本研究でも比較手法として PLS モデルを用いた．

4.　R-STEINER

本節では，与えられた特定遺伝子の翻訳エンハンサーを発見するための手法である R-STEINER (generate nucleotide sequences Randomly and Select a TrEmendous 5'-untranslated retion that Increase the amount of traNslated protEins of a ceRtain gene) を提案する．R-STEINER は，B-step と G-step の 2 つのステップから構成されている．B-step では PR 値の予測モデルを作成する．G-step では与えられた遺伝子の翻訳量を最も向上させた k 個の 5'UTR 配列を選択して，翻訳エンハンサーとして出力する．B-step と G-step の詳細はそれぞれ 4.1 節と 4.2 節に記す．

4.1　B-step

B-step は以下の二つのステップから構成される：

(B1)　特徴量設計,

(B2)　予測モデルの構築.

表 1　Con でのデータセット ($N_{\mathrm{Con}} = 24915$)

Gene ID	5'UTR	CDS	3'UTR	PR-value
1	GUU...GAG	AUGU...AUGA	UGA...UGC	0.9229
2	GAA...UAU	AUGA...GUAA	GAG...GUC	1.0054
⋮	⋮	⋮	⋮	⋮
N_{Con}	$s^{5'\mathrm{UTR}}_{N_{\mathrm{Con}}}$	$s^{\mathrm{CDS}}_{N_{\mathrm{Con}}}$	$s^{3'\mathrm{UTR}}_{N_{\mathrm{Con}}}$	$y_{N_{\mathrm{Con}}}$

表 2　HS でのデータセット ($N_{\mathrm{HS}} = 21786$)

Gene ID	5'UTR	CDS	3'UTR	PR-value
1	GAA...UAU	AUGA...GUAA	GAG...GUC	1.1293
2	AGG...GCC	AUGG...UUGA	GUG...UUC	0.7600
⋮	⋮	⋮	⋮	⋮
N_{HS}	$s^{5'\mathrm{UTR}}_{N_{\mathrm{HS}}}$	$s^{\mathrm{CDS}}_{N_{\mathrm{HS}}}$	$s^{3'\mathrm{UTR}}_{N_{\mathrm{HS}}}$	$y_{N_{\mathrm{HS}}}$

(B1) では mRNA 配列から特徴ベクトルを設計する (4.1.1 節)．(B2) では，ランダムフォレスト [2]，勾配ブースティング [4]，XGBoost [3] を使って PR 値の予測モデルを構築する (4.1.2 節)．

4.1.1　(B1) 特徴量設計

PR 値予測モデルのための特徴量を設計する．用いる特徴量は次の 3 種類である．

(F1)　5'UTR, CDS, 3'UTR それぞれの配列長

(F2)　5'UTR, CDS, 3'UTR それぞれの自由エネルギー

(F3)　5'UTR, CDS, 3'UTR それぞれにおける A, U, G, C, AA, AU, ..., CC, AAA, AAU, ..., CCC のカウンタ

(F1) と (F2) は，mRNA の翻訳量に影響をおよぼすことが知られている [8]．これらの特徴量から特徴ベクトルを式 (1) のように構成する．

$$\boldsymbol{x} = \mathrm{concat}\left[\boldsymbol{x}_{F_1} \quad \boldsymbol{x}_{F_2} \quad \boldsymbol{x}_{F_3}\right] \in \mathbb{R}^{238} \tag{1}$$

ただし，式 (1) において

$$\boldsymbol{x}_{F_1} = \left[\mathrm{len}(5'\mathrm{UTR}) \quad \mathrm{len}(\mathrm{CDS}) \quad \mathrm{len}(3'\mathrm{UTR})\right] \in \mathbb{R}^3,$$
$$\boldsymbol{x}_{F_2} = \left[G(5'\mathrm{UTR}) \quad G(\mathrm{CDS}) \quad G(3'\mathrm{UTR})\right] \in \mathbb{R}^3,$$
$$\boldsymbol{x}_{F_3} = \mathrm{concat}\left[\boldsymbol{c}^{5'\mathrm{UTR}} \quad \boldsymbol{c}^{\mathrm{CDS}} \quad \boldsymbol{c}^{3'\mathrm{UTR}}\right] \in \mathbb{R}^{232}$$

