別紙１

厚生労働科学研究費補助金

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

総合研究報告書

研究代表者　李 謙一

令和５（２０２３）年　４月

別紙２

目　　次

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

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「EHEC 菌株の全ゲノム解析および MLVA との比較」
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

機械学習の基礎となるデータを得るために、腸管出血性大腸菌（enterohemorrhagic *Escherichia coli* : EHEC）O157、O26、および O111 においてそれぞれ計 1,636 株、585 株、および 285 株の全ゲノム配列から単一塩基多型（single nucleotide polymorphism : SNP）を抽出した。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。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孝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。そこで本分担研究では、機械学習に供するための EHEC O157、O26、および O111 の WGS 解読を行い、各菌株間の遺伝的距離を単一塩基多型（single nucleotide polymorphism）にて算出した。得られた結果と MLVA 結果を比較し、機械学習の基礎となるデータを得た。また、従来のように MLVA のみで近縁株を抽出した場合と比べた際と、機械学習モデルを用いた際との結果を比較することで、同モデルの評価を行った。

B. 研究方法

2014 年から 2021 年に分離された EHEC O157 384 株について、ゲノム DNA 抽出を行い、Nextera XT DNA Library Prep Kit（illumina）または 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。得られたショートリードは、これまでに感染研・細菌第一部で既に解読したデータと合わせ、計 1,636 株で解析を行った。O26 および O111 では、それぞれ 585 株および 285 株についての全ゲノム配列解析を行った。一部の菌株については上記の方法で新たに WGS を解読した。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。

また、モデル構築に用いた株のデータを用いて、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。

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

計 1636 株の WGS 解析を行い、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た。さらに、*in silico* で clade を決定した（表 1）。過去の同様の解析では、MLVA での差異が 1 か所以内の株間では少数の SNP のみ存在することが示されている。今回の解析は、散発事例株が含まれるため、SNP のばらつきはより大きく表れた。MLVA での差異が 2 座位以内の場合には、cgSNP の中央値は 10 以内に収まり、近縁な株が大部分であった（図 1）。しかし、MLVA が同一でも SNP が 400 か所以上存在する株や、MLVA の差異が 11 か所存在する場合にも、SNP が 8 か所である株が存在した。経時的な SNP の蓄積速度を調べるために、MLVA の差異ごとに SNP と分離日の間隔を用いて、回帰分析を行った。その結果、異なる MLVA 座位数が大きくなるにつれ、回帰式の傾きが小さくなる傾向が認められ、5 座位が異なる株間では相関は認められなくなった。

また、機械学習にて近縁株を抽出した後に、病原性や国内での分布を予測するための Perl プログラムを作製した。本プログラムでは、まず SNP 情報に基づいて 5 か所または 10 か所以内の株同士をクラスター化する。クラスター化された株について、菌株情報をもとに重症化率（溶血性尿毒症症候群および血便の割合）、無症状保菌の割合、分離地の中央値、最小値、および最大値を算出した。結果例を表 2 に示す。本プログラムによって、機械学習モデルによって近縁株を抽出した後に、関連株の病原性等を予測することが可能となった。

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 3）。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（表 4）。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

国内株の O157 の SNP 解析データをさらに蓄積し、機械学習の基礎となるデータを得た。これまでのデータでは、集団感染株や関連する MLVA 型の株の割合が高かったが、本研究では散発事例株も含む株の解析を行った。この結果、MLVA と SNP の相関関係は先行研究と同様に認められたが、例外的な株（MLVA で類似しているが多数の SNP が存在する、または MLVA での差異が大きいが少数の SNP のみ存在する）が多数認められた。これらの株については、差異が存在する MLVA の座位やリピート数についてより詳細に検討する必要がある。また、異なる MLVA 座位数別に経時的な SNP の蓄積を回帰分析で解析したところ、強い相関は認められなかった。さらに、異なる MLVA 座位数が大きくなるにつれて、SNP と分離日間隔の相関性が弱くなる傾向が認められた。これは、経時的に MLVA 型も変化しているためと考えられる。つまり、分離日が数年離れている同一型のケースは、特殊な事例（冷凍保存食品など）の影響が強く出ている可能性がある。このことから、SNP 数は分離日と MLVA 型の差異数から単純に予測することはできず、多変量解析や機械学習等のより複雑なモデルによる予測が必要と考えられた。O157 は、遺伝的に多様であり、複数の亜系統（clade）に分けられる。本研究では、国内分離株の 9 割以上を占める clade 2, 3, 7 および 8 を対象にして機械学習モデルを構築することで、解析の精度を高めることが可能であった。

