Vol.2016-MPS-108 No.25
Vol.2016-BIO-46 No.25
2016/7/5

機械学習を用いた非モデル生物におけるミトコンドリア及び関連オルガネラタンパク質の予測手法

久米　慶太郎[1,2]　天笠　俊之[3]　橋本　哲男[2]　北川　博之[3]

概要：本研究では、ミトコンドリアおよび関連オルガネラに輸送されるタンパク質を予測する分類手法を提案する。既存の分類手法は、文字列で表現されるタンパク質配列データを教師データとした機械学習によって作成されてきた。しかしそのほとんど全てがモデル生物というごく一部の生物で得られた知見をもとに、特徴量及び教師データを用意しているため、それ以外の生物に対しては精度に劣っていた。これは、広範な生物におけるミトコンドリア進化の研究を行うにあたって解決すべき問題となっている。そこで、本研究では、非モデル生物を対象とした分類手法を提案する。この分類手法は、既存の分類手法と比べて同等以上の実用的な高い精度を示すことを確認した。

Keitaro Kume[1,2]　Toshiyuki Amagasa[3]　Tetsuo Hashimoto[2]　Hiroyuki Kitagawa[3]

1. はじめに

1.1 ミトコンドリア（Mt）とは

我々ヒトを含む真核生物は例外なくミトコンドリアと呼ばれる細胞小器官（オルガネラ）を保持している。このオルガネラは、太古に真核生物の共通祖先と当時の真正細菌（バクテリア）の一種が細胞内共生することによって獲得された、という仮説が広く支持されている [1], [2]。このオルガネラの獲得は、真核生物に様々な恩恵をもたらし、その後の進化に大きな影響を与えたと考えられている。そのため、このオルガネラは進化学の分野において重要な研究対象として注目を集めてきた [3], [4]。

ミトコンドリアは二重の生体膜で包まれた構造であり、その内部の物質は細胞内の他の物質と明確に区画分けされている。そのためミトコンドリアの内部には様々な酵素（タンパク質の一種であり、特定の化学反応を触媒（促進）する）が細胞内に拡散することなく高濃度で存在することが可能である。また、これにより効率的に化学反応を触媒している。ミトコンドリアの内部に存在する酵素がミトコンドリアの機能を決定づけており、すなわちミトコンドリアの機能を知るためにはそこに存在する酵素を知る必要がある。

[1] 筑波大学大学院システム情報工学研究科
[2] 筑波大学大学院生命環境科学研究科
[3] 筑波大学大学院計算科学研究センター

1.2 Mt の機能を解析する手法

そのためには、ミトコンドリアを細胞から精製し、質量分析等の手法によってそこに含まれる全タンパク質を直接知る手法（プロテオーム解析）が考えられる。しかし、この手法は金銭的に高コストであり、また、先行研究が存在しない生物に適用する場合、ミトコンドリア精製プロトコルの条件検討からはじめる必要があるため、時間的にも人的にも高いコストがかかる。

そのため、ミトコンドリアを含む細胞を丸ごと処理し、細胞に含まれる全 RNA 情報を取得する方法が考えられる。全 RNA 情報の取得は上記の手法に比べて低コストかつ容易に行うことができ、タンパク質は RNA から翻訳されるため RNA の情報から間接的にどのようなタンパク質が存在するのかを知ることができる。ただし、細胞全体を処理したことから、この全 RNA 情報にはミトコンドリアタンパク質以外の情報も含まれるため、何らかの方法でミトコンドリアタンパク質の情報を分離する必要がある。この分離手法として、機械学習を用いた分類によるアプローチが考えられる [5]。

1.3 既存研究の問題点および本研究の目的

これまでにミトコンドリアタンパク質を予測する分類器は様々なものが提案されてきたが、そのほとんど全てがモデル生物というごく一部の生物で得られた知見をもとに特徴量抽出やデータセットの構築を行っていた。そのため、

