Vol.2017-BIO-49 No.2
2017/3/23

アンサンブル学習を用いたmRNA配列からの翻訳効率予測

田中 宏昌[1,a)]　鈴木 優[1]　山崎 将太朗[1]　加藤 晃[1]　吉野 幸一郎[1]　中村 哲[1]

概要：本稿では，メッセンジャー RNA 配列から機械学習を用いて翻訳効率を予測する手法を提案する．従来は単回帰や PLS モデルなどが翻訳効率の予測に用いられていたが，十分な精度で翻訳効率を予測することが困難であった．一方で，ある特徴から数値を予測する手法は機械学習分野で数多く提案されており，それらの手法を用いることにより翻訳効率の予測精度を改善することが可能となると考えた．そこで，翻訳効率の予測精度を改善するための方法として，近年予測モデルとして用いられている Random Forest Regression および Gradient Boosting Regression を利用することを考えた．これらの予測モデルはアルツハイマー病の臨床スコア予測をはじめとしてバイオインフォマティクス分野でよく用いられているため，予測精度の向上が期待できる．構築した予測モデルによりシロイヌナズナのメッセンジャー RNA の翻訳効率の予測を行った結果，従来手法と比較して高精度であることが分かった．

キーワード：mRNA，ランダムフォレスト回帰，勾配ブースティング回帰，翻訳効率予測，シロイヌナズナ

Hiroaki tanaka[1,a)]　Yu Suzuki[1]　Syotaro Yamasaki[1]　Ko Kato[1]　Koichiro Yoshino[1]
Satoshi Nakamura[1]

1. はじめに

製薬において，メッセンジャー RNA (mRNA) から効率良くタンパク質を生成することは重要な目的の一つである．しかし，mRNA の働きに関しては未知の部分も多いため，どのような配列パターンの mRNA が多量のタンパク質を生成するかという点は明らかになっていない．

mRNA からタンパク質を生成するまでの過程を翻訳と呼ぶ．そして，翻訳の効率を表す指標として，本稿では PR (Polysome Ratio) 値を用いる．翻訳に関する説明は 3.2 節に記載している．生命科学分野においてこれまでに得られた知見に従って翻訳効率の良い mRNA を演繹的に求める方法もあるが，それ以外の新たな方法が求められている．

例えば，多数の mRNA を用いて探索的にタンパク質翻訳効率の良い配列を探す方法が考えられる．しかし，この方法では多くの mRNA の作製を要するために大規模な実験が必要になる．そこで，計算機上で mRNA 配列の合成と翻訳効率をシミュレーションし，その配列が翻訳効率の良い配列であれば実際にその配列を用いて実証するという方法が考えられる．このような方法を実現するためには，mRNA 配列から高精度で翻訳効率を予測出来る予測器が必要不可欠である．そのため本研究では，mRNA 配列から翻訳効率を予測する予測器を構築した．

本研究で取り組む問題は，数学的には適当な特徴量から PR 値を予測するという問題，つまり回帰の問題であると考えられる．バイオインフォマティクスの分野ではこのような問題に対して，ランダムフォレスト回帰が用いられる事が多い．例えば，Huang ら [2] はアルツハイマー病の臨床スコアを予測するためにランダムフォレスト回帰を用いている．そこで，mRNA の翻訳効率を予測するためにもランダムフォレスト回帰を用いることが有効であると考えた．さらに，ランダムフォレスト回帰以外にも様々な予測器が有用である可能性が考えたことから，回帰木を用いる手法である勾配ブースティング回帰を翻訳効率予測に用いた．

本研究により，mRNA 配列からの PR 値予測には，今回検討した手法の中で，勾配ブースティング回帰が最良の手法であることや，翻訳効率の制御に重要と思われる配列パターンを特定することが可能となった．

2. 関連研究

本節では，Kawaguchi ら [4] と Matsuura ら [5] の研究を

[1] 奈良先端科学技術大学院大学
Nara Institute Science and Technology
[a)] tanaka.hiroaki.sy2@is.naist.jp