さらに、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。

また、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。

E. 結論

本研究では、国内 EHEC O157、O26、および O111 において、MLVA 結果からゲノムレベルで近縁な株を高精度で抽出するプログラムを構築した。本モデルは、散発事例株や地理的に離れた株の関連性を推定するうえで有用となると考えられる。

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

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

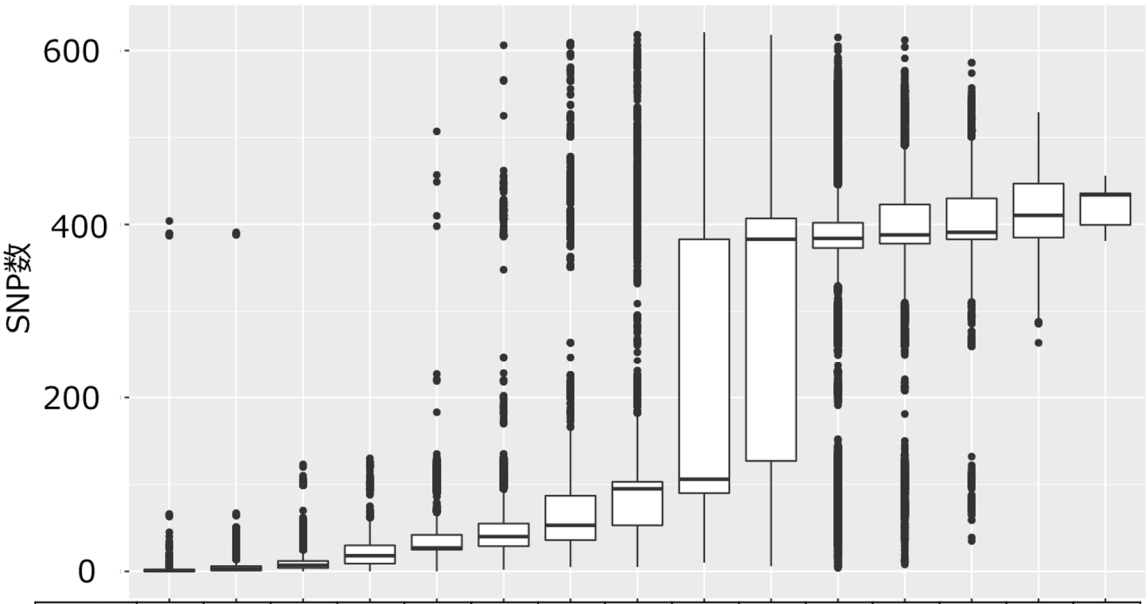
なし

２. 実用新案登録

なし

図 1. MLVA と SNP の関連性

MLVA の異なる座位数別に見た SNP の分布を箱ひげ図で示す。



MLVA差異	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Median	1	3	7	18	27	40	53	95	106	383	384	388	391	410.5	434
Min	0	0	0	0	0	2	5	5	10	6	4	8	35	264	381
Max	404	391	123	130	507	606	609	618	621	618	615	612	586	529	456