Vol.2016-MPS-108 No.25
Vol.2016-BIO-46 No.25
2016/7/5

既存の分類器は典型的なミトコンドリアタンパク質の予測を念頭に置いたものであり、それ以外の生物がもつミトコンドリア及びその関連オルガネラタンパク質の予測精度には問題があった。これは広範な真核生物を対象に取るミトコンドリア進化研究を行うにあたって障害となっている。本研究では、モデル生物以外の、非モデル生物も対象に入れた分類器について検討する。

2. 関連研究

2.1 Mt タンパク質配列の特徴

タンパク質はアミノ酸が方向をもって一列に連なった構造をしている。そのためアミノ酸 1 つを 1 文字で表記すれば、タンパク質は文字列で表現される（図 1）。

```
>SequenceName1
MTIRNLDRLFKPSSIALIGASRHPQSIGQVVARNLFNAGF...
>SequenceName2
MGVLAYNSVAALPLTPDLAVIATPPQTFIPGLIAELIPGL...
...
```

図 1　タンパク質配列ファイル

はじめに述べたようにミトコンドリアへ輸送されるタンパク質とそうでないタンパク質は明確に区別されている。この区別のために、典型的なミトコンドリアタンパク質配列には特徴的な配列（シグナル配列）が利用されていることが知られている。

その特徴とは以下のようなものである [6]。

- シグナル配列はタンパク質配列の先頭に存在する
- その配列長は 10-80 文字である場合が多い
- 正に帯電しているアミノ酸を多く含む
- アルギニン (R) を多く含む
- ミトコンドリアに輸送後、切断される部位（切断サイト）がある
- 立体構造をとったとき、疎水性面、親水性面の両方が出現するらせん構造をとる
- .....

これらの特徴に注目すると、そのタンパク質がミトコンドリア及び関連オルガネラへ輸送されるか否かを、文字列で表現されるタンパク質配列情報を用いて判定する二値分類問題として考えられる。

2.2 典型的なミトコンドリアタンパク質予測ソフトウェア

ここでは関連研究として、PSORT II[7], TargetP[8]、そして 2016 年 6 月現在最も精度が高い予測ソフトウェアである Mitofates[9] の、3 つのソフトウェアを紹介する。

PSORT II[7] は Nakao らによって開発されたソフトウェアであり、トレーニングデータとして SwissProt データベースより取得した酵母菌のタンパク質配列 1531 レコードを用いており、各配列の先頭 30 番目までのアミノ酸出現頻度を特徴量として用いている。このレコードにはミトコンドリアに輸送されるか否かのラベルが付与されている。これを利用し、新たな配列が与えられたとき、k 近傍法に基づく分類を行う。

TargetP[8] は Emanuelsson らによって開発されたソフトウェアであり、PSORT II と同様に SwissProt から取得されたラベル付きタンパク質配列 3678 レコードをトレーニングデータとしている。各配列の先頭 130 番目までのアミノ酸出現頻度を特徴量とし、ニューラルネットワークによって学習及び分類を行う。

Mitofates[9] は Fukasawa らによって開発されたソフトウェアであり、上記 2 手法と同様 SwissProt から取得された植物、酵母菌のラベル付きタンパク質配列 7177 レコードをトレーニングデータとしており、サポートベクターマシン（SVM）によって学習及び分類を行う。