紹介する．

Kawaguchi ら [4] は，シロイヌナズナを対象として ribosome loading と呼ばれる指標を用いている．ribosome loading は mRNA に結合しているリボソームの量を数値化したもので，翻訳効率の指標値として用いることができる．彼らは，C,..., UU 含量の相関分析を始めとして様々な分析を行っているが，例えば，5'UTR[*1]の長さと ribosome loading の関係の分析や，3'UTR[*2]の長さと ribosome loading の関係の分析などいずれも 2 変量間の分析である．このような分析方法は，変量間の関係性を考察しやすいという利点はあるが，一方で複数の変量が複雑に絡みあった関係性を捉えることはできない．

Kawaguchi ら [4] が 2 変量間での分析を主に行っていたのに対し，Matsuura ら [5] は，PLS 回帰モデルを用いて複数の特徴量から relative F-Luc activity と呼ばれる値の予測器を作成している．relative F-Luc activity は，5'UTR 配列が熱ストレス時の翻訳へ与える影響を評価する指標である．より具体な予測器の作成方法は，5'UTR を様々な領域に区切り，その領域内での AAA, AAC,..., UUU の頻度を特徴量として PLS 回帰モデルを構築するというものである．領域の区切り方と同数の PLS 回帰モデルが作成されるが，

$$Q^2 = 1 - \frac{\sum_y (y - \widehat{y})^2}{\sum_y (y - \overline{y})^2} \tag{1}$$

を用いてモデル選択を行っている．(1) において，y は目的変数の観測値，$\widehat{y}$ は y の推定値，$\overline{y}$ は y の標本平均である．なお，Matsuura ら [5] は，翻訳効率予測それ自体を目的としているわけではなく，熱ストレス環境下での翻訳効率の決定に重要な 5'UTR 領域を特定するために，予測器を構築している．したがって，本研究とは最終的な目的が異なっている．

3. 提案手法

本節では，ランダムフォレスト回帰を用いて mRNA 配列から翻訳効率を予測する手法を提案する．まずはランダムフォレスト回帰と勾配ブースティング回帰について説明し，その後どのようにランダムフォレスト回帰と勾配ブースティング回帰による予測機を構築したかを説明する．

3.1 mRNA からタンパク質への翻訳

まずは，mRNA からタンパク質が生成される過程（翻訳）について説明する．mRNA は A, C, G, U が並べられた文字列であり，mRNA 配列と呼ぶ．配列長および，A, C, G, U の並び方も mRNA 種によって異っており，このような mRNA 配列からアミノ酸が合成（翻訳）される．mRNA 配列は 5'UTR, CDS, 3'UTR という 3 つの領域に分けられている．この 3 領域の中で，実際にアミノ酸へと翻訳される領域は CDS のみである．そして，リボソームと呼ばれる翻訳装置が 5'UTR 側から mRNA を読み取り，アミノ酸を合成する．合成されるアミノ酸は mRNA の A, C, G, U の並びによって決定される．例えば，CAU という並びからはヒスチジンが合成される．これらのアミノ酸が複雑に絡み合い，タンパク質としての機能を発揮する．翻訳に関してより詳しい説明は，瀬々ら [6] を参照されたい．

3.2 翻訳効率

ここでは，mRNA からタンパク質への翻訳効率を表す指標を導入する．活発に翻訳が行われている mRNA には複数のリボソームが結合し，ポリソームを形成している．このポリソームの形成度合いを表す指標として PR 値を定義する．PR 値は細胞内に存在する mRNA がポリソームを形成している比率として計算され，PR 値が高いほど，翻訳効率が良い．

3.3 ランダムフォレスト回帰

ランダムフォレスト回帰は，複数の回帰木の平均によって目的変数を予測するアンサンブル学習の一種である．新しいデータの特徴量ベクトル $\boldsymbol{x}$ に対する予測アルゴリズムを以下に記述する．

Step 1 For $r = 1$ to R

(1) トレーニングセットから大きさ N のブートストラップ標本 X_r を抽出する．

(2) 標本 X_r から回帰木 T_r を作成する．

Step 2 予測値

$$\widehat{y}(\boldsymbol{x}) = \frac{1}{R} \sum_{r=1}^{R} T_r(\boldsymbol{x}) \tag{2}$$

