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

分担研究課題　「機械学習モデルの構築・評価」
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

本研究では、腸管出血性大腸菌のあるペアにおいて MLVA 型の差異から SNP 数を予測・類別指標とすることによって、MLVA によるサーベイランスを高精度化することを目指した。

本報告期間では、TSUBAME3.0 上に機械学習モデルの構築に必要な環境を整え、現在までに得られている 882 株、約 40 万ペアの MLVA 型のデータ全体から、各ペアの SNP 数を予測することを試みた。機械学習モデルには線形回帰モデル、回帰木モデル、勾配ブースティング回帰木を使用したが、勾配ブースティング回帰木モデルが最も精度が良く、R 二乗値が 0.8 以上の機械学習モデルを作成することができた。一方で、近縁株判定の指標となる SNP 数が 10 以下のペアについての予測精度が悪かった。これは SNP 数が 10 以下のペアがデータ全体の 3.7%程度しかないことに由来すると考えられる。

今後、MLVA 座位以外の株のデータ（毒素産生遺伝子の有無、分離地、clade など）を予測のための特徴量としてモデルに組み込むことで、SNP 数が 10 以下のペアの精度向上を試みる。

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 2019 年に分離された EHEC O157 の約 882 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 360 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。

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 TSUBAME3.0 の環境を利用した。

機械学習モデルの最適化指標には RMSE（二乗平均平方根誤差）を用いた。

C. 研究結果

1. 各 MLVA 座位の差異の有無を学習データに用いた機械学習モデル

任意の 2 株間の SNP 数のデータで指定された 2 株において、各 MLVA 座位の差異の有無（17 座位）を特徴量として用い、回帰木および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。

回帰木のアルゴリズムを利用した結果を図 1 に示す。学習時のパラメーターとして回帰木の深さを深さ 2、深さ 5 を利用した。結果、深さ 2 の回帰木モデルでは、RMSE が 91.1、深さ 5 の回帰木モデルでは RMSE が 69.0 となった。これは直感的には各ペアの SNP 数の実測値に対し、深さ 2 の回帰木では 91 個、深さ 5 の回帰木モデルでは 69 個程度、SNP 数がずれた予測を行なっていることを示している。

また勾配ブースティング回帰木のアルゴリズムを利用した結果を図 2 に示す。学習時のパラメーターとして、勾配ブースティング回帰木の深さ 3 を利用した。RMSE が 61.0、R^2 値は 0.87 となり、回帰木のアルゴリズムを用いた場合よりも予測精度が向上した。

2. 各 MLVA 座位データを学習データに用いた機械学習モデル

任意の 2 株間の SNP 数のデータで指定された 2 株において、2 株の各 MLVA 座位データ（34 座位）を特徴量として用い、線形回帰および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。

線形回帰のアルゴリズムを利用した場合、実測値と予測値の関係は図 3 のようになり、RMSE 値は 88.8 となった。

また勾配ブースティング回帰木のアルゴリズムを利用した結果を図 4 に示す。学習時のパラメーターとして、勾配ブースティング回帰木の深さ 3 を利用した。RMSE が 22.9、R^2 値は 0.98 となり、線形回帰のアルゴリズムを用いた場合よりも予測精度が向上した。

D. 考察

各 MLVA 座位の差異の有無を特徴量として用いた場合、および各 MLVA 座位データを特徴量として用いた場合の両者において勾配ブースティング回帰木を用いた場合が最も精度が良かった。これは線形回帰や回帰木に比べ、勾配ブースティング回帰木のアルゴリズムが MLVA 型か

らの SNP 数の予測に適している可能性を示唆する。

また、各 MLVA 座位の差異の有無では特徴量として 17 座位のデータのみを用いていたが、各 MLVA 座位データを用いた場合では任意の 2 株の MLVA 座位全て（17 座位×2 ＝ 34 座位）を用いることで予測精度が向上したと考えられる。各 MLVA 座位データを用いた場合の予測において、特徴量の重要度を比較すると、2 株間で対称的でない部分があり、MLVA 座位間に相互的な関係性が存在する可能性がある。