図 2. 異なる MLVA の座位数別にみた SNP と分離日間隔の関連性

各カラムの右上に回帰式および決定係数（R^2）を示す。

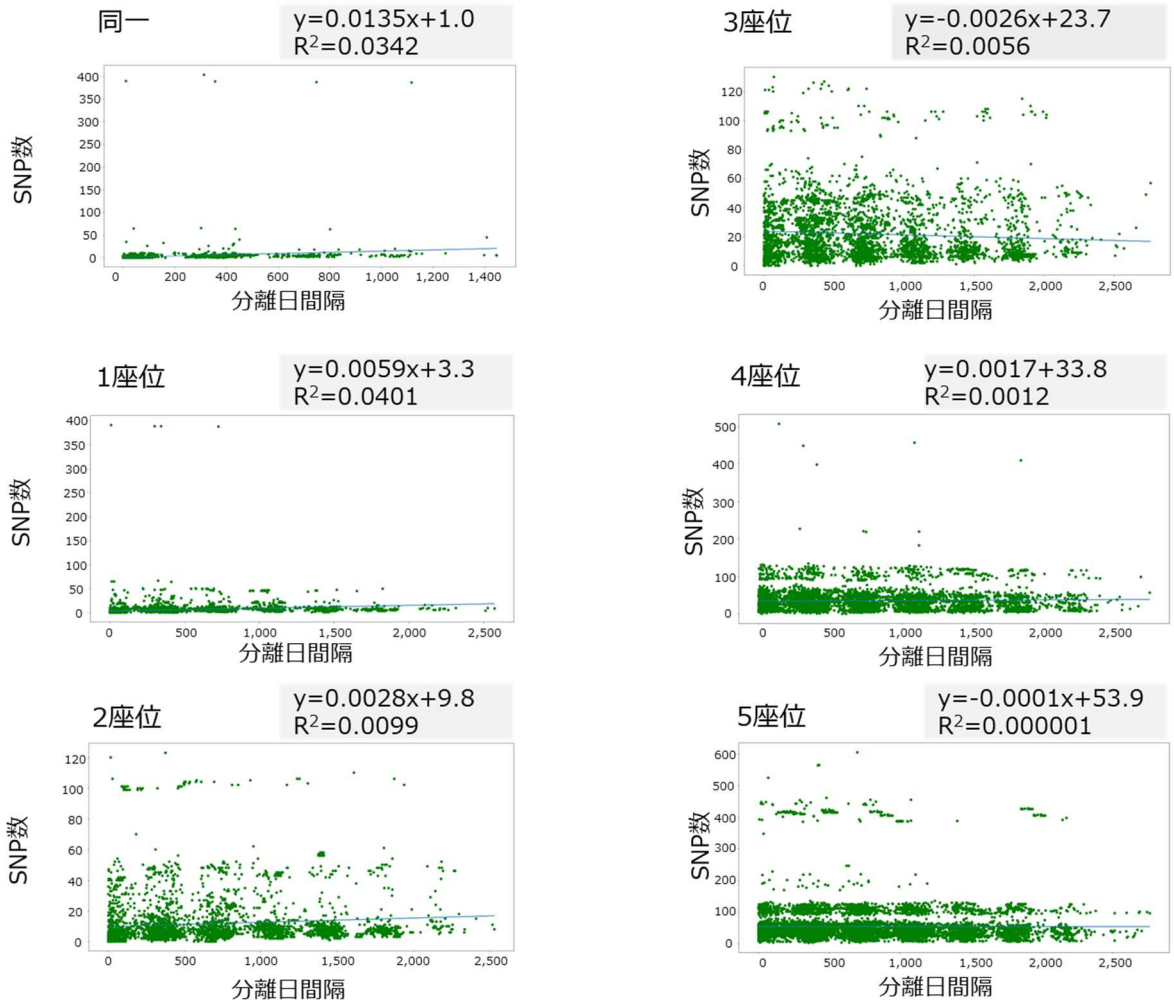


図 1. 解析菌株の clade 分布

Clade	株数
1	1
2	396
3	471
4/5	8
6	7
7	346
8	395
9	2
同定不可	10
計	1636

表 2. クラスター検出プログラムの出力例

株名	SNP5_cluster							SNP10_cluster						
	クラスター株	株数	重症化率(%)	無症状保菌者率(%)	距離中央値(km)	距離最小値(km)	距離最大値(km)	クラスター株	株数	重症化率(%)	無症状保菌者率(%)	距離中央値(km)	距離最小値(km)	距離最大値(km)
JNE130772	NA							NA						
JNE130856	JNE131493,JNE131896,JN	118	61.9	11.0	351.4	10.9	891.9	JNE131493,JNE131896,JN	229	61.1	11.8	304.9	10.9	891.9
JNE131070	NA							JNE160912,JNE171012,JN	6	50.0	33.3	318.5	37.4	872.1
JNE131158	NA							NA						
JNE131281	JNE131486,JNE131487,JN	25	52.0	32.0	53.0	14.3	913.7	JNE131486,JNE131487,JN	42	57.1	21.4	77.8	14.3	1032.7

表 3. 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 4. 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

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

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

腸管出血性大腸菌（EHEC）の高精度なサーベイランスを実現するためには、従来法である MLVA 型よりも詳細かつ安価で迅速な類別法が必要である。本研究では、EHEC のあるペアにおいて MLVA 型の差異から機械学習を用いて SNP 数を予測・類別指標とすることによりこれを実現し、高精度なサーベイランスに役立てることを目指した。機械学習モデルの構築には、東京工業大学が保有するスーパーコンピューター TSUBAME3.0 を用いた。

本研究では、まず日本の代表的な EHEC の血清型である O157 について機械学習モデルの構築を目指した。1 年目には 890 株、2 年目には 764 株のデータを追加し、最終的には合計 1636 株のデータを用いて機械学習モデルの作成を試みた。機械学習アルゴリズム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、SNP 数を連続値で予測する場合には 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。また株のペアの MLVA 型のデータを各 Clade ごとに分割し、各ペアの SNP 数を予測することも試みた。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。