以下では、本研究に最も関連が深い Mitofates の概要について記述する。

2.3 Mitofates の概要

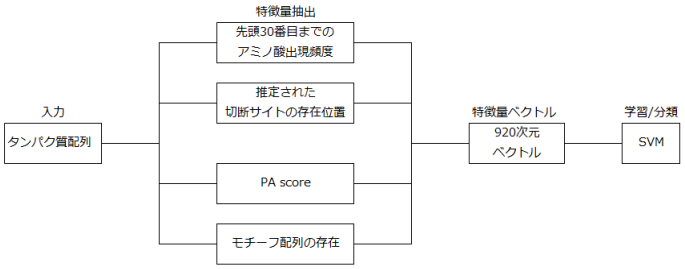

図 2　Mitofates の概要

Mitofates が用いる特徴量を以下に示す。

- 推定された切断サイトの位置
- 先頭 30 番目までのアミノ酸出現頻度
- PA score
- モチーフ配列の有無

これらのうち PA score およびモチーフ配列は以下のように定められる。

2.3.1 PA score

PA score は、疎水性面をもち、その反対側に正電荷を帯びた親水性面をもつらせん構造をとるような配列に、最も高いスコアを与えるようにした特徴量であり、以下の式で算出される。

$$PA = \frac{1}{n}(\sqrt{(\Sigma_i H_i \cos(\delta i))^2 + (\Sigma_i H_i \sin(\delta i))^2}$$

$$-r\cos\theta\sqrt{(\Sigma_i C_i \cos(\delta i))^2 + (\Sigma_i C_i \sin(\delta i))^2})$$

入力されたタンパク質配列の先頭 30 番目までを対象と

Vol.2016-MPS-108 No.25
Vol.2016-BIO-46 No.25
2016/7/5

n	ウィンドウの大きさ
r	重み係数
H_i	i 番目のアミノ酸の疎水性度
δ_i	定数
θ	疎水性、電荷モーメントベクトル間の角度
C_i	i 番目のアミノ酸の電荷 $(-1, 0, 1)$

して、10-20 の幅のウィンドウをスライドさせながら、全てのパターンにおいて PA を算出し、最も大きな PA を PA score とする。

2.3.2 モチーフ配列の有無

タンパク質を構成する 20 種類のアミノ酸を、その性質によって 5 群に分け、各々を別の記号で置き換え縮重表記する（表 1）。このとき入力されたタンパク質配列の先頭 30 文字中に、ミトコンドリアタンパク質に偏って出現する 14 組の 6 縮重アミノ酸配列 (モチーフ配列、例：$\phi\phi\sigma\beta\phi\phi$) のうち 1 つあるいは複数が存在した場合、その p-value P を用いて、

$$-\sum \log_{10}(P)$$

とした値を特徴量として用いる（表 2）。

表 1 アミノ酸の縮重表記

アミノ酸	縮重記号	性質
RKH	α	塩基性アミノ酸
ED	β	酸性アミノ酸
PG	γ	構造破壊性アミノ酸
LFIVWYMCA	ϕ	疎水性アミノ酸
STNQ	σ	中性極性アミノ酸

表 2 モチーフ配列例

モチーフ配列	p-value
$\phi\phi\sigma\beta\phi\phi$	5.7×10^{-13}
$\phi\phi\beta\sigma\phi\phi$	1.2×10^{-11}
$\phi\phi\phi\sigma\beta\phi$	1.1×10^{-9}
$\phi\phi\sigma\beta\phi\beta$	1.8×10^{-9}
$\phi\beta\phi\phi\beta\gamma$	6.1×10^{-9}
$\beta\phi\phi\sigma\sigma\sigma$	9.3×10^{-8}

3. 提案手法

本研究では、一部のモデル生物を対象に開発されてきたミトコンドリアタンパク質予測ソフトウェアをベースに、非モデル生物を対象としたミトコンドリア及びその関連オルガネラのタンパク質予測手法の開発を目的とする。

3.1 非モデル生物における Mt タンパク質の予測

まず最初に既存の手法を利用し、トレーニングデータセットを非モデル生物のデータセットに更新することによって、非モデル生物がもつタンパク質配列を対象とした場合の予測精度を検証した。既存の手法としては現在最も予測精度が高いソフトウェアである Mitofates を利用した。