本年度では最終的に RMSE が 22.9、R^2 値が 0.98 となる高精度な予測モデルの作成に成功したが、これは各株ペアの全体に対する予測精度である。本研究では、今後、近縁株の指標として 2 株間の SNP 数の差異が 10SNP 以下とするが、10SNP 以下の株のペアのみに着目した場合には予測精度は 3 割程度となり、十分な予測精度とは言えなかった。この予測精度は任意の 2 株の SNP 数のデータのうち 96%程度が 10SNP 以上のペアのデータであることに由来すると考えられる。そのため、今後近縁株の予測精度を向上させるためには、MLVA 以外の指標から近縁株候補を絞り込み、その中で学習モデルを作成する必要があると考えられる。近縁株候補の絞り込みに用いる指標としては、毒素産生遺伝子の有無、分離地、clade などが考えられる。

E. 結論

本研究では、2013 年から 2019 年に分離された国内 EHEC O157 の 890 株の MLVA 型・および任意の 2 株の SNP 数データから、MLVA 座位のデータを特徴量として 2 株間の SNP 数を予測する機械学習モデルの作成を試みた。

勾配ブースティング回帰木のアルゴリズムを用いた機械学習モデルは R^2 値で 0.98 を示し、高精度なモデルとなった。このことは MLVA 型と 2 株間の SNP 数の間に非線形の関係性があることを示唆している。

一方で、本機械学習モデルの実用化のためには SNP 数が 10 以下のペアを正確に予測する必要がある。本研究で得られた機械学習モデルは、SNP 数が 10 以下のペアだけに着目した場合、予測精度が不十分であった。

今後、近縁株の予測精度の向上のため、MLVA 型以外の特徴量を機械学習モデルに組み込み、SNP 数が 10 以下のペアだけに特化した機械学習モデルの作成が必要である。

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. 各 MLVA 座位の差異の有無を特徴量として回帰木を利用したモデル

