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 882 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。この結果、単純な線形回帰では MLVA から SNP を予測することは困難であることが判明した。そこで、複数の機械学習モデルを試行した結果、勾配ブースティング回帰木モデルを用いることによって、R^2 値が 0.98 となるモデル構築が可能となった。今後は、実用化に向け系統情報などを加えることで精度の向上を目指す予定である。

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 882 株の WGS 解析

研究代表者 李 謙一の分担研究として、国内で 2014 年から 2020 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 882 株の SNP 解析を行った。この結果、大部分の株間では SNP と MLVA の結果は正の相関関係を示したが、大きなばらつきが認められた。さらに、MLVA に加えて分離日間隔を含めても、SNP を線形的に予測することは困難であり、機械学習等のより複雑なモデル化が必要であることが明らかとなった。

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。

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。

E. 結論

本研究では、SNP 予測を目的とした機械学習モデルを構築した。実用化に向けて、病原性遺伝子や系統情報をモデルに加えるなどを行って、精度の向上を図る予定である。

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