3.1.1 データセット

Uniprot データベースより、非モデル生物のポジティブデータ（ミトコンドリア局在とされているタンパク質配列）を 398 レコード、ネガティブデータ（ミトコンドリア局在ではないとされているタンパク質配列）を 7071 レコード取得した。

3.1.2 特徴量抽出とトレーニング

Mitofates のアルゴリズムを利用して、トレーニングデータセットから特徴量を抽出した（参考：図 2）。(1) この特徴量をスケーリングし、libsvm3.2[10] を用いてサポートベクターマシン (SVM) によるトレーニングを行った。カーネル関数は RBF カーネルを用い、パラメータはデフォルトのまま行った。(2) この特徴量を用いて、R[11] のライブラリの一つである xgboost[12] を用いて、勾配ブースティング (GBM) によるトレーニングを行った。パラメータは木の数を最適化し、その他はデフォルトのまま行った。

3.1.3 精度の評価

作成した分類器の精度は 10-fold cross-validation により評価した。また、より詳細な評価を行うために ROC 曲線を描き、ROC AUC を算出した。また GBM でトレーニングしたものについては、Precision Recall 曲線も描いた。

3.1.4 実験結果

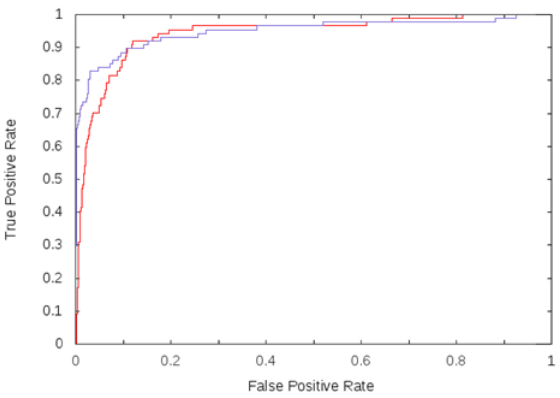

図 3 非モデル生物における Mt タンパク質の予測 (SVM)：ROC 曲線

表 3 ROC AUC

分類器のトレーニングデータ	ROC AUC
モデル生物データ	0.9409
非モデル生物データ	0.9479

作成した非モデル生物における Mt タンパク質の分類器についてその性能を検証した結果、10-fold cross-validation accuracy は 94.7%となった。また、ROC 曲線（図 3、青線：モデル生物データによるトレーニングを行った分類器、

赤線：非モデル生物データによるトレーニングを行った分類器)、及び ROC AUC（表 3）の評価から、特にパラメータチューニングを施さずとも、モデル生物をトレーニングデータとした分類器と同等の高い精度で分類できることが分かった。また、GBM によるトレーニングを行った場合 (図 4, 5)、ROC AUC が 0.9719 となり、さらに良い性能を示した。