であり，$\mathrm{len}(R)$ は領域 R の長さ，$G(R)$ は領域 R の自由エネルギーを表す．また，$\boldsymbol{c}^{5'\mathrm{UTR}}, \boldsymbol{c}^{3'\mathrm{UTR}} \in \mathbb{R}^{84}$ はそれぞれ 5'UTR, 3'UTR における A, U, G, C, AA, AU, ..., UU, AAA, AAU, ..., UUU のカウンタを要素に持つベクトルであり，$\boldsymbol{c}^{\mathrm{CDS}} \in \mathbb{R}^{64}$ は AAA, AAU, ..., UUU のカウンタを要素に持つベクトルである．CDS では 3 つの連続したヌクレオチドが 1 つのアミノ酸に対応することが分かっている．したがって，ヌクレオチドの 1-gram, 2-gram のカウンタは特徴量に用いないことにした．

4.1.2　(B2) 予測モデル

予測モデルの作成に用いたデータセットは 2 種類ある．一つ目は自然な状態（Con）で採取されたイネのデータセット，二つ目は熱ストレス下（HS）で採取されたイネのデータセットである．これらのデータセットの概要を表 1, 表 2 に示す．Gene ID は mRNA の ID，PR-value は mRNA に対応する PR 値，$s^{5'\mathrm{UTR}}_{N_{\mathrm{Con}}}$ は Con における 5'UTR 配列を表しており，表内における各領域の配列長は一般的に異なっている．

PR 値の予測モデルを，ランダムフォレスト・勾配ブースティング・XGBoost のアンサンブルモデルとして構築する．すなわち，新たに与えられた mRNA の PR 値を 式 (2) で推定する．

$$h^{(\mathrm{HS})}(\boldsymbol{x}^*) = \frac{1}{3}\left(h^{(\mathrm{HS})}_{\mathrm{rf}}(\boldsymbol{x}^*) + h^{(\mathrm{HS})}_{\mathrm{gb}}(\boldsymbol{x}^*) + h^{(\mathrm{HS})}_{\mathrm{xgb}}(\boldsymbol{x}^*)\right),$$
$$h^{(\mathrm{Con})}(\boldsymbol{x}^*) = \frac{1}{3}\left(h^{(\mathrm{Con})}_{\mathrm{rf}}(\boldsymbol{x}^*) + h^{(\mathrm{Con})}_{\mathrm{gb}}(\boldsymbol{x}^*) + h^{(\mathrm{Con})}_{\mathrm{xgb}}(\boldsymbol{x}^*)\right),$$
$$h(\boldsymbol{x}^*) = \frac{1}{2}\left(\hat{h}^{(\mathrm{HS})}(\boldsymbol{x}^*) + \hat{h}^{(\mathrm{Con})}(\boldsymbol{x}^*)\right). \tag{2}$$

ここで，$\boldsymbol{x}^*$ は新たに与えられた mRNA 配列の特徴ベクトル，$h^{(\mathrm{HS})}_{\mathrm{rf}}(\cdot), h^{(\mathrm{HS})}_{\mathrm{gb}}(\cdot), h^{(\mathrm{HS})}_{\mathrm{xgb}}(\cdot)$ はそれぞれ，HS のデータを使ってランダムフォレスト，勾配ブースティング，XGBoost によって構築された予測モデルで，$h^{(\mathrm{Con})}_{\mathrm{rf}}(\cdot), h^{(\mathrm{Con})}_{\mathrm{gb}}(\cdot), h^{(\mathrm{Con})}_{\mathrm{xgb}}(\cdot)$ は Con のデータを使って構築された予測モデルである．HS, Con それぞれの状態での各予測モデルの学習に関する詳細は 5.1 節–5.2 節に記した．

4.2　G-step

G-step では，翻訳エンハンサーを発見する．手順の概要は以下のようになっている．

(G1)　ℓ 個のヌクレオチド A, U, G, C をランダムに生成し，それらを連結して 5'UTR 配列を計算機上で作成する．

(G2)　生成した 5'UTR 配列を，特定遺伝子に対応する CDS, 3'UTR に結合させて mRNA 配列を作成し，特徴ベクトルを作成する．

(G3)　B-step で作成した予測モデルを用いて各 mRNA 配列の PR 値を推定し，PR 値が高いものから順に k 個を選択して出力とする．

ただし，(G1) において部分列 AUG, AAUAAU は使用できないとする．これは，AUG, AAUAAU が存在する部分で 5'UTR が切れてしまうためである．このような状況下で翻訳エンハンサーを発見するアルゴリズムを Algorithm 1 に示す．

5.　予 備 実 験

PR 値の予測モデルを，ランダムフォレスト・勾配ブースティング・XGBoost のアンサンブル学習器によって構築する．これら三つの学習器はいずれも回帰木ベースの加法的モデルで，様々な分野で使用されている [17], [20], [21] ．さらに，4.1.1 節節で設計した特徴ベクトルは離散型の変数を多数含んでいる．このような特徴ベクトルに対して予測を行う場合，先述のような回帰木ベースの予測器が，他の主要な手法よりも高い精度を発揮すると期待できる [7].

