厚生労働省科学研究費補助金　食品の安全確保推進研究事業
「ゲノム情報および機械学習を用いた腸管出血性大腸菌の高精度型別モデルの構築」
(20KA3002)
研究分担報告書

研究分担者　伊澤　和輝（東京工業大学　情報理工学院）

研究要旨

腸管出血性大腸菌の高精度なサーベイランスを実現するためには、従来法である MLVA 型よりも詳細かつ安価で迅速な類別法が必要である。本研究では、腸管出血性大腸菌の株のペアについて MLVA 型の差異から SNP 数を予測・類別指標とすることによりこれを実現し、高精度なサーベイランスに役立てることを目指す。

本報告期間では、前年度までに得られていた O157 の機械学習モデル構築に用いた知見を利用し、O26、O111 の MLVA データを用いた学習・予測を試みた。O26 については 585 株、O111 については 285 株のデータを用いて、それぞれ約 17 万ペア、約 4 万ペアの MLVA 型のデータを用いた。

前年度までの結果から、機械学習アルゴリズムとして勾配ブースティング法を使用した。学習・予測の方針として、2 株間の SNP 数を連続値で予測する場合と、近縁株判定の指標である SNP 数 10 以下のペアか否かを予測するカテゴリの予測の場合を比較した。結果として、カテゴリの予測の場合の方が、連続値の予測の場合よりも精度が高かった。

今後は本研究で作成した予測モデルを実際に新規株の近縁株判別に使用し、予測モデルの有用性について検討する。

A. 研究目的

腸管出血性大腸菌（enterohemorrhagic Escherichia coli: EHEC）は、国内で年間 3,000 名以上の感染者が報告され、毎年数名の死者が報告されている公衆衛生上重要な食中毒菌である。そのため、発生源の特定や伝播経路を明らかにするために、高精度なサーベイランス法が必要とされている。

従来のサーベイランスで用いられている分子型別手法（反復配列多型解析法：MLVA 法）はゲノム中に存在する複数のリピート配列のパターンによって菌株を型別する手法であり、迅速かつ安価であるが、ゲノム中の特定部分だけを用いるため、型別能には限界がある。一方、高精度なサーベイランスを実現する手法として、全ゲノム情報を用いた単一塩基多型（SNP）解析が存在するが、高い型別能を有する一方で迅速性や費用面で従来法に劣っている。

本研究では、MLVA 型および菌株情報から、全ゲノムレベルの型別情報を推測するモデルを、人工知能の一種である機

械学習を用いて構築することを目指す。

B. 研究方法

2013 年から 2021 年に分離された EHEC O26 の 585 株、O111 の 285 株についての MLVA 型データと任意の 2 株間の SNP 数のデータ（それぞれ約 17 万ペア、約 4 万ペア）を研究代表者の李謙一氏から提供いただいた。

任意の 2 株間の SNP 数のデータのうち、25%を機械学習モデルの評価用として分割し、残りの 75%を機械学習モデルの構築用のデータとして用いた。

予測結果として、各株ペア間の SNP 数を直接計算する連続値の予測と、各株ペアが 10 SNP または 20 SNP を閾値とした場合に近縁株であるか否かを予測するカテゴリの予測を行った。

機械学習モデルの構築には東京工業大学が有するスーパーコンピューターである TSUBAME 3.0 の環境を利用した。

連続値予測の最適化関数には平均二乗誤差（squared error）、カテゴリ予測の最適化関数には逸脱度（deviance）を用いた。

C. 研究結果

1. O26 株ペアの SNP 数を連続値で予測する機械学習モデル

O26 の任意の株ペアにおいて、2 株の各 MLVA 座位データ（34 座位）、*stx1,2* 遺伝子の有無を、分離日間隔を特徴量として用い、勾配ブースティング回帰木のアルゴリズムを利用して機械学習モデルを構築した。この結果を図 1 に示す。連続値の予測においては、二乗平均平方根誤差（RMSE）は 44.8 となり、これは O26 血清型内の各ペアの SNP 数の実測値に対し平均的に±45 ヶ所程度増減した予測が行われていることを表す。

また、連続値予測において、近縁株の基準を 10 SNP 以内とした場合の混同行列を図 2 に示す。

再現率（Recall）（赤字）は、実測値から近縁株と判定される株ペアのうち、どの程度を予測から近縁株と判定できるかを表した数値であり、本研究で最も重要視している数値である。

