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 890 株に加え、年度内に 746 株のデータを追加し、合計 1636 株のデータを用いた。これらの株のペア（約 130 万ペア）の MLVA 型のデータを各 Clade に分割し、各ペアの SNP 数を予測することを試みた。

前年度の結果から、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 O26、O111 の MLVA データを用いた学習・予測を同様の枠組みで行い、本研究で探索した機械学習の枠組みの汎用性について議論する。

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

O157 の 1636 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（約 130 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、Clade 2、3、7、8 の各 Clade 内のペアのみを抽出した。各 Clade において、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. 株ペアの SNP 数を連続値で予測する機械学習モデル

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、分離地の緯度・経度情報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、Clade 2 では二乗平均平方根誤差（RMSE）は 3.8 となり、これは直感的には Clade 2 内の各ペアの SNP 数の実測値に対し±4 ヶ所程度増減した予測が行われていることを表す。同様に Clade 3 では RMSE が 4.8、Clade 7 では RMSE が 35.6、Clade 8 では RMSE が 4.9 となった。

また、近縁株の基準を 10 SNP、20 SNP とした場合の混同行列を図 2、3 に示す。

再現率（Recall）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。

Clade 2、3、8 でも、近縁株の基準を 10 SNP から 20 SNP に広げると再現率が上昇していた。一方、Clade 7 では他の Clade に比べて著しく再現率が低かった。

2. 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

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、分離地の緯度・経度情報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 4、5 に示す。

カテゴリの予測においてはどの Clade 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おり、特に連続値の予測では難しかった Clade 7 における再現率が著しく上昇た。

D. 考察

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、Clade 7 での予測精度が他の Clade に比べて悪かった。これは、Clade 7 のデータセットには他の Clade

ではそれほど多くない 200 SNP 以上の株ペアデータが多かったことが原因であると考えられる。連続値の予測においては、株ペアデータ全体に対して SNP 数の予測が最適化されるため、200 SNP 以上のペアの学習・予測にあった最適化がなされることになる。この結果、RMSE が 36 程度と大きくなり、近縁株の閾値を大きく超えたため、近縁株の予測精度が悪かったと考えられる。

一方、カテゴリの予測では Clade 7 においても 60%以上の再現率が見られた。こちらの予測では、近縁株か否かの〇×問題を解く学習・予測のため、データセットの中で 1%以下の近縁株についても、今回用いた特徴量から学習・予測が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 O157、1636 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、分離地の緯度・経度情報を特徴量として株ペアの SNP 数を予測する機械学習モデルの作成を試みた。

連続値の学習・予測においては、特に Clade 7 において、SNP 数の大きい株ペアのデータに学習・予測全体が影響を受け、近縁株の予測がうまくいかない部分が見られた。

一方で、カテゴリでの学習・予測においては、Clade 7 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に注力したソフトウェアの開発を進める。

また今後は、O157 以外で主要な血清型である、O26、O111 の MLVA データを用いた学習・予測を、O157 で得られた知見を用いて行い、本研究で行っている SNP 数の機械学習での予測が他血清型でも応用可能かどうかについて議論したい。

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

MLVA結果と機械学習モデルを用いた腸管出血性大腸菌の遺伝的距離の予測
伊澤和輝、李謙一、泉谷秀昌、伊豫田淳、大西真、明田幸宏
（第42回日本食品微生物学会学術総会・2021年9月21日（火）～10月20日（水））