実際に，回帰木ベースの予測器が他の手法よりも優れていることを示すために，予測モデルの精度比較を行った．比較手法には PLS モデル [22]・線形 Lasso [5], [6], [19]・多層ニューラルネットワークを用いた．

翻訳量予測モデルを作成する際には，表 1, 表 2 の 50% を学

Algorithm 1 Algorithm of mRNA Generation

Require: $\hat{h}(\cdot)$: (2) によって構築された PR 値予測モデル，特定の CDS 配列, 3'UTR 配列

Ensure: k 個の 5'UTR 配列 $S^{5'\mathrm{UTR}}$

1: 特徴ベクトルの一部である $\boldsymbol{x}_{F_2}$ と $\boldsymbol{x}_{F3}$ を作成
2: **for** $t = 1, 2, \cdots, B$ **do**
3: 5'UTR の配列長 $L \in \mathbb{N}$ を区間 $(22, 49)$ からランダムに固定
4: L 個のヌクレオチド $\{ s_\ell \}_{\ell=1}^{L}$ をランダムに生成
5: $S_t^{5'\mathrm{UTR}} \leftarrow \mathrm{concat}\{ s_\ell \}_{\ell=1}^{L}$
6: 特徴ベクトルの一部である $\boldsymbol{x}_{F_3}$ を $S_t^{5\mathrm{utr}}$ から作成
7: $\boldsymbol{x}_{F_1}, \boldsymbol{x}_{F_2}, \boldsymbol{x}_{F_3}$ を結合して，特徴ベクトル $\boldsymbol{x}_t^*$ を作成
8: 現在のステップ t における 5'UTR 配列 $S_t^{5'\mathrm{UTR}}$ の PR 値を

$$\hat{y}_t = \hat{h}\left(\boldsymbol{x}_t^*\right)$$

で推定する
9: **end for**
10: **return** $\{ S_t^{5'\mathrm{UTR}} \}_{t=1}^{T}$ の中から，推定された PR 値が高いものを上から順に k 個

習セットとし，25% を開発セット，25% をテストセットとした．学習セットで学習器を学習させ，開発セットでハイパーパラメータのチューニングを行った（**??**節）．

5.1 ハイパーパラメータのチューニング

各学習器にはいくつかのハイパーパラメータが存在している．PLS モデルでは，主軸の本数 η がハイパーパラメータで，線形 Lasso では正則化パラメータ α がハイパーパラメータである．ニューラルネットワークには層の数，各層におけるユニット数や，学習率・ドロップアウト率・最適化手法などの様々なハイパーパラメータがある．一般的に，ニューラルネットワークの構造は該当分野の先行研究を参考にして構築するが，我々が探した限り，本研究に関する先行研究の中には多層ニューラルネットワークを用いた翻訳量予測の研究は存在しなかった．そのため，本研究では最も単純な 3 層のフィードフォワードニューラルネットワークを用いた．回帰木ベースの 3 手法では，弱学習器である回帰木の本数 M，各回帰木の深さの最大値 $d_{\max}$ をハイパーパラメータとしてチューニングする．

一般的に，ハイパーパラメータ $\theta_1, \theta_2, \cdots, \theta_p$ の決定に関しては開発セットでの経験損失を最小化するように $\boldsymbol{\theta} = [\theta_1, \theta_2, \cdots, \theta_p]$ を決定する；

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \frac{1}{N_{\mathrm{vali}}} \sum_{n=1}^{N_{\mathrm{vali}}} l\left(h(\boldsymbol{x}_n; \boldsymbol{\theta}), y_n\right). \qquad (3)$$