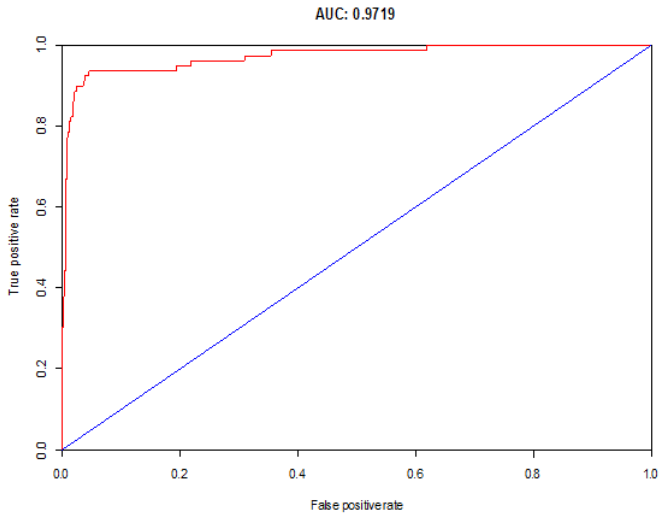

図 4　非モデル生物における Mt タンパク質の予測 (GBM)：ROC 曲線

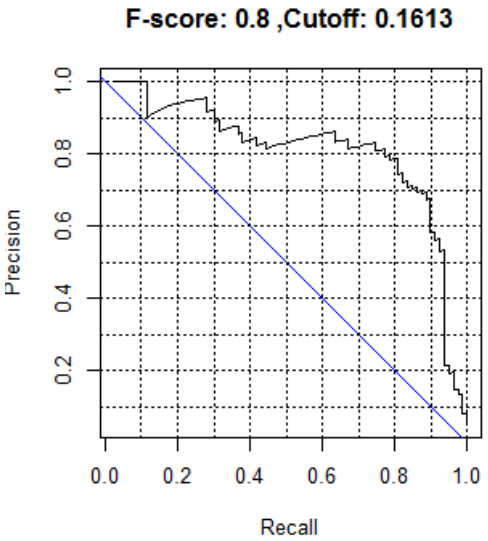

図 5　非モデル生物における Mt タンパク質の予測 (GBM)：PR 曲線

3.2　退化的な Mt をもつ生物における Mt タンパク質の予測

3.2.1　退化的 Mt タンパク質の予測精度

退化的なミトコンドリアをもつ生物におけるミトコンドリアタンパク質の予測において、既存のソフトウェアや、本研究においてミトコンドリアをもつ非モデル生物のデータをもとに作成した分類器で精度を検証したところ、共に 20%前後と低い値を示し、問題があることが分かった (表 4)。そこで、退化的なミトコンドリアを持つ生物に限定した分類器を作成した。

表 4　退化的な Mt をもつ生物における Mt タンパク質の予測精度

分類器のトレーニングデータ	Accuracy
モデル生物データ	18.8%
非モデル生物データ	25.7%

3.2.2　データセット

GiardiaDB 及び TrichoDB より、退化的ミトコンドリアタンパク質のポジティブデータ（退化的ミトコンドリア局在とされているタンパク質配列）を 102 レコード、ネガティブデータ（退化的ミトコンドリア局在ではないとされているタンパク質配列）を 271 レコード取得した。

3.2.3　特徴量抽出とトレーニング

Mitofates のアルゴリズムを利用して、トレーニングデータセットから特徴量を抽出した（参考：表 2)。(1) この特徴量をスケーリングし、libsvm3.2 を用いて SVM によるトレーニングを行った。カーネル関数は RBF カーネルを用い、パラメータはデフォルトのまま行った。(2) この特徴量を用いて、R ライブラリの xgboost を用いて、GBM によるトレーニングを行った。パラメータは木の数を最適化し、その他はデフォルトのまま行った。

3.2.4　精度の評価

作成した分類器の精度は 4-fold cross-validation により評価した。また、より詳細な評価を行うために ROC 曲線を描き、ROC AUC を算出した。また GBM でトレーニングしたものについては、Precision Recall 曲線も描いた。

3.2.5　実験結果

作成した退化的 Mt をもつ生物における Mt タンパク質の分類器についてその性能を検証した結果、(図 6,7)、ROC AUC が 0.9491 となり、十分に実用的な性能を示した。

3.3　新たな特徴量の探索

近年、退化的なミトコンドリアをもつ生物はミトコンドリアタンパク質の輸送にシグナル配列を利用していないという可能性を示唆する報告が Zimorski、Garg らによって行われており [13][14]、従来のものとは異なる輸送経路が利用されている可能性がある。従って、このようなミトコンドリアタンパク質の予測に際して、既存の特徴量抽出アルゴリズムを適用すべきでなく、新たな特徴量を検討する余地があると思われる。そこで、退化的なミトコンドリアタンパク質において、そうでないタンパク質との間にアミノ酸出現頻度に偏りがあるか否か、あるいはモチーフ配列が存在するか否かの探索を行った。