で予測を行う．

ランダムフォレスト回帰のアルゴリズムでは，ハイパーパラメータとして回帰木の本数 R を決定する必要がある．本稿の実験では，R はグリッドサーチを用いて決定した．これらハイパーパラメータの決定方法についての詳細は，3.5 節で述べる．

また，ランダムフォレスト回帰の特徴として，特徴量の重要度を算出することが出来る．それぞれの回帰木のそれぞれの分割で，分割基準（本稿の実験では平均 2 乗誤差）をどれだけ改善できたのかを，その特徴量の重要度とみなす．4.5 節ではこの重要度を用いて，翻訳効率を制御する因子に関する議論を行う．

3.4 勾配ブースティング回帰

ランダムフォレスト回帰がバギングと呼ばれるアンサン

[*1] mRNA の末端部の名称．詳細は 3.1 節を参照．
[*2] mRNA の末端部の名称．詳細は 3.1 節を参照．

ブル学習の一種であるのと同様に，勾配ブースティング回帰もアンサンブル学習の一種である．ここでは勾配ブースティング回帰に関する詳細な説明は避け，直感的な説明を行う．勾配ブースティング回帰に関しての詳細な説明はビショップ [1] や Hastie ら [8] を参照されたい．

ブースティングは，逐次的に弱学習器を構築していく手法である．新しい弱学習器を構築する際，それまでに構築されたすべての弱学習器の結果を利用する．そのためすべての弱学習器が独立に学習されるバギングと比べて，計算を並列化できないので学習に時間がかかる．ブースティングでは各ステップごとに弱学習器を構築して損失関数を最小化する．その際，前のステップで損失が大きかった学習データへの重みを大きくして，次のステップで学習を行う．

ブースティングの各ステップで行っていることは，損失関数の最小化である．この最小化におけるパラメータの最適化に勾配降下法を用いて学習する方法を，勾配ブースティングと呼ぶ．なお，勾配降下法に関する詳細は Trevor ら [8] を参照されたい．

また，勾配ブースティングではハイパーパラメータとして逐次学習のステップ数 s，個々の回帰木の深さの最大値 d_G，学習率 η を決定しなければならない．本稿の実験ではこれらのハイパーパラメータをグリッドサーチを用いて決定した．ハイパーパラメータの決定に関する詳細は，4.3 節で説明する．

3.5 ランダムフォレスト回帰による予測機の構築

領域 R における塩基配列 x の出現回数を c_x^R で表し，$V = \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{U}, \mathrm{AA}, \mathrm{AC}, \cdots, \mathrm{UU}, \mathrm{AAA}, \mathrm{AAC}, \cdots, \mathrm{UUU}\}$ とする．ランダムフォレスト回帰に使用する特徴量は，$c_v^{5'\mathrm{UTR}}, c_v^{\mathrm{CDS}}, c_v^{3'\mathrm{UTR}}$ である．ただし，v は V のすべての元にわたる．また，ハイパーパラメータの決定に関しては，$R \in \{10, 50, 100, 150\}$ でバリデーションセットを用いたグリッドサーチを行って $R = 150$ に決定した．

3.6 勾配ブースティング回帰による予測器の構築

特徴量はランダムフォレスト回帰と同様の特徴量を用いた．ハイパーパラメータは，それぞれ $s \in \{50, 100, 150, 200\}, d_G \in \{5, 10, 15\}, \eta \in \{0.01, 0.05, 0.1\}$ の範囲でバリデーションセットを用いたグリッドサーチによって決定した．その結果，$s = 200, d_G = 5, \eta = 0.05$ となった．

4. 実験

本節では，実際のデータを用いて提案手法の精度を確認する．さらに，いくつかの予測器を作成し，提案手法と比較する．最後に実験結果に対する考察を与える．

表 1: mRNA 配列の例

遺伝子	5'UTR	CDS	3'UTR
A_1	GAGAGCA	AUGGCG...	UCCCA...
A_2	AGAGAGCA	AUGGCG...	UCCCA...
⋮	⋮	⋮	⋮
A_{N_A}	GAGAGAG	AUGGCG...	UCCCA...
B_1	GCCAAU	AUGAGU...	AGCUUC...
⋮	⋮	⋮	⋮