ただし，式 (3) において N_{vali} は開発セットのサンプル数，$l(\cdot, \cdot)$ は損失関数，$h(\boldsymbol{x}_n)$ を $\boldsymbol{x}_n$ に対する予測値，y_n を $\boldsymbol{x}_n$ に対応する目的変数とし，開発セットに含まれる標本に対して損失の和をとる．損失関数 l を $\boldsymbol{\theta}$ の関数と見たとき，回帰木ベースの学習器では l を陽に書くことは出来ない．すなわち，解析的に $\hat{\boldsymbol{\theta}}$ を求めることは出来ない．この問題を解決するため，ハイパーパラメータの最適化手法にはグリッドサーチ・ランダムサーチ [14], [15], [16]・ベイズ最適化 [11], [12], [13] など様々な方法が存在する．これらの手法はいずれも，与えらた探索範囲内で $\hat{\boldsymbol{\theta}}$ を探索する．そのため，探索範囲の設定の仕方によっては最適解とは全く異なった $\hat{\boldsymbol{\theta}}$ を選択する可能性がある．さらに最悪の場合，選択された $\hat{\boldsymbol{\theta}}$ は任意軸方向の極小点ですらない場合もある．そこで本研究では，ハイパーパラメータの探索範囲を以下のようにして決定した．

step 1. ハイパーパラメータの初期値を $(\theta_1, \theta_2, \cdots, \theta_p) = \left(\hat{\theta}_1, \hat{\theta}_2, \cdots, \hat{\theta}_p\right)$ とする．これらの初期値は与えられているとする．

step 2. 式 (4) のようにして θ_1 を更新する．

$$\hat{\theta}_1 = \arg\min_{\theta_1 \in I_1} \frac{1}{N_{\mathrm{vali}}} \sum_{n=1}^{N_{\mathrm{vali}}} l\left(h(\boldsymbol{x}_n), y_n\right) \qquad (4)$$

ただし，I_1 は θ_1 方向の極小点を少なくとも 1 つ含む区間とする．

step 3. $\theta_2, \cdots, \theta_p$ を同様の方法で更新する．

step 4. 区間 $\prod_{i=1}^{p} [\theta_i - \varepsilon_i, \theta_i + \varepsilon_i]$ を探索範囲として，グリッドサーチでハイパーパラメータを決定する．

このようにハイパーパラメータの探索範囲を決定することで，損失関数 l が探索範囲で凸である場合は少なくとも一つの極小点を含む．

ハイパーパラメータを一つしか持たない学習器の場合は θ_1 方向に最適な値を探索するだけでよいが，複数のハイパーパラメータを持つ学習器を使って予測モデルを構築する場合は，上記のような手順でハイパーパラメータの決定を行う．ただし，多層ニューラルネットワークに関してはベイズ最適化によってハイパーパラメータを決定した．これは，多層ニューラルネットワークは他の学習器に比べてハイパーパラメータの数が著しく多く，グリッドサーチでそれらを決定するには組み合わせが多すぎると考えたためである．ハイパーパラメータの探索を行う範囲と，決定された値を表 3，表 4 に示した．

5.2 予測モデルの評価

??節の方法でチューニングした各予測モデルの精度を図 4 に示した．図 4 では，縦軸は予測 PR 値と観測された PR 値 (すなわち，真の PR 値) の相関係数であり，1 に近いほど予測精度が良い．図 4 に示されているように，回帰木ベースの三つの予測モデルが他の予測モデルよりも良い性能を示している．また，これら三つの予測モデルの相関係数に統計学的に有意な差があるかを確認するため，

帰無仮説 H_0　$\rho_{\mathrm{rf}} = \rho_{\mathrm{gb}} = \rho_{\mathrm{xgb}}$

対立仮説 H_1　$\neg H_0$

として有意水準 5% で検定を行った．その結果，H_0 を棄却することはできなかったため，三つの予測モデルの精度に差が無い可能性がある．したがって，本研究では三つの予測モデルを全て用いて，最終的な予測モデルを構築した．

6. 合成実験と評価

R-STEINER では予測モデルを評価関数として配列選択を行っている．しかし，ここで用いられている予測モデルは天然

表 3　ニューラルネットワークのハイパーパラメータ

layer	hyperparameter	candidate area	selected (HS)	selected (Con)
input	activation	None	tanh	tanh
hidden	number of units	$\{256, 512, 1024, 2148\}$	2048	512
	drop out rate	$[0, 0.5]$	0	0.27
	activation function	None	relu	relu
hidden	number of units	$\{256, 512, 1024, 2148\}$	1024	512
	drop out rate	$[0, 0.5]$	0	0
	activation function	None	linear	linear
output	number of units	None	1	1