3 年目においては、前年度までに得られていた O157 の機械学習モデル構築に用いた知見を利用し、O26、O111 の MLVA データを用いた学習・予測を試みた。O26 については 585 株、O111 については 285 株のデータを用いて、それぞれ約 17 万ペア、約 4 万ペアの MLVA 型のデータを用いた。前年度までの結果から、機械学習アルゴリズムとして勾配ブースティング法を使用した。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。

今後は本研究で作成した予測モデルを実際に新規株の近縁株判別に使用し、予測モデルの有用性について検討する。

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

1. O157 の機械学習モデル構築

2013年から2021年に分離されたEHEC O157の約1636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 130 万ペア）を研究代表者の李謙一氏から提供いただいた。

全データを用いた学習・予測においては、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。

Clade ごとの予測においては、任意の 2 株間の SNP 数のデータのうち、Clade 2、3、7、8 の各 Clade 内のペアのみを抽出した。各 Clade において、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。

予測結果として、各株ペア間の SNP 数を直接計算する連続値の予測と、各株ペアが 10 SNP または 20 SNP を閾値とした場合に近縁株であるか否かを予測するカテゴリの予測を行った。

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 TSUBAME 3.0 の環境を利用した。

連続値予測の最適化関数には平均二乗誤差（squared error）、カテゴリ予測の最適化関数には逸脱度（deviance）を用いた。

2. O26・O111 の機械学習モデル構築

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MLVA型データと任意の2株間のSNP数のデータ（それぞれ約17万ペア、約4万ペア）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。

予測結果として、各株ペア間の SNP 数を直接計算する連続値の予測と、各株ペアが 10 SNP または 20 SNP を閾値とした場合に近縁株であるか否かを予測するカ

テゴリの予測を行った。

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 TSUBAME 3.0 の環境を利用した。

連続値予測の最適化関数には平均二乗誤差（squared error）、カテゴリ予測の最適化関数には逸脱度（deviance）を用いた。

C. 研究結果

1. O157 株ペアの各 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. O157 株ペアの各 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。

3. O157 株ペアの SNP 数を連続値で予測する機械学習モデル

任意の株ペアにおいて、2 株の各 MLVA 座位データ（34 座位）、*stx1,2* 遺伝子の有無、分離地の緯度・経度情報を特徴量として用い、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。