O26 においての連続値予測においては再現率は 61.4%となり、O157 において Clade を分けて学習・予測した場合に比べて再現率が低かった。

2. O26 株ペアを近縁株か否かのカテゴリで予測する機械学習モデル

O26 血清型の任意の 2 株において、2 株の各 MLVA 座位データ（34 座位）、*stx1,2* 遺伝子の有無、分離日間隔を特徴量として用い、勾配ブースティング決定木のアルゴリズムを利用して機械学習モデルを構築した。

近縁株の基準を 10 SNP、20 SNP とした場合の混同行列を図 2 に示す。

カテゴリの予測においてはどちらの近縁株基準においても連続値の予測の場合よりも再現率が上昇していた。

3. O26 株ペアの SNP 数を連続値で予測する機械学習モデル

O111 の任意の株ペアにおいて、2 株の各 MLVA 座位データ（34 座位）、*stx1,2* 遺伝子の有無を、分離日間隔を特徴量として用い、勾配ブースティング回帰木のア

ルゴリズムを利用して機械学習モデルを構築した。この結果を図 3 に示す。連続値の予測においては、二乗平均平方根誤差（RMSE）は 38.7 となり、これは O111 血清型内の各ペアの SNP 数の実測値に対し平均的に±39 ヶ所程度増減した予測が行われていることを表す。

また、連続値予測において、近縁株の基準を 10 SNP 以内とした場合の混同行列を図 4 に示す。

O111 においての連続値予測において再現率は 27.7%となり、O157 において Clade を分けて学習・予測した場合や O26 において学習・予測した場合に比べて著しく再現率が低かった。

4. O111 株ペアを近縁株か否かのカテゴリで予測する機械学習モデル

O111 血清型の任意の 2 株において、2 株の各 MLVA 座位データ（34 座位）、*stx1,2* 遺伝子の有無、分離日間隔を特徴量として用い、勾配ブースティング決定木のアルゴリズムを利用して機械学習モデルを構築した。

近縁株の基準を 10 SNP、20 SNP とした場合の混同行列を図 2 に示す。

カテゴリの予測においてはどちらの近縁株基準においても連続値の予測の場合よりも再現率が上昇していた。

D. 考察

O26・O111 においては株ペアの SNP 数を連続値で予測する機械学習モデルの場合、O157 で Clade を分けて 予測した場合よりも予測精度が悪かった。これは、O157 においては Clade を分けることで、Clade 間ペアの SNP 数が大きいと思われるデータを学習・予測データから除くことができたことに起因している。現状では O26・O111 血清型においてはこのような Clade の別はなく、図 1 からもわかるように 2 株間で SNP 数が 1000 前後の遠縁株間データが含まれている状態である。また O26・O111 血清型においては、O157 今後に比べて既存のデータが少ないことも予測精度低下の原因と考えられる。今後、O26・O111 血清型のデータの追加により部分的に機械学習・予測の精度が上がることが期待される。

一方、カテゴリの予測においては、O26・O111 においても 75%以上の再現率が見られた。こちらの予測では、近縁株か否かの学習・予測のため、データセットの少なさ、また遠縁株が含まれたデータセットであっても、今回用いた特徴量から学習・予測が十分に可能であったと考えられる。

E. 結論

本研究では、2013 年から 2021 年に分離された国内 EHEC O26・O111 血清型株についての MLVA 型データと任意の株ペアの SNP 数のデータから、MLVA 座位、*stx1,2* 遺伝子の有無、分離日間隔を特徴量として株ペアの SNP 数及び近縁株か否かのカテゴリを予測する機械学習モデルの作成を行った。

連続値の学習・予測においては、近縁株の予測がうまくいかない部分が見られたが、カテゴリでの学習・予測においては、どちらの血清型においても精度良く近縁株を予測することができた。そのため、今後はカテゴリでの予測モデルを用いて、

