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「O26 および O111 菌株の全ゲノム解析および機械学習モデルの評価」

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

機械学習モデルを構築するために、腸管出血性大腸菌（enterohemorrhagic *Escherichia coli*：EHEC） O26 および O111 において、それぞれ 585 株および 285 株の全ゲノム配列解析を行い、単一塩基多型（single nucleotide polymorphism : SNP）を抽出した。また、分担研究者が構築したモデルの評価を行い、O157、O26、および O111 のいずれにおいても敏感度（SNP で 10 以内のペアを「近縁株」として検出している割合）の顕著な増加が認められた。以上の結果から、本研究の機械学習モデルは、「MLVA での差異がある程度あるがゲノムレベルでは近縁な株」を効率よく抽出可能であることが示された。

A. 研究目的

腸管出血性大腸菌（enterohemorrhagic *Escherichia coli*: EHEC）の全国サーベイランスでは、現在反復配列多型解析（multi locus variable tandem repeat analysis：MLVA）法が用いられている。これまでに EHEC O157 を対象にした、MLVA 法と全ゲノム配列（whole-genome sequence : WGS）解析法との比較では、MLVA 法は短期間の集団感染調査には十分高い型別能を有することが示されている。しかしながら、MLVA 型が 2 座位以上異なる株間では、近縁な株と遠縁な株が混在していることが明らかとなっている。

そこで本研究では、機械学習によって MLVA および菌株情報から菌株間の距離を推定するモデルの作製を目的とした。本課題の 1、2 年度目では EHEC O157:H7 を対象とするモデルの作製を行い、良好な成績が得られたため、同血清型に次ぐ発生数が報告される O26:H11 および O111:H8 についても同様の解析を行った。

また、1，2 年度目に解析した O157 の結果を含めて、3 血清型における機械学習モデルの評価を行った。

B. 研究方法

2013 年から 2021 年に分離された EHEC O26 の 585 株、O111 の 285 株についての全ゲノム配列解析を行った。一部の菌株は、ゲノム DNA 抽出を行い、QIAseq FX DNA Library Kit（QIAGEN）を用いてライブラリー調製を行った。作製したライブラリーを使用して、HiSeqX（illumina）に

よってペアエンドシークエンシング（150-mer×2）を行った。SNP 抽出は、BactSNP および snippy などを用いた解析パイプラインを用いて行い、Gubbins によって組換え領域の検出・削除を行った。

機械学習モデルの評価として、近縁株を検出する能力を敏感度、特異度、陽性的中率、および陰性的中率の 4 種の指標を用いた。近縁株の定義としては、O157 の MLVA では 1 アリール以内の差異、O26 および O111 では同一の MLVA 型、を用いた。機械学習モデルでは、最も成績の良かった 10 か所以内・11 か所以上のカテゴリー分けデータを用いた。

C. 研究結果

O26 の 585 株、O111 の 285 株について、全株総当たりのペアを作製し、各ペアでの SNP 数および MLVA で異なる座位数を算出し、モデル構築に用いた。

機械学習の評価では、O157 では MLVA のみで近縁株を抽出した場合、敏感度以外の指標は 0.95 以上と非常に高い値を示した（表 1）。一方、敏感度（SNP で 10 以内のペアを「近縁株」として検出している割合）は比較的低い値（0.61）であった。機械学習の結果を用いると、clade 7 以外の敏感度は 0.88 以上となり、より多くの近縁株の検出が可能であった。また、実際のサーベイランスにおける運用の際には、MLVA で 1 段階目の近縁株の抽出を行い、次に機械学習によって 2 段階目の抽出を行うと考えられる。そこで、MLVA 結果と機械学習結果を組み合わせた際の、敏感度などの指標を計算した。その結果、いずれの clade においても敏感度の値が向上した。

次に、O26 および O111 において O157 と同様の評価を行った。これらの血清型では、MLVA の結果のみを用いる場合には敏感度の値が低く、O26 では 0.51、O111 では 0.16 であった。しかし、機械学習モデルを適用することによって、それぞれ 0.90 および 0.78 に向上した。その他の指標は、MLVA および機械学習に関わらず 0.95 以上と高値であった。

D. 考察

本研究では、過去 2 年間に構築した機械学習モデルを用いて近縁株の抽出を行った際の敏感度などの指標を評価した。その結果、機械学習モデルを用いることで、敏感度の顕著な増加が認められた。この結果は、「MLVA での差異がある程度あるがゲノムレベルでは近縁な株」を効率よく抽出出来ていることを示している。詳細な機序は不明であるが、差異のある MLVA アリールやリピート数の違いによって、ゲノムレベルでの遺伝的距離を推測していると考えられる。Clade 7 では敏感度の上昇は認められなかったが、MLVA 結果と組み合わせることによって、成績の向上が可能であった。実際のサーベイランスでは、そのような従来法（O157 では MLVA で 1 か所以内の差異）と組み合わせた運用がなされると考えられるため、実用的にも高い型別能を有すると考えられた。

E. 結論

構築した機械学習モデルは EHEC O157、O26、および O111 のいずれにおいても精

度よく近縁株を抽出できることが判明した。

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

1. **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.
2. 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.
3. **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.
4. 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

表 1. O157 における MLVA と機械学習（ML）結果との比較

Stats	MLVA 1 アリール以内	Clade							
		2		3		7		8	
		ML のみ	MLVA＋ML	ML のみ	MLVA＋ML	ML のみ	MLVA＋ML	ML のみ	MLVA＋ML
敏感度	0.61	0.88	0.90	0.95	0.97	0.62	0.89	0.99	1.00
特異度	1.00	0.93	0.93	1.00	0.99	1.00	1.00	1.00	0.99
陽性的中率	0.97	0.82	0.82	0.98	0.96	0.90	0.87	1.00	0.99
陰性的中率	0.98	0.95	0.96	0.99	0.99	1.00	1.00	1.00	1.00

表 2. O26 および O111 における MLVA と機械学習（ML）結果との比較

Stats	O26		O111	
	MLVA	ML	MLVA	ML
敏感度	0.51	0.90	0.16	0.78
特異度	1.00	1.00	1.00	1.00
陽性的中率	0.95	0.99	1.00	0.95
陰性的中率	0.99	1.00	1.00	1.00