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。研究初年度に構築した機械学習の精度を向上させるために、前年度までに得られていた O157 の 890 株に加え、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。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。

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 O157 1,636 株の WGS 解析およびクラスター検出プログラムの開発

研究代表者 李 謙一の分担研究として、国内で 2020 年から 2021 年に分離された EHEC O157 192 株の WGS を新たに解読し、国立感染症研究所・細菌第一部で既に解読済みのデータと合わせ、計 1,636 株の SNP 解析を行った。さらに、クラスター化された株について病原性等の情報を自動的に得られるプログラムによって、集団感染等が起こった際の危険度を予測することが可能になると考えられる。

2. 機械学習モデルの精度向上

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、勾配ブースティング回帰木を使用した。MLVA 型のデータを各 Clade ごとに分割し、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。

D. 考察

モデル構築の際には、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。今後、本モデルでの近縁株予測精度について、従来の方法（主に MLVA 型のみで判断）との差異を検証する必要がある。

E. 結論

本研究では、SNP 予測を目的とした機械学習モデルの改善を行った。今後は、実際の集団感染事例を対象に解析やモデルの改善を行うことで、本モデルの実用化を目指す。

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