4.1 実験準備

本研究で使用したデータは，熱ストレス状況下（37°C）で採取されたシロイヌナズナの mRNA データである．CDS, 3'UTR のサンプル数は 7,986 配列，5'UTR のサンプル数は 162,181 配列である．すなわち，**表 1** のように CDS・3'UTR 一つに対して複数の 5'UTR が対応している．一方で，一つの遺伝子 A に対して一つの PR が対応している．つまり，遺伝子 A, B, ... に対しては PR 値が一対一で対応しているが，$\mathrm{A}_1, \mathrm{A}_2, \ldots, \mathrm{A}_{N_A}$ に対しては PR 値が一対一に対応していない．

表 1 のようなデータから，3.5 節に記載した特徴量を算出する方法を説明する．まず CDS と 3'UTR に関しては，PR 値と配列が一対一に対応してるので，各領域における出現回数をそのままカウントすればよい．次に $c_v^{5'\mathrm{UTR}}, c_v^{\mathrm{CDS}}, c_v^{3'\mathrm{UTR}}$ に関してであるが，遺伝子 A における塩基配列 v の出現頻度は

$$c_v^{5'\mathrm{UTR}} = \sum_{n=1}^{N_A} f_n^{(\mathrm{A}_n)} c_v^{5'\mathrm{UTR}-\mathrm{A}_n} \tag{3}$$

で算出する．ただし，$f_n^{(\mathrm{A})}$ は A_n の存在比率，$c_v^{5'\mathrm{UTR}-\mathrm{A}_n}$ は A_n の 5'UTR における塩基配列 v の出現回数を表す．最後に，mRNA 全体における $v \in V$ の出現回数は，

$$c_v^{\mathrm{mRNA}} = c_v^{3'\mathrm{UTR}} + c_v^{\mathrm{CDS}} + c_v^{5'\mathrm{UTR}} \tag{4}$$

で算出する．

4.2 実験の手順

まずは実験の手順を記す．CDS 及び 3'UTR が異なっている 7,986 配列の 50% をトレーニングセット，25% をバリデーションセット，25% をテストセットとして実験を行った．なお，データの分割方法は Trevor ら [8] を参考にした．トレーニングセットと開発セットを用いて構築した予測機でテストセットの PR 値を予測し，観測値との平均 2 乗誤差を用いて評価した．

4.3 比較手法

ランダムフォレスト回帰による予測精度と比較するため，線形回帰モデル（Linear Regression; LR）・PLS 回帰モデル（PLS Regression; PLS）による PR 値予測器を作

成した．

線形回帰モデル　線形回帰モデル [7] を用いて回帰を行った．特徴量はランダムフォレスト回帰と同様の特徴量を用いた．回帰式は

$$\text{PR-value} = \sum_{v\in V} \alpha_v c_v^{5'\mathrm{UTR}} + \sum_{v\in V} \beta_v c_v^{\mathrm{CDS}} + \sum_{v\in V} \gamma_v c_v^{3'\mathrm{UTR}} \quad (5)$$

である．

PLS 回帰モデル　特徴量はランダムフォレスト回帰と同様の特徴量を用いた．主軸の本数は，バリデーションセットにおけるグリッドサーチを用いて $[2, 62]$ の範囲から決定した．PLS 回帰モデルに関しては橋本ら [3] が詳しい．

4.4 結果と考察

実験の結果は**表 2** のようになった．表 2 から，ランダムフォレスト回帰は線形回帰モデルよりも平均 2 乗誤差にして 0.27×10^{-2}，PLS モデルよりも 0.95×10^{-2} だけ精度が良くなっていることが分かる．また，勾配ブースティング回帰はランダムフォレスト回帰よりも 0.17×10^{-2} だけ精度が良くなっていることが分かる．

より詳細な考察のため，**図 1** に各種法による予測値と観測値の分布を示す．図 1 は，ヒストグラムをカーネル密度推定したものである．なお，カーネルはガウシアンカーネルを使用した．各図において実線はテストセットの観測値を，破線がテストセットでの予測値を，点線がテストセットでの観測値の標本平均を表している．

まず**図 1a** から分かるように，PLS 回帰モデルによる予測値は観測値の標本平均周りに集中していて，分布の広がりを捉えきれていない．したがって，PLS 回帰モデルは各領域におけるすべての $v \in V$ から PR 値を予測するという問題に対しては，表現力が十分でないと言える．