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. 連続値予測の機械学習モデルの予測結果

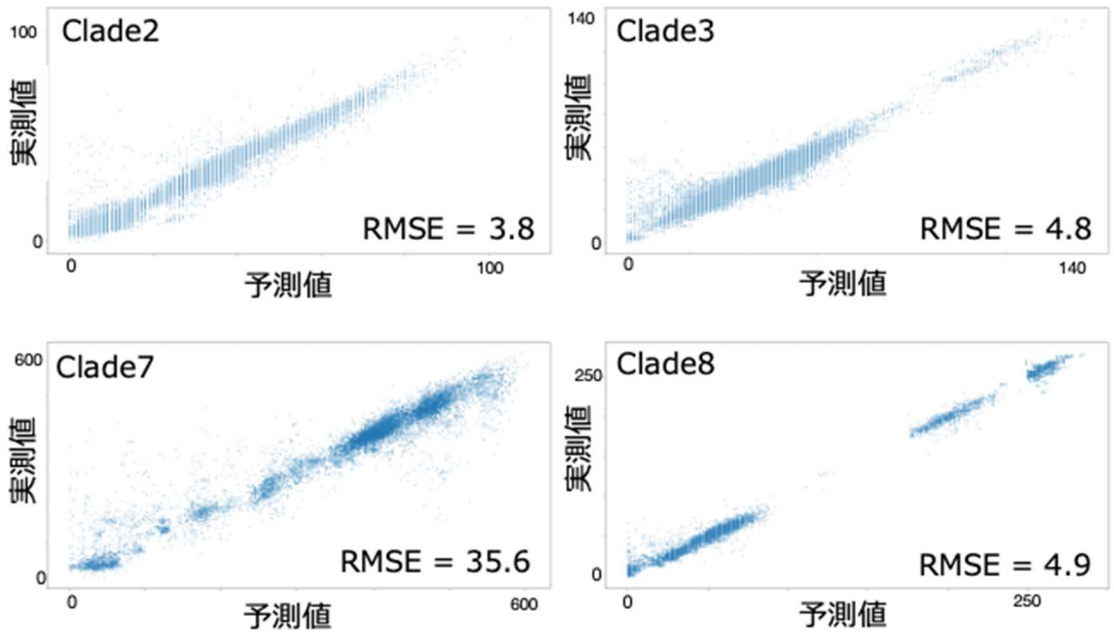

図 2. 連続値予測の機械学習モデルの予測結果（閾値 10 SNP の混同行列）

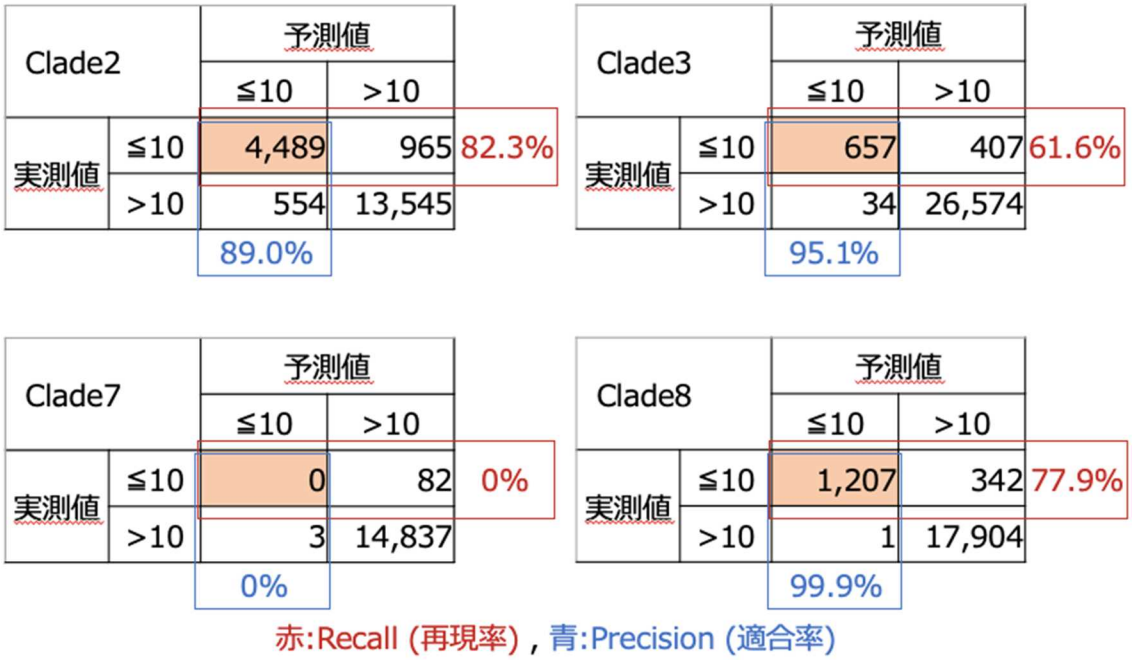