この結果を図 5 に示す。連続値の予測においては、Clade 2 では二乗平均平方根誤差（RMSE）は 3.8 となり、これは直感的には Clade 2 内の各ペアの SNP 数の実測値に対し±4 ヶ所程度増減した予測が行われていることを表す。同様に Clade 3 では RMSE が 4.8、Clade 7 では RMSE が 35.6、Clade 8 では RMSE が 4.9 となった。

また、近縁株の基準を 10 SNP、20 SNP

とした場合の混同行列を図 6、7 に示す。

再現率（Recall）は、実測値から近縁株と判定される株ペアのうち、どの程度を予測から近縁株と判定できるかを表した数値であり、本研究で最も重要視している数値である。

Clade 2、3、8 でも、近縁株の基準を 10 SNP から 20 SNP に広げると再現率が上昇していた。一方、Clade 7 では他の Clade に比べて著しく再現率が低かった。

4. O157 株ペアを近縁株か否かのカテゴリで予測する機械学習モデル

任意の 2 株において、2 株の各 MLVA 座位データ（34 座位）、*stx1,2* 遺伝子の有無、分離地の緯度・経度情報を特徴量として用い、勾配ブースティング決定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。

近縁株の基準を 10 SNP、20 SNP とした場合の混同行列を図 8、9 に示す。

カテゴリの予測においてはどの Clade においても連続値の予測の場合よりも再現率が上昇しており、特に連続値の予測では難しかった Clade 7 における再現率が著しく上昇た。

5. O26 株ペアの SNP 数を連続値で予測する機械学習モデル

O26 の任意の株ペアにおいて、2 株の各 MLVA 座位データ（34 座位）、*stx1,2* 遺伝子の有無を、分離日間隔を特徴量として用い、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。この結果を図 10 に示す。連続値の予測においては、二乗平均平方根誤差（RMSE）は 44.8 となり、これは O26 血清型内の各ペアの SNP 数の実測値に対し平均的に±45 ヶ所程度増減した予測が行われていることを表す。

また、連続値予測において、近縁株の基準を 10 SNP 以内とした場合の混同行列を図 11 に示す。

再現率（Recall）（赤字）は、実測値から近縁株と判定される株ペアのうち、どの程度を予測から近縁株と判定できるかを表した数値であり、本研究で最も重要視している数値である。

O26 においての連続値予測においては再現率は 61.4%となり、O157 において Clade を分けて学習・予測した場合に比べて再現率が低かった。

6. O26 株ペアを近縁株か否かのカテゴリで予測する機械学習モデル

O26 血清型の任意の 2 株において、2 株の各 MLVA 座位データ（34 座位）、*stx1,2* 遺伝子の有無、分離日間隔を特徴量として用い、勾配ブースティング決定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。

近縁株の基準を 10 SNP、20 SNP とした場合の混同行列を図 11 に示す。

カテゴリの予測においてはどちらの近縁株基準においても連続値の予測の場合よりも再現率が上昇していた。

7. O111 株ペアの SNP 数を連続値で予測する機械学習モデル

O111 の任意の株ペアにおいて、2 株の各 MLVA 座位データ（34 座位）、*stx1,2* 遺伝子の有無を、分離日間隔を特徴量とし

て用い、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。この結果を図 12 に示す。連続値の予測においては、二乗平均平方根誤差（RMSE）は 38.7 となり、これは O111 血清型内の各ペアの SNP 数の実測値に対し平均的に±39 ヶ所程度増減した予測が行われていることを表す。

また、連続値予測において、近縁株の基準を 10 SNP 以内とした場合の混同行列を図 13 に示す。

O111 においての連続値予測において再現率は 27.7%となり、O157 において Clade を分けて学習・予測した場合や O26 において学習・予測した場合に比べて著しく再現率が低かった。

8. O111 株ペアを近縁株か否かのカテゴリで予測する機械学習モデル

O111 血清型の任意の 2 株において、2 株の各 MLVA 座位データ（34 座位）、*stx1,2* 遺伝子の有無、分離日間隔を特徴量として用い、勾配ブースティング決定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。

