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）
研究総括報告書

研究代表者　李　謙一（国立感染症研究所　細菌第一部）

研究要旨

現在、腸管出血性大腸菌（enterohemorrhagic *Escherichia coli*：EHEC）のサーベイランスでは主に multi locus variable tandem repeat analysis（MLVA）が用いられている。本研究では、MLVA を用いたサーベイランスの精度を向上するために、機械学習モデルを用いて SNP の予測を試みた。まず、国内 EHEC O157 計 1636 株の全ゲノム配列（whole-genome sequence：WGS）解析を行い、単一塩基多型（single nucleotide polymorphism：SNP）と MLVA との関連性を解析した。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。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。さらに、菌株間の SNP が 5 または 10 か所以内の株をクラスター化し、重症化率等を計算するプログラムを Perl にて作製した。次に、O157 で構築したモデルしたモデルを EHEC O26 および O111 についても適用した。すなわち、両血清型のゲノム情報を新たに取得し、両血清型での機械学習モデルの構築および評価を行った。この結果、O157 に比べると精度は下がるものの、75%以上の再現度で近縁株の抽出が可能となった。さらに、これまでに構築した 3 血清型（O157、O26、および O111）における機械学習モデルの評価を行った。その結果、いずれの血清型でも MLVA 単独で近縁株の抽出を行った場合よりも、敏感度（SNP で 10 以内のペアを「近縁株」として検出する割合）の顕著な増加が認められた。

**研究分担者**
李 謙一（国立感染症研究所 細菌第一部）
伊澤和輝（東京工業大学 情報理工学院）

## 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る公衆衛生上重要な食中毒菌である。EHEC 感染症は胃腸炎症状を主徴とし、時として血便や急性腎不全である溶血性尿毒症症候群を引き起こし、毎年数名の死者が報告されている。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。

現在、国内分離株の 95%以上を占める主要 8 血清群（O157, O26, O111 など）では、反復配列多型解析 (multilocus variable-number tandem-repeat analysis: MLVA) 法を用いたサーベイランスが、国立感染症研究所を中心に行われている。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、全ゲノム情報を用いた単一塩基多型（single nucleotide polymorphism: SNP）解析は、高い型別能を有するが、迅速性や費用面で劣るため、当面は MLVA 法を用いたサーベイランスが主流であり続けると考えられる。

そこで本研究では、従来のサーベイランスで用いられている MLVA 法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的とした。

**B. 研究方法**

各分担研究報告書に記載。

**C. 研究結果**

**1. 国内 EHEC の WGS 解析およびモデルの評価**

研究代表者 李 謙一の分担研究として、国内で 2014 年から 2021 年に分離された EHEC O157 計 1,636 株の SNP 解析を行った。さらに国内で 2013 年から 2021 年に分離された EHEC O26 の 585 株および O111 の 285 株について SNP 解析を行った。これらの株について、総当たりのペアを作製し、SNP 距離等を計算し、機械学習用のデータを作製した。また、2 年度目および 3 年度目に構築したモデルの評価を行った。この結果、敏感度（SNP で 10 以内のペアを「近縁株」として検出する割合）の顕著な増加が認められた。

**2. 機械学習モデルの構築および評価**

研究分担者 伊澤和輝の分研究として、研究代表者 李が作成した SNP データセットを用いた機械学習モデルの構築を行った。モデルとしては、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。入力データとしては、MLVA 型の差異数、各座位でのリピート数、分離日間隔を用い、出力データとしては SNP 数とした。この結果、勾配ブースティング回帰木モデルで精度の良い（$R^2$ 値が 0.8 以上）機械学習モデルを作製が可能であった。さらに精度を向上させるために、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結果、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。また、clade 2,3,および 8 では、80%以上の再現性で近縁株を予測できることが明らかとなった。加えて、EHEC O26 および O111 の SNP データセットを用いた機械学習モデルの構築を行った。モデルとしては、O157 で用いたものと同様の勾配ブースティング回帰木を使用した。この結果、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。いずれの血清型においても、再現度が 75%以上となり、高精度に近縁株を推定することが可能であった。

なし

**D. 考察**

EHEC O157 における MLVA と SNP の関連性の解析で、両者は経時的に変化しており、単純な線形回帰ではないことが明らかとなった。機械学習モデル（勾配ブースティング回帰木）を利用した SNP 予測を行ったところ、$R^2$ 値が 0.98 となるモデルを作製することができた。以上の結果から、SNP の予測には機械学習モデルが有効であることが明らかとなった。さらに、clade の細分類後に SNP の予測をすることで、著しく精度の向上が認められることが明らかとなった。各 clade での精度では、clade 7 で精度が比較的低かったが、これは同 clade では近縁株が比較的少なく、学習が十分でなかったことが原因として考えられる。

O26 および O111 では、O157 のモデル構築で用いた clade のような細分類は存在しないため、O157 に比べて推定の精度は低かった。しかし、MLVA 単独で近縁株を予測する場合に比べて、より多くの近縁株を抽出することが可能であった。

**E. 結論**

本研究では、EHEC O157、O26、および O111 を対象に SNP 予測を目的とした機械学習モデルを構築し、MLVA 結果から、ゲノムレベルでの近縁株を抽出することが可能となった。今後サーベイランスで本モデルを活用しながら精度を改善させることが望ましいと考えられた。

**F. 健康危険情報**

**G. 研究発表**

1) 誌上発表

なし

2) 学会発表

1. 伊澤和輝, 李 謙一, 泉谷秀昌, 伊豫田 淳, 大西 真, 明田幸宏. 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, 第 42 回日本食品微生物学会学術総会
2. 李 謙一. 腸管出血性大腸菌の全ゲノム解析法について. 第 34 回 地方衛生研究所全国協議会 関東甲信静支部細菌研究部会. 横浜, 2023.
3. 泉谷秀昌, 李 謙一, 伊豫田 淳, 明田幸宏. 腸管出血性大腸菌の MLVA による分子疫学解析. 第 43 回日本食品微生物学会学術総会. 東京, 2022.
4. 李 謙一. 全ゲノム配列解析を用いた腸管出血性大腸菌サーベイランスとクラスター検出事例 衛生微生物技術協議会 42 回研究会. Web, 2022.
5. 泉谷秀昌, 李 謙一, 伊豫田 淳, 大西 真. 2021 年に分離された腸管出血性大腸菌の MLVA 法による解析. 2022. Infectious Agents Surveillance Report 43:108-109.

**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