表 4　学習器のハイパーパラメータ

Model	hyperparameter	searched area (HS)	determined value (HS)	searched area (Con)	determined value (Con)
PLS モデル	η	$\{2, 3, \cdots, 221\}$	123	$\{2, 3, \cdots, 221\}$	90
線形 Lasso	α	$\{2^i \mid i = 1, 0, \cdots, -4\}$	2^{-4}	$\{2^i \mid i = 1, 0, \cdots, -4\}$	2^{-4}
ランダムフォレスト	M	$\{10, 20, \cdots, 100\}$	94	$\{270, 275, \cdots, 285\}$	285
	$d_{\max}$	$\{1, 6, \cdots, 31\}$	20	$\{11, 12, \cdots, 20\}$	16
勾配ブースティング	M	$\{35, 40, \cdots, 50\}$	50	$\{180, 185, \cdots, 195\}$	195
	$d_{\max}$	$\{9, 10, \cdots, 14\}$	9	$\{1, 2, \cdots, 10\}$	5
XGBoost	M	$\{2^i \mid i = 11, 12, \cdots, 15\}$	2^{11}	$\{974, 984, \cdots, 1074\}$	1054
	$d_{\max}$	$\{1, 6, \cdots, 46\}$	11	$\{1, 2, \cdots, 10\}$	3

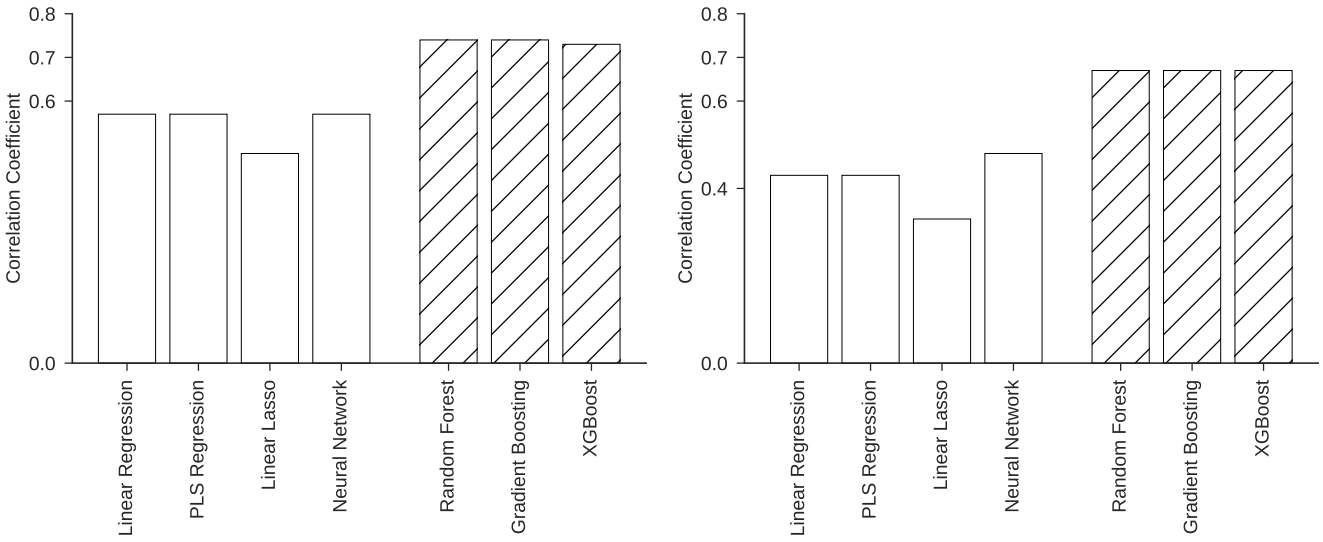

図 4　PR 値の実測値と予測値の相関係数. 左が HS，右が Con.