次に，**図 1b** について考える．図 1b から分かるように，線形回帰モデルは PLS 回帰モデルよりも分布の広がりを捉えることが出来ている．しかし，やはり線形回帰モデルによる予測値は標本平均周りに集中している．さらに，予測値の分布が対称になっており，観測値の分布の歪さを捉えきれていない．

次に，**図 1c** について考える．図 1c から，ランダムフォレスト回帰は線形回帰モデルよりも観測値の分布の広がりや歪さを捉えられている事が分かる．

表 2: 各モデルによる平均 2 乗誤差

モデル	トレーニングセット	テストセット
PLS	2.85×10^{-2}	2.73×10^{-2}
LR	2.00×10^{-2}	2.17×10^{-2}
RFR	2.71×10^{-3}	1.90×10^{-2}
GBR	6.74×10^{-3}	1.73×10^{-2}

最後に，**図 1d** について考える．勾配ブースティング回帰が他のどの手法よりも，観測値の分布の広がりや歪さを捉えている．さらに，他の手法に比べて，予測値の分布が標本平均周りに集中するという現象も幾分緩和されている事が分かる．

以上の考察から，$c_v^{5'\mathrm{UTR}}, c_v^{\mathrm{CDS}}, c_v^{3'\mathrm{UTR}}$ を特徴量として PR 値を予測するという問題に対しては，勾配ブースティング回帰による予測が最も優れていると言える．