図 3. 連続値予測の機械学習モデルの予測結果（閾値 20 SNP の混同行列）

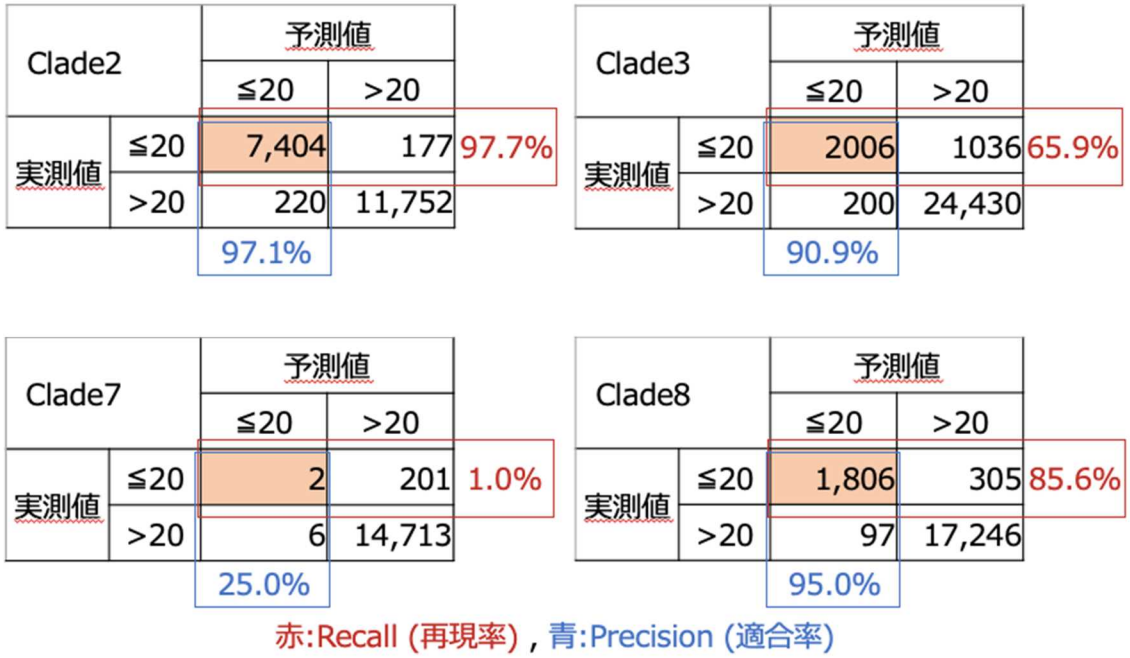

図 4. カテゴリ予測の機械学習モデルの予測結果（閾値 10 SNP の混同行列）

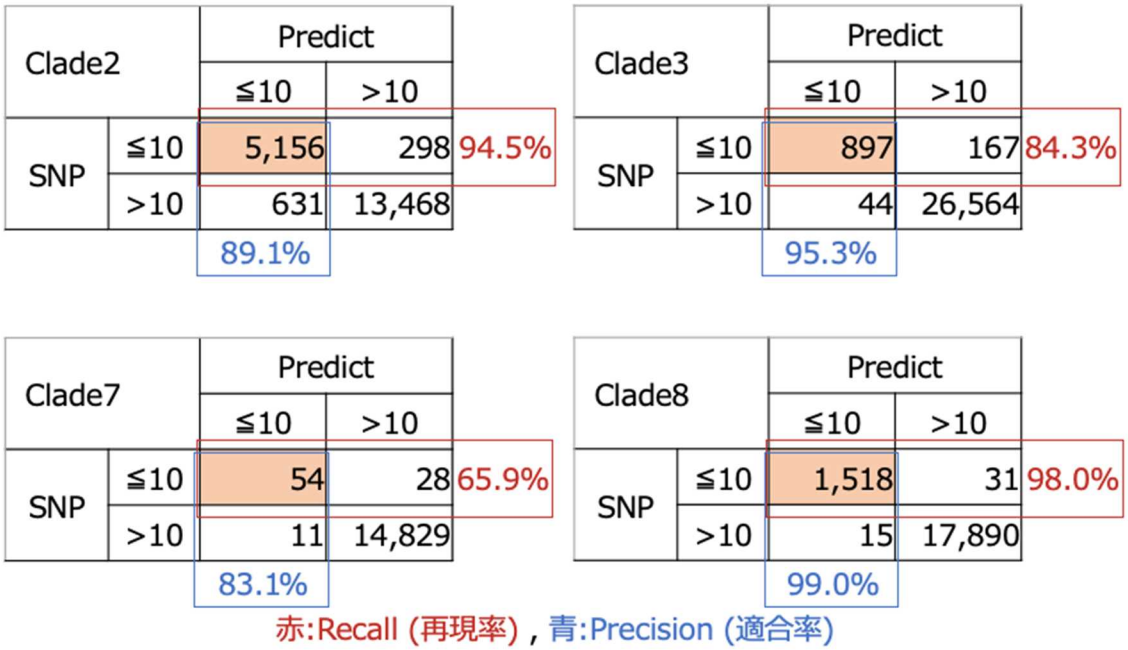

図 5. カテゴリ予測の機械学習モデルの予測結果（閾値 20 SNP の混同行列）