の mRNA 配列に適合するように学習されたものであるため，人工的に合成した mRNA 配列に対しても翻訳量を高精度に予測できるとは限らない．また，R-STEINER によって生成された 5'UTR はヌクレオチドの順序変換に対して PR 値が不変である．すなわち，特徴量ベクトルに対して高い PR 値を推定した場合でも，実際に生成した 5'UTR が遺伝子の翻訳量を増加させない可能性もある．

これらの問題が実際に起きているかどうかを検証するため，F-luc と呼ばれる CDS 配列と固定された 3'UTR に対して R-STEINER と同様の手順—Algorithm 1 の 9 行目まで—で 5'UTR を生成し，これらを結合して出来る mRNA を実際に合成する実験を行った．本実験で 5'UTR のみを生成する理由は，現実の合成実験を考えた際に，現実的に編集可能な領域が 5'UTR のみだからである．CDS は合成されるタンパク質を決定するため，CDS を改変してしまうと合成されるタンパク質も変化してしまう．これは応用上の観点から避けるべきであるため，本研究では CDS は生成しない．また，3'UTR 配列には 3'UTR の末端を決定する情報が含まれているため，3'UTR 配列を編集すると実際の 3'UTR 配列は想定とは異なる部分で切れてしまう可能性がある．したがって，本研究では 3'UTR 領域の変更も行わないこととした．

合成実験では合成した mRNA の翻訳量を測定するため， F/R-luc activity という指標を用いた．PR 値は $\log_{10}$(F/R-luc activity) と線形の関係にあることが知られている [25] ので，PR 値と $\log_{10}$(F/R-luc activity) の相関係数を用いて実験結果の評価を行う．

6.1　実験設定

合成実験では，mRNA が合成される環境は Con を理想とし

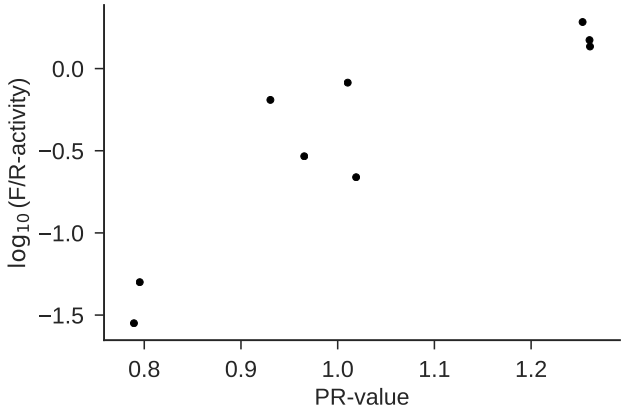

図 5　合成実験の結果: 横軸が PR 値, 縦軸が $\log_{10}$(F/R-luc activity). 相関係数は 0.89.

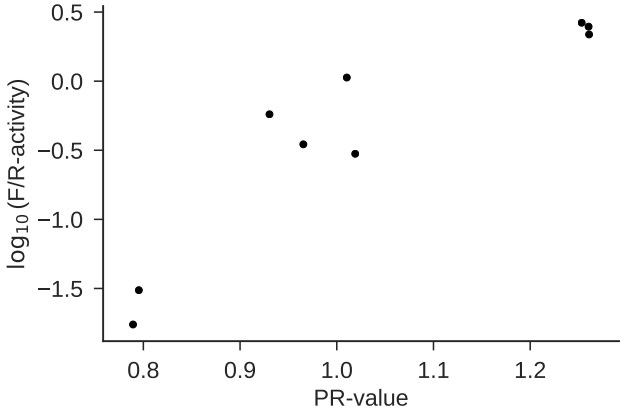

図 6　合成実験 (再現) の結果: 横軸が PR 値, 縦軸が $\log_{10}$(F/R-luc activity). 相関係数は 0.91

ているが, 実験中に様々な要因で細胞にストレスがかかり得る. そのため, (2) のように Con と HS での PR 値予測値を平均して最終的な予測値としている.

評価実験の手順を以下に示す.

step 1　R-STEINER と同様の方法で 5'UTR の配列を生成する. すなわち, Algorithm 1 の 9 行目までを行う.

step 2　生成された 5'UTR の中から, 予測 PR 値が高いものから三つ, 低いものから二つ, 生成された配列全体から無作為に四つ (ただし, 左記の配列と同一でないもの) を選択する.

このようにして選択された 5'UTR に CDS と 3'UTR を連結した mRNA を実際に合成し, 予測 PR 値と観測された F/R-luc activity の相関係数を計算する.

6.2　実験結果と評価

図 5 と図 6 に実験結果を示す.　図 6 は合成実験の再現性を確認するため—すなわち, 実験上のミス等によって偶然図 5 のような結果が得られのではなく, 同じ実験をすると同様の結果が得られることを確認するため—に, 図 5 の結果を出した実験と同様の実験を行ったものである. 図 5 から分かるように, 予測 PR 値と $\log_{10}$(F/R-luc activity) の観測値は強い相関がある (相関係数は 0.89). すなわち, PR 値の予測値が大きくなるほど真の翻訳量も大きくなる. また, 図 5 と図 6 は散布図の外形がおおむね一致している. このことから, 本実験に関する再現性は担保されている (注1).