Vol.2016-MPS-108 No.25
Vol.2016-BIO-46 No.25
2016/7/5

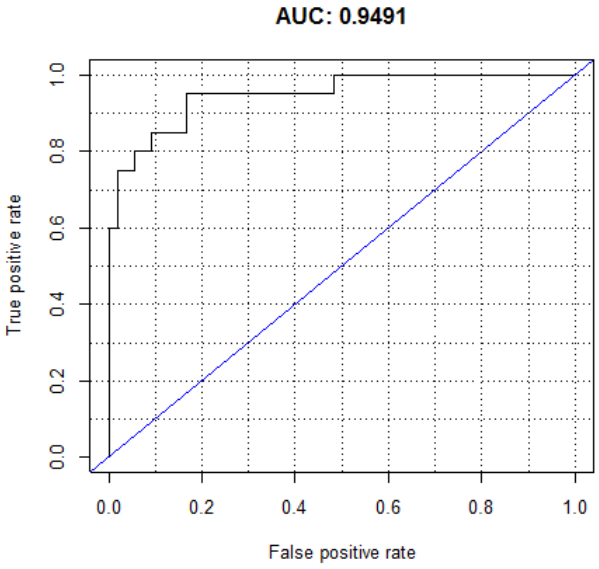

図 6　退化的 Mt タンパク質の予測 (GBM)：ROC 曲線,

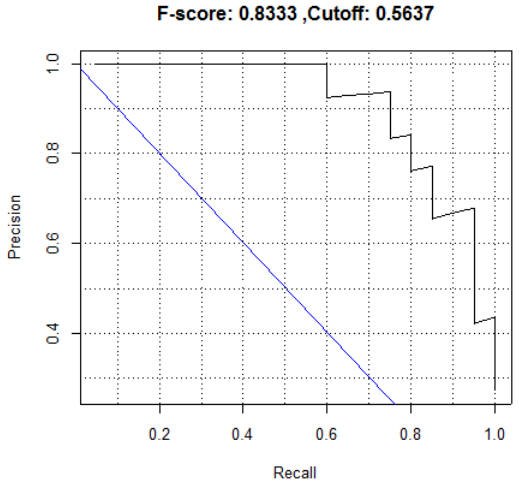

図 7　退化的 Mt タンパク質の予測 (GBM)：PR 曲線,

3.3.1　データセット

GiardiaDB 及び TrichoDB より、退化的ミトコンドリアタンパク質のポジティブデータ（退化的ミトコンドリア局在とされているタンパク質配列）を 102 レコード、ネガティブデータ（退化的ミトコンドリア局在ではないとされているタンパク質配列）を 271 レコード取得した。

3.3.2　モチーフ配列の探索

データセットの全タンパク質配列について、表 1 に従ってタンパク質配列を縮重アミノ酸で置換した。そして 6 縮重アミノ酸の全順列 (例：$\alpha\alpha\alpha\alpha\alpha\alpha$, $\alpha\alpha\alpha\alpha\alpha\phi$…) について、それぞれを含むか否かの対応表を作成した。そして chi square test を行い、多重検定補正法 LAMP (Limitless Arity Multiple testing Procedure) [15] で p-value を補正した。そして $p < 0.05$ となった配列を退化的なミトコンドリアタンパク質に特有のモチーフ配列であるとみなした。

また、フェニルアラニン (F) およびシステイン (C) について、典型的なミトコンドリアにおいて輸送に関わるモチーフに多く含まれていることが示唆されているため、この 2 アミノ酸を縮重アミノ酸に置換せずに、それぞれ同様にモチーフ配列を探索した。

その結果、図に示すモチーフ配列及びその組み合わせが、退化的ミトコンドリアタンパク質に偏って出現することが示唆された (表 5, 6)。C については、検出されたモチーフ配列は存在しなかった。