横軸は SNP 数の実測値、縦軸は SNP 数の予測値を示す。

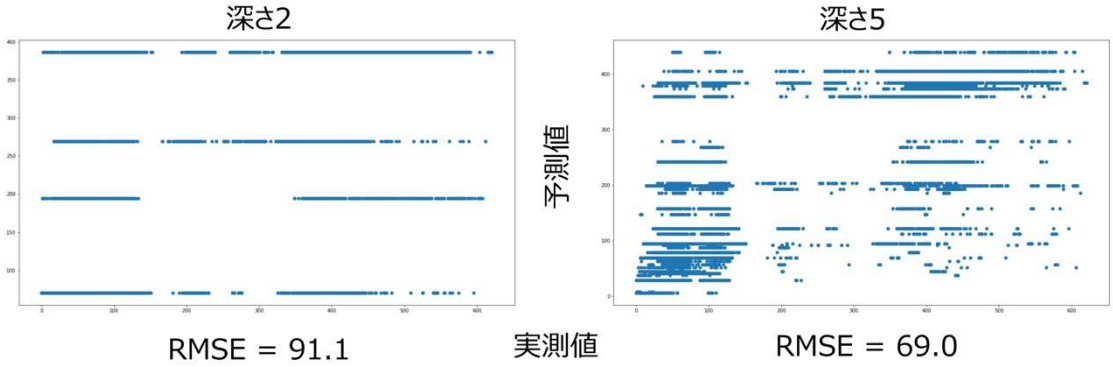

図 2. 各 MLVA 座位の差異の有無を特徴量として勾配ブースティング回帰木を利用したモデル

横軸は SNP 数の実測値、縦軸は SNP 数の予測値を示す。

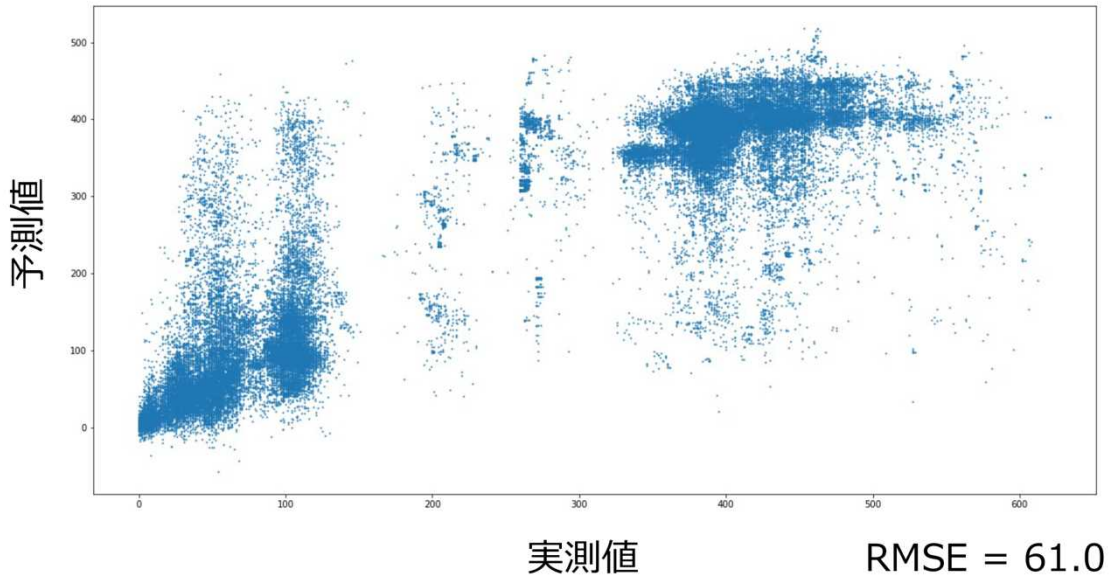

図 3. 各 MLVA 座位を特徴量として線形回帰を利用したモデル

横軸は SNP 数の実測値、縦軸は SNP 数の予測値を示す。

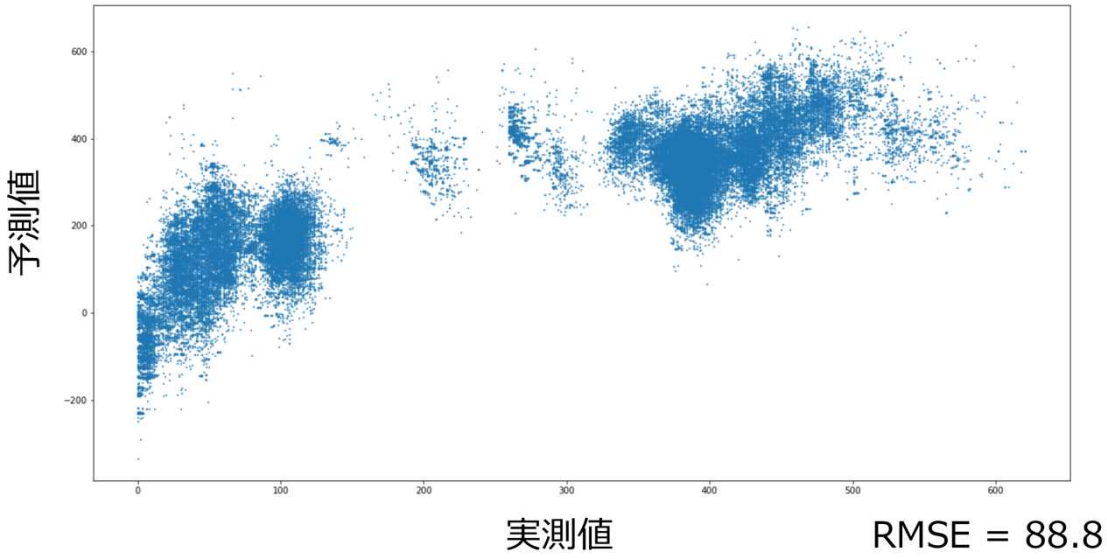

図 4. 各 MLVA 座位を特徴量として勾配ブースティング回帰木を利用したモデル

横軸は SNP 数の実測値、縦軸は SNP 数の予測値を示す。