本合成実験の結果から, R-STEINER で使われている PR 値予測モデルは人工的に作られた mRNA に対しても高精度に翻訳量を予測できていることが分かった. したがって, Algorithm 1 において生成配列の個数 B を増加させることにより, R-STEINER は特定遺伝子の翻訳エンハンサーを生成することができる.

(注1): 実験手順等に関するミスで図 5 の結果が出たのではないということ

7. おわりに

本稿では, 特定遺伝子の翻訳エンハンサー発見方法である R-STEINER を提案した. R-STEINER での B-step において, PR 値の予測モデルを構築した. PR 値の予測モデルには, 決定木ベースの予測器によるアンサンブル予測モデルが, 他の手法に比べて良い予測精度を発揮した. これは, **??**節で用いた特徴量の内, 自由エネルギーを除く全ての特徴量が離散型の特徴量であるからだと考えられる. また, この結果は HS と熱ストレス下 Con で共通であったため, 植物が置かれている環境に依存していないと考えられる. そして, R-STEINER で使われている予測モデルが, 人工 mRNA に対しても高い精度で翻訳量を予測できることを示した. つまり, R-STEINER を用いて発見される翻訳エンハンサーは, 実際に特定遺伝子の翻訳量を向上させることができ, Algorithm 1 における B を十分大きくして生成された翻訳エンハンサーのみに対して合成実験を行うことで, 合成実験のコスト・時間・労力を削減することができる.

本研究で残された課題として, G-step における配列生成方法が挙げられる. 本手法の G-step では, ヌクレオチドをランダムに生成し, それらをつなぎ合わせて翻訳エンハンサーの候補を生成している. 本来ならば

$$\boldsymbol{x}' = \arg\max_{\boldsymbol{x}} \hat{h}(\boldsymbol{x}) \tag{5}$$

を求めて, $\boldsymbol{x}'$ から 5'UTR を生成するという方法を採るべきである. しかし, このような方法には二つの課題がある. 一つ目は, 式 (5) を求めることは容易ではないということである. 二つ目は, 4.1.1 節節での特徴量設計方法から分かるように, 特徴ベクトルとそれに対応する mRNA 配列が一対一対応ではないという点である.

一つ目の課題を解決するためには, 238 次元のベクトルを変数とする明示的に書くことができない関数の最適化問題を解く手法が必要である. そのような手法が開発された後, 二つ目の課題を解決する方法が必要となる. 本研究における特徴量設計の方法では, 一つの特徴ベクトルに複数の mRNA 配列が対応する. したがって, 二つ目の課題を解決する具体的な方法とし

て，生成された複数の mRNA から何らかの基準によって一つの mRNA 配列を選択するという方法が考えられる．これら二つの課題が解決された場合，特定遺伝子の翻訳エンハンサーをより短時間で発見することが可能となると考えられる．

謝　　辞

本研究は NAIST 多元ビッグデータプロジェクトの助成および，新エネルギー・産業技術総合開発機構の「植物や他の生物のスマートセルを用いた高機能生体材料の製造技術の開発」による支援を受けたものである．また，本研究で用いたデータセットは DANAFORM の助力を得て作成された．合成実験に関しては，奈良先端科学技術大学院大学バイオサイエンス研究科の鈴木淳展氏に合成実験を行って頂いた．

文　　献

[1] Bruce Alberts, Dennis Bray, Karen Hopkin, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter. *Essential cell biology*. Garland Science, 2013.

[2] Leo Breiman. Random forests. *Machine learning*, Vol. 45, No. 1, pp. 5–32, 2001.

[3] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794. ACM, 2016.

[4] Jerome H Friedman. Stochastic gradient boosting. *Computational Statistics & Data Analysis*, Vol. 38, No. 4, pp. 367–378, 2002.

[5] Jerome Friedman, Trevor Hastie, Holger Höfling, Robert Tibshirani, et al. Pathwise coordinate optimization. *The Annals of Applied Statistics*, Vol. 1, No. 2, pp. 302–332, 2007.

[6] Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, Vol. 33, No. 1, p. 1, 2010.

[7] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. The elements of statistical learning, 2001.