表 5　検出されたモチーフ配列

Rank	Adjusted p-value	Combination		
1	0.00135	$\alpha\phi\phi\phi\phi\gamma$		
2	0.00782	$\phi\phi\phi\phi\alpha\phi$	$\alpha\phi\phi\phi\phi\gamma$	
3	0.01044	$\phi\gamma\phi\phi\sigma$	$\alpha\phi\phi\phi\phi\gamma$	
4	0.01044	$\phi\sigma\phi\gamma\phi\phi$	$\phi\phi\phi\phi\alpha\phi$	$\alpha\phi\phi\phi\phi\gamma$
5	0.01044	$\phi\phi\phi\phi\alpha\phi$	$\phi\phi\gamma\phi\sigma\phi$	$\alpha\phi\phi\phi\phi\gamma$
6	0.01044	$\alpha\alpha\phi\phi\phi\phi$	$\alpha\phi\phi\phi\phi\gamma$	
7	0.01044	$\alpha\phi\alpha\phi\alpha\phi$	$\alpha\phi\phi\phi\phi\gamma$	
8	0.01044	$\gamma\phi\sigma\phi\phi\beta$		
9	0.02879	$\phi\sigma\phi\gamma\phi\phi$	$\phi\phi\phi\phi\alpha\phi$	

表 6　検出されたモチーフ配列 (F)

Rank	Adjusted p-value	Combination		
1	0.001735	$\phi\theta F\theta\phi\theta$	$\phi\phi\phi\phi F\theta$	$\theta\phi\phi\phi F$
2	0.001735	$\phi\phi\theta F\theta\phi$	$\phi\phi\phi\phi F\theta$	$\theta\phi\phi\phi F$
3	0.035203	$\phi\theta F\theta\phi\theta$	$\phi\phi\phi\phi F\theta$	
4	0.035203	$\phi\phi\phi\theta F\theta$	$\phi\phi\phi\phi F\theta$	$\theta\phi\phi\phi\phi F$
5	0.035203	$\phi\phi\phi F\theta\theta$	$F\theta\theta\theta\phi\theta$	
6	0.035203	$\phi\phi\theta F\theta\phi$	$\theta\phi\phi\phi\phi F$	
7	0.035203	$\phi\phi\theta F\theta\phi$	$\phi\phi\phi F\theta\theta$	

3.3.3　アミノ酸出現頻度の偏りの検証

既存手法はタンパク質配列の先頭 30 アミノ酸のみを出現頻度の計算対象としている。しかし、配列の内部または末尾にシグナル配列が存在する可能性が実験的に示唆されており [13][14]、従って典型的な mt タンパク質の特徴をもっていないと考えられるため、特徴量抽出において先頭 30 アミノ酸に限定する根拠は無いと考えられる。そこでタンパク質配列全体を対象としてアミノ酸出現頻度の偏りを検証した。このとき、タンパク質の機能が類似していれば、共通する配列が存在する可能性が高く、それによって誤った結論が導かれる可能性がある。例えば、Heat shock protein ファミリーのタンパク質が退化的ミトコンドリアタンパク質に多いことが知られているため、この機能に関わる部位のアミノ酸組成が反映される可能性がある。従って、MOTIF SEARCH[16] によって、機能をもつ領域と判定されなかった部分配列のみを用いた。そして、機能領域除去処理後のデータセットに含まれる各アミノ酸の数を集計し (表 7)、chi square test を行った。

その結果、$p < 2.2 \times 10^{-16}$ となり、退化的ミトコンドリアタンパク質とそうでないタンパク質の間に差があることが示唆された。

表 7 退化的ミトコンドリア/非ミトコンドリアタンパク質のアミノ酸出現頻度の集計結果

	A	C	D	E	F	G	H
Mt	1439	203	972	1045	485	1236	242
non-Mt	9828	1680	6744	8183	3135	5731	2493