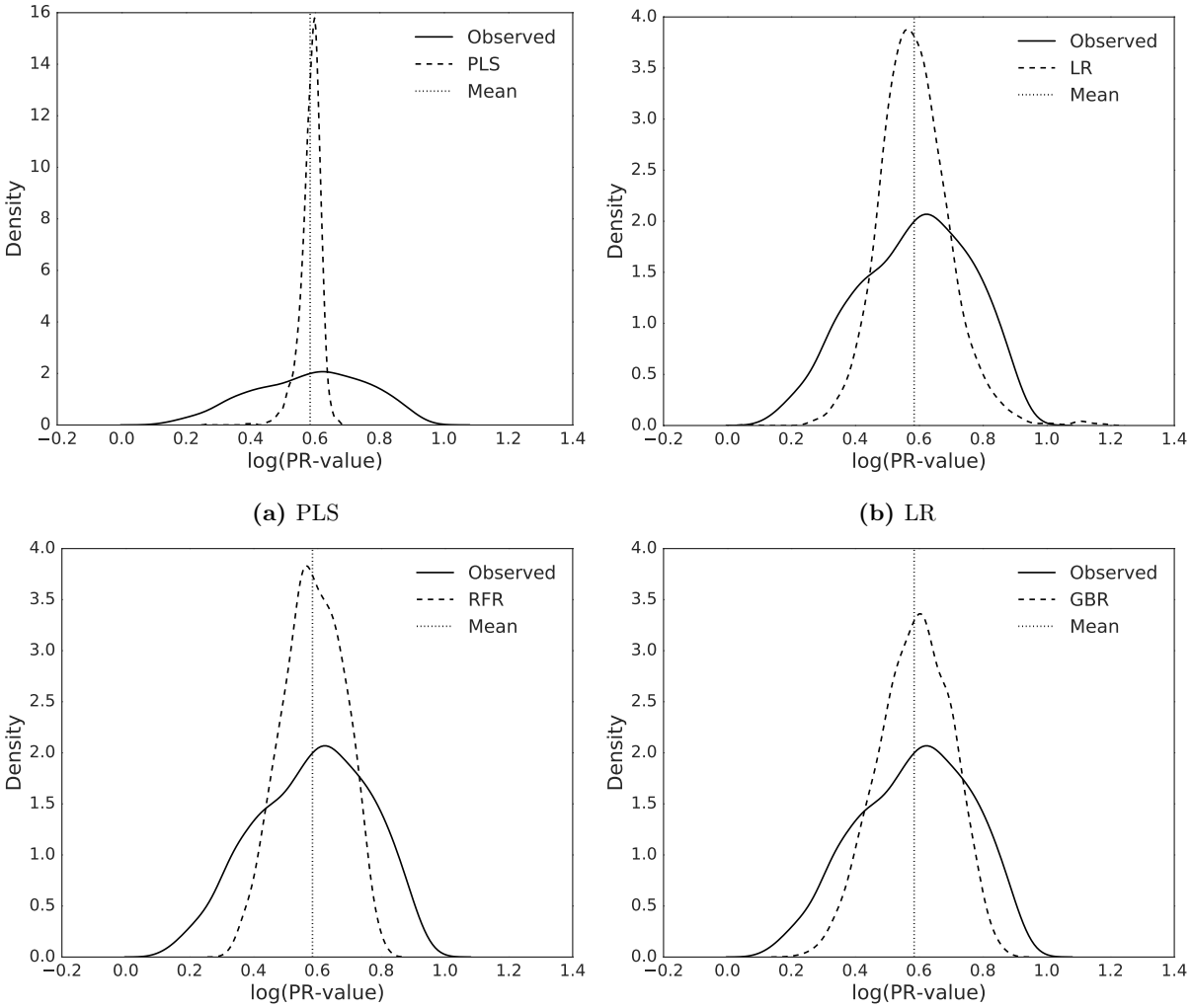
4.5 議論

ランダムフォレスト回帰と勾配ブースティング回帰は，使用した特徴量の重要度を算出することが出来る．本稿で行った実験で使用した 336 個の特徴量に対して，特徴量の重要度を算出した．**図 2a** はランダムフォレスト回帰による特徴量重要度，**図 2b** は勾配ブースティング回帰による特徴量重要度である．両図とも，重要度を昇順に並べ替えた時の上位 5% に当たる特徴量のみを描画した．横軸が塩基配列を表し，縦軸が重要度を表している．さらに，塩基配列が含まれる領域によって色分けされている．図 2a, 図 2b から，5'UTR における塩基の特徴量が重要である事が分かる．ランダムフォレスト回帰では，選ばれた特徴量の中で AA 以外のすべての特徴量は 5'UTR での塩基出現回数である．同様に，勾配ブースティング回帰では，選ばれたすべての特徴量が 5'UTR での塩基出現回数となっている．これらの結果から，5'UTR での塩基出現回数は PR 値予測に関して重要な働きをしている事がわかる．したがって，5'UTR の構造は PR 値制御に関して重要な働きを持つことが予想される．なお，5'UTR の重要性を示す結果は Kawaguchi ら [4] や Matsuura ら [5] でも示されている．特に，Kawaguchi ら [4] では A,..., U, AC, AG, AU, CU, GC, GU の重要度を算出している．本稿では塩基配列の重要度を 3 塩基までの塩基パターンで考えることにより，より詳細に塩基配列の翻訳効率への影響を考察することができた．

一方，図 2a, 図 2b の両図において，3'UTR での塩基出現回数が選ばれていない．このとこから，3'UTR での塩基出現回数は PR 値予測に関する重要度が低いことが分かる．つまり，3'UTR の構造は PR 値制御に関してあまり重要でないと予想される．