[8] Riki Kawaguchi and Julia Bailey-Serres. mrna sequence features that contribute to translational regulation in arabidopsis. *Nucleic Acids Research*, Vol. 33, No. 3, pp. 955–965, 2005.

[9] Ronny Lorenz, Stephan H Bernhart, Christian Hoener Zu Siederdissen, Hakim Tafer, Christoph Flamm, Peter F Stadler, and Ivo L Hofacker. Viennarna package 2.0. *Algorithms for Molecular Biology*, Vol. 6, No. 1, p. 26, 2011.

[10] Hideyuki Matsuura, Shinya Takenami, Yuki Kubo, Kiyotaka Ueda, Aiko Ueda, Masatoshi Yamaguchi, Kazumasa Hirata, Taku Demura, Shigehiko Kanaya, and Ko Kato. A computational and experimental approach reveals that the 5′ -proximal region of the 5′ -utr has a cis-regulatory signature responsible for heat stress-regulated mrna translation in arabidopsis. *Plant and cell physiology*, Vol. 54, No. 4, pp. 474–483, 2013.

[11] J Močkus. On bayesian methods for seeking the extremum. In *Optimization Techniques IFIP Technical Conference*, pp. 400–404. Springer, 1975.

[12] Jonas Mockus. On bayesian methods for seeking the extremum and their application. In *IFIP Congress*, pp. 195–200, 1977.

[13] Jonas Mockus. *Bayesian approach to global optimization: theory and applications*, Vol. 37. Springer Science & Business Media, 2012.

[14] LA Rastrigin. The convergence of the random search method in the extremal control of a many parameter system. *Automaton & Remote Control*, Vol. 24, pp. 1337–1342, 1963.

[15] Günther Schrack and Mark Choit. Optimized relative step size random searches. *Mathematical Programming*, Vol. 10, No. 1, pp. 230–244, 1976.

[16] M Schumer and Kenneth Steiglitz. Adaptive step size random search. *IEEE Transactions on Automatic Control*, Vol. 13, No. 3, pp. 270–276, 1968.

[17] Jamie Shotton, Toby Sharp, Alex Kipman, Andrew Fitzgibbon, Mark Finocchio, Andrew Blake, Mat Cook, and Richard Moore. Real-time human pose recognition in parts from single depth images. *Communications of the ACM*, Vol. 56, No. 1, pp. 116–124, 2013.

[18] Tadatoshi Sugio, Junko Satoh, Hideyuki Matsuura, Atsuhiko Shinmyo, and Ko Kato. The 5'-untranslated region of the oryza sativa alcohol dehydrogenase gene functions as a translational enhancer in monocotyledonous plant cells. *Journal of bioscience and bioengineering*, Vol. 105, No. 3, pp. 300–302, 2008.

[19] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

[20] Stephen Tyree, Kilian Q Weinberger, Kunal Agrawal, and Jennifer Paykin. Parallel boosted regression trees for web search ranking. In *Proceedings of the 20th international conference on World wide web*, pp. 387–396. ACM, 2011.

[21] Paul Viola and Michael J Jones. Robust real-time face detection. *International journal of computer vision*, Vol. 57, No. 2, pp. 137–154, 2004.

[22] H Wold. Estimation of principal components and related models by iterative least squares. multivariate analysis. edited by: Krishnaiah pr. 1966.

[23] Jian Yao, Yunqi Weng, Alexia Dickey, and Kevin Yueju Wang. Plants as factories for human pharmaceuticals: applications and challenges. *International journal of molecular sciences*, Vol. 16, No. 12, pp. 28549–28565, 2015.

[24] 山崎将太朗, 上田清貴, 加藤晃. 環境ストレスの影響を考慮した植物発現ベクターの開発 (< 特集 > 植物形質転換に関する最新技術). 生物工学会誌: seibutsu-kogaku kaishi, Vol. 91, No. 8, pp. 456–460, 2013.

[25] 山崎将太郎. 植物 mRNA の翻訳機構に関する研究. PhD thesis, 奈良先端科学技術大学院大学, 2016.

[26] 田辺三菱製薬株式会社. 次世代新規ワクチンの共同研究契約締結について（２０１２年３月７日発表）. `https://www.mt-pharma.co.jp/shared/show.php?url=../release/nr/2012/MTPC120307.html`.

[27] 藤山和仁. 植物を使った医療用タンパク質生産の挑戦. 生物工学会誌: seibutsu-kogaku kaishi, Vol. 90, No. 9, pp. 563–566, 2012.