	I	K	L	M	N	P	Q
Mt	1016	1093	1426	364	611	572	588
non-Mt	5856	6463	11251	3025	4890	4896	4991

	R	S	T	V	W	Y	Total
Mt	716	1198	967	1188	54	316	15731
non-Mt	6878	9765	6876	6294	618	3297	112894

4. 終わりに

本研究では、モデル生物で得られた知見を基にした特徴量抽出アルゴリズムを用いても、非モデル生物のミトコンドリアタンパク質を実用的な精度で予測することができる可能性を示した。同様に、一部の退化的なミトコンドリアをもつ生物がもつミトコンドリアタンパク質についても、データセットを更新した上で、SVM によるトレーニングではなく、GBM によるトレーニングを行うことによって既存の特徴量アルゴリズムを用いても、実用的な精度を示す分類器が作成できる可能性を示した。また、退化的ミトコンドリアタンパク質の予測に用いることのできる新たな特徴が存在することを示唆した。

今後は、本稿で提案した分類器の実証実験を行い、既存の予測ソフトウェアに対する優位性を検証していく予定である。また、新たに発見した特徴を特徴量抽出アルゴリズムに組み込み、精度向上に寄与するかどうかを検証する予定である。

最後に、退化的ミトコンドリアタンパク質の予測について、既存の特徴量抽出アルゴリズムでは対応できないことが分子生物学的な見地から予見されたものの [13][14]、それに反して十分な精度が示された (図 4,5)。この点について、各特徴量の重要度や主成分分析を通して生物学的な解釈が与えられるかどうか検討していく予定である。

参考文献

[1] Andersson et al., "The genome sequence of Rickettsia prowazekii and the origin of mitochondria.", Nature, 1998.
[2] Fitzpatrick et al., "Genome Phylogenies Indicate a Meaningful α-Proteobacterial Phylogeny and Support a Grouping of the Mitochondria with the Rickettsiales.", Molecular Biology and Evolution, 2005.
[3] Pittis & Gabaldn, "Late acquisition of mitochondria by a host with chimaeric prokaryotic ancestry", Nature, 2016.
[4] Koonin, "The origin and early evolution of eukaryotes in the light of phylogenomics", Genome Biology, 2010.
[5] Klein et al, "The detection and classification of membrane-spanning proteins.", Biochimica et Biophysica Acta, 1985.
[6] Pfanner & Geissler, "Versatility of the mitochondrial protein import machinery.", Nature Reviews Molecular Cell Biology, 2001.
[7] Nakao et al, "Improvement of PSORT II Protein Sorting Prediction for Mammalian Proteins", Genome Informatics, 2002.
[8] Emanuelsson et al, "Locating proteins in the cell using TargetP, SignalP and related tools.", Nature Protocols, 2007.
[9] Fukasawa et al, "MitoFates: improved prediction of mitochondrial targeting sequences and their cleavage sites.", Molecular & cellular proteomics, 2015.
[10] C.-C. Chang and C.-J. Lin, "LIBSVM : a library for support vector machines.", ACM Transactions on Intelligent Systems and Technology, 2011.
[11] R Core Team, "R: A language and environment for statistical computing", R Foundation for Statistical Computing, 2015.
[12] Tianqi Chen, Carlos Guestrin, "XGBoost: A Scalable Tree Boosting System.", ACM Transactions on Intelligent Systems and Technology, 2016.
[13] Zimorski et al, "The N-terminal sequences of four major hydrogenosomal proteins are not essential for import into hydrogenosomes of Trichomonas vaginalis.", The Journal of eukaryotic microbiology, 2013.
[14] Garg et al, "Conservation of Transit Peptide-Independent Protein Import into the Mitochondrial and Hydrogenosomal Matrix.", Genome Biology and Evolution, 2015.
[15] Terada et al, "Statistical significance of combinatorial regulations.", Proceedings of the National Academy of Sciences, 2013.
[16] http://www.genome.jp/tools/motif/.