5. おわりに

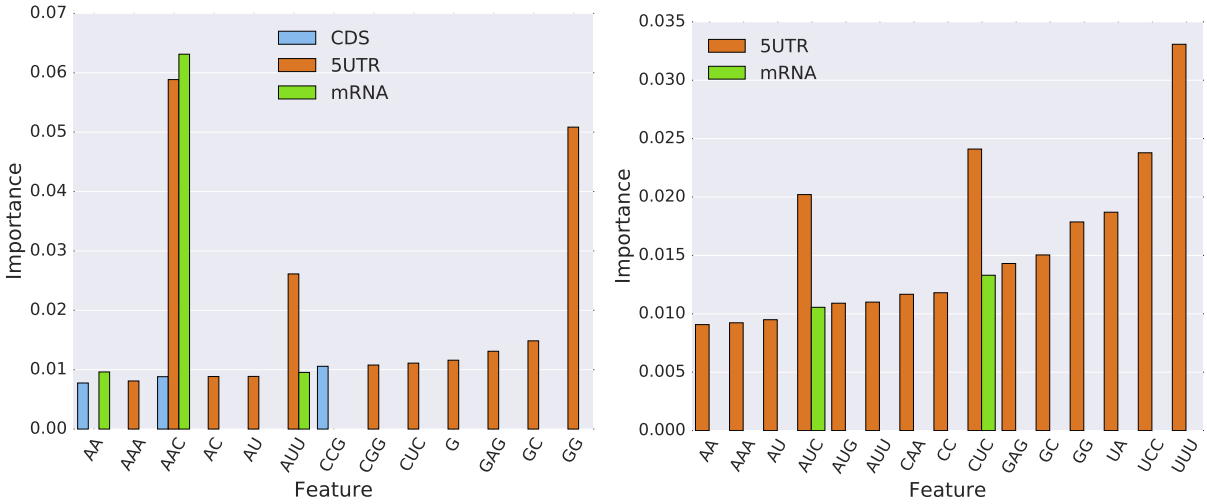
本稿では，mRNA 配列から翻訳効率を予測する手法を提案した．評価実験の結果，PR 値の予測には勾配ブースティング回帰が優れた手法であることが分かった．また，ランダムフォレスト回帰と勾配ブースティング回帰による特徴量重要度を算出することによって，5'UTR における塩基配列の出現回数が PR 値制御に関して重要であるという予測を得た．



(a) PLS　**(b)** LR

(c) RFR　**(d)** GBR

図 1: 予測値と観測値の分布



(a) RFR　**(b)** GBR

図 2: 特徴量の重要度

しかし，本研究だけでは PR 値の予測は可能になっただけであり，PR 値の高い mRNA 配列を設計するためには不十分である．なぜなら，ランダムフォレスト回帰や勾配ブースティング回帰は PR 値の予測式を解析的に書くこと

ができず，「どの塩基配列を増やせば PR 値が増加するか」や「どの塩基配列を減らせば PR 値が増加するか」といったことが分からないからである．また，ランダムフォレスト回帰や勾配ブースティング回帰による予測器からは，各塩基の順序を決定するための情報を得ることができないことも，本研究だけから配列を設計することが不可能な理由の一つである．

したがって，今後の研究として，本稿で提案した予測器を評価関数として用いて mRNA を設計する研究が必要である．また，本稿では特徴量として 3-gram までの塩基配列を用いたが，4-gram, 5-gram というように，より長い塩基配列を特徴量として考慮に入れることも今後の課題である．さらに，4.5 節で考察を行った塩基出現回数の重要度に関してバイオサイエンスの立場からより深い考察及び実験を行うことも，価値ある結果に繋がる．

謝辞　本研究で用いたデータは，奈良先端科学技術大学院大学バイオサイエンス研究科よりご提供頂いた．また，本研究は NAIST ビッグデータプロジェクトの助成を受けたものである．

参考文献

[1] C.M. ビショップ：パターン認識と機械学習 (下) ベイズ理論による統計的予測，丸善出版 (2008).
[2] Huang, L., Jin, Y., Gao, Y., Thung, K.-H., Shen, D., Initiative, A. D. N. et al.: Longitudinal clinical score prediction in Alzheimer's disease with soft-split sparse regression based random forest, *Neurobiology of aging*, Vol. 46, pp. 180–191 (2016).
[3] 橋本淳樹，田中　豊：PLS 回帰におけるモデル選択，アカデミア 情報理工学編，Vol. 10, pp. 39–49 (2010).
[4] Kawaguchi, R. and Bailey-Serres, J.: mRNA sequence features that contribute to translational regulation in Arabidopsis, *Nucleic Acids Research*, Vol. 33, No. 3, pp. 955–965 (2005).
[5] Matsuura, H., Takenami, S., Kubo, Y., Ueda, K., Ueda, A., Yamaguchi, M., Hirata, K., Demura, T., Kanaya, S. and Kato, K.: 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).
[6] 瀬々　潤，浜田道昭：生命情報処理における機械学習: 多重検定と推定量設計，講談社 (2015).
[7] 佐和隆光：回帰分析，朝倉書店 (1979).
[8] Trevor, H., Robert, T. and Jerome, F.: 統計的学習の基礎: データマイニング・推論・予測，共立出版 (2014).