本機械学習モデルの実応用性を検討したい。

F. 健康危険情報

なし

G. 研究発表

1) 誌上発表

なし

2) 学会発表

なし

H. 知的財産権の出願・登録状況

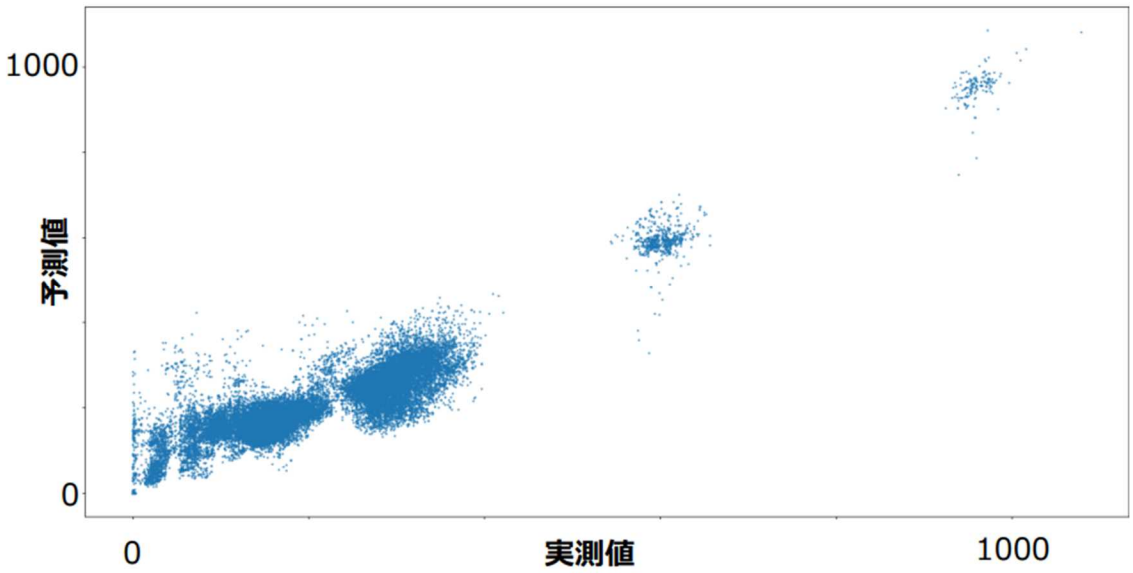
1. 特許取得

なし

２. 実用新案登録

なし

図 1. O26 についての連続値予測の機械学習モデルの予測結果



RMSE=44.8

図 2. O26 についての連続値･カテゴリ予測の機械学習モデルの予測結果（混同行列）

赤: 再現率（Recall）、青:適合率（Precision）

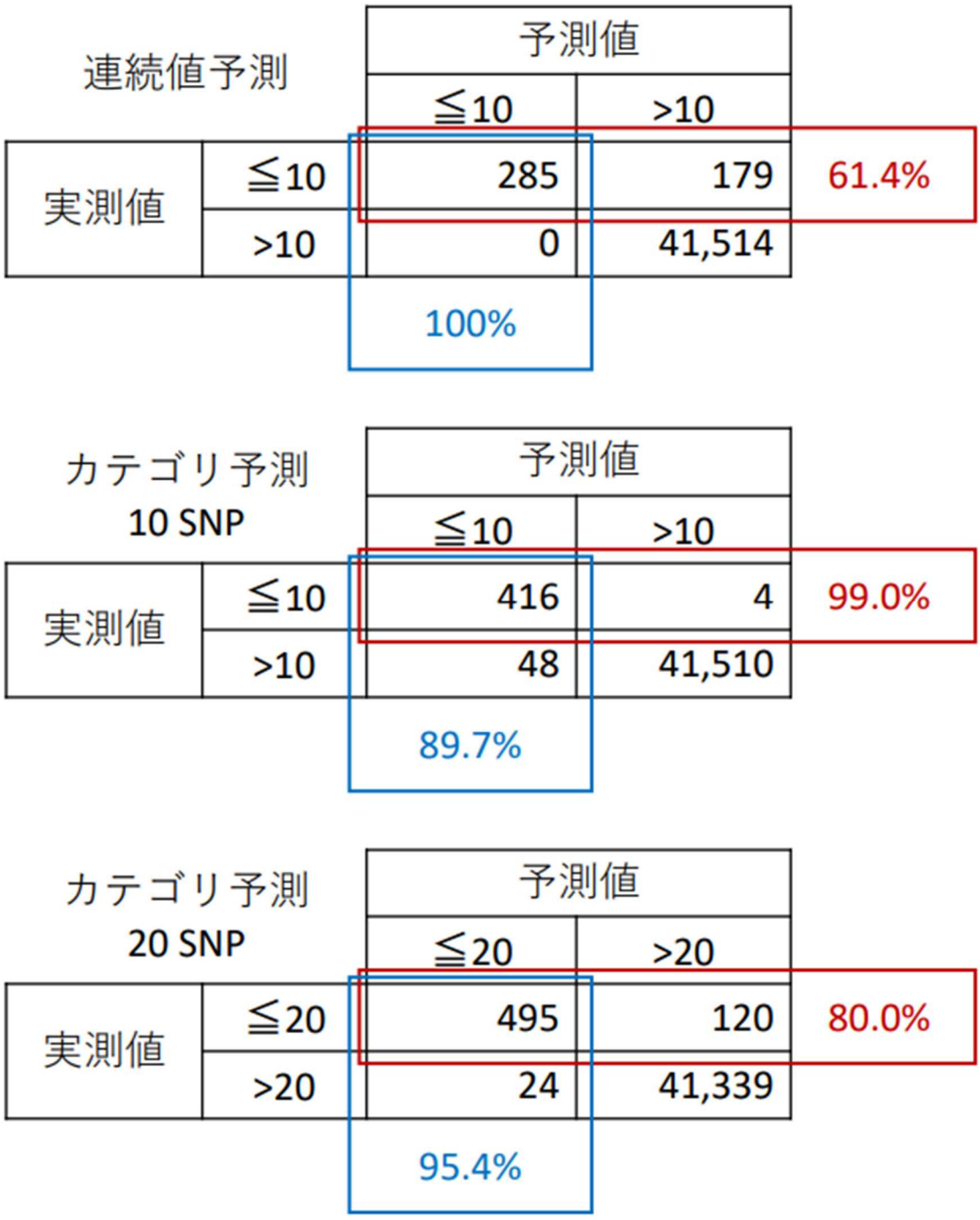
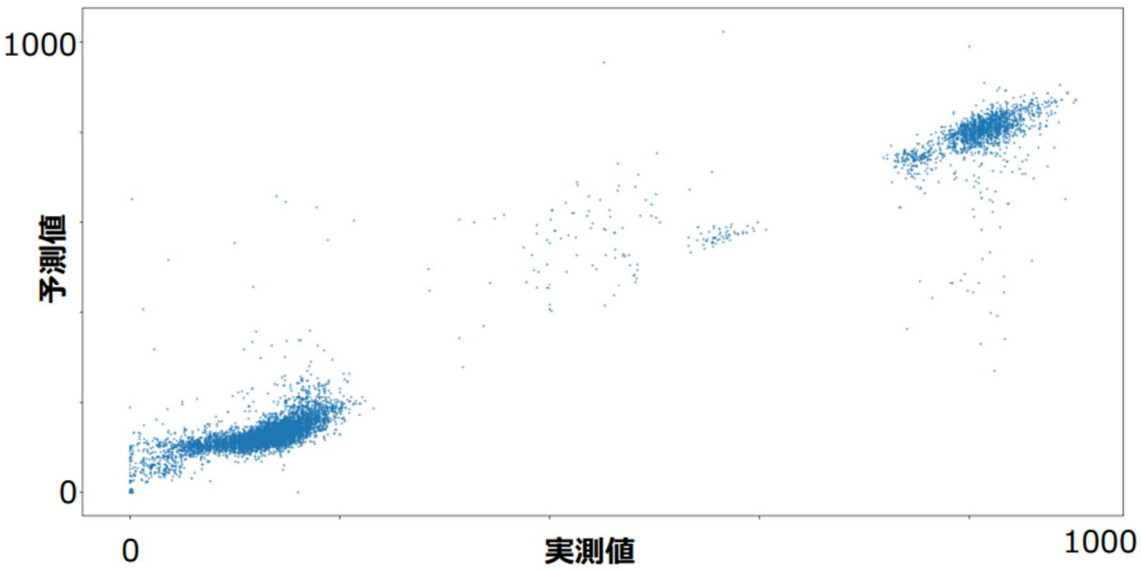


図 3. O111 についての連続値の機械学習モデルの予測結果



RMSE=38.7

図 4. O111 についての連続値・カテゴリ予測の機械学習モデルの予測結果（混同行列）

赤: 再現率（Recall）、青:適合率（Precision）