近縁株の基準を 10 SNP、20 SNP とした場合の混同行列を図 13 に示す。

カテゴリの予測においてはどちらの近縁株基準においても連続値の予測の場合よりも再現率が上昇していた。

D. 考察

1. O157 の機械学習モデルについて

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。

株全体からの学習・予測においては、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。

そこで、本研究では O157 の株ペアを各 Clade に分けて学習・予測を行った。株ペアの SNP 数を連続値で予測する機械学習モデルの場合、Clade 7 での予測精度が他の Clade に比べて悪かった。これは、Clade 7 のデータセットには他の Clade ではそれほど多くない 200 SNP 以上の株ペアデータが多かったことが原因であると考えられる。連続値の予測においては、株ペアデータ全体に対して SNP 数の予測が最適化

されるため、200 SNP 以上のペアの学習・予測にあった最適化がなされることになる。この結果、RMSE が 36 程度と大きくなり、近縁株の閾値を大きく超えたため、近縁株の予測精度が悪かったと考えられる。

一方、カテゴリの予測では Clade 7 においても 60%以上の再現率が見られた。こちらの予測では、近縁株か否かの○×問題を解く学習・予測のため、データセットの中で 1%以下の近縁株についても、今回用いた特徴量から学習・予測が可能であったと考えられる。

2. O26・O111 の機械学習モデルについて

O26・O111 においては株ペアの SNP 数を連続値で予測する機械学習モデルの場合、O157 で Clade を分けて 予測した場合よりも予測精度が悪かった。これは、O157 においては Clade を分けることで、Clade 間ペアの SNP 数が大きいと思われるデータを学習・予測データから除くことができたことに起因している。現状では O26・O111 血清型においてはこのような Clade の別はなく、図 10、12 からもわかるように 2 株間で SNP 数が 1000 前後の遠縁株間データが含まれている状態である。また O26・O111 血清型においては、O157 今後に比べて既存のデータが少ないことも予測精度低下の原因と考えられる。今後、O26・O111 血清型のデータの追加により部分的に機械学習・予測の精度が上がることが期待される。

一方、カテゴリの予測においては、O26・O111 においても 75%以上の再現率が見られた。こちらの予測では、近縁株か否かの学習・予測のため、データセットの少なさ、また遠縁株が含まれたデータセットであっても、今回用いた特徴量から学習・予測が十分に可能であったと考えられる。

E. 結論

1. O157 の機械学習モデルについて

本研究では、2013 年から 2021 年に分離された国内 EHEC O157、1636 株についての MLVA 型データと任意の株ペアの SNP 数のデータから、MLVA 座位、*stx1,2* 遺伝子の有無、分離地の緯度・経度情報を特徴量として株ペアの 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。

また Clade ごとにデータを分けた場合の連続値の学習・予測においては、特に Clade 7 において、SNP 数の大きい株ペアのデータに学習・予測全体が影響を受け、近縁株の予測がうまくいかない部分が見られた。

一方で、カテゴリでの学習・予測においては、Clade 7 においても精度良く近縁株を予測することができた。そのため、今後はカテゴリでの予測モデルを用いて、本機械学習モデルの実応用性を検討したい。

2. O26・O111 の機械学習モデルについて

本研究では、2013 年から 2021 年に分離された国内 EHEC O26・O111 血清型株についての MLVA 型データと任意の株ペア

のSNP数のデータから、MLVA座位、*stx1,2*遺伝子の有無、分離日間隔を特徴量として株ペアのSNP数及び近縁株か否かのカテゴリを予測する機械学習モデルの作成を行った。

連続値の学習・予測においては、近縁株の予測がうまくいかない部分が見られたが、カテゴリでの学習・予測においては、どちらの血清型においても精度良く近縁株を予測することができた。そのため、今後はカテゴリでの予測モデルを用いて、本機械学習モデルの実応用性を検討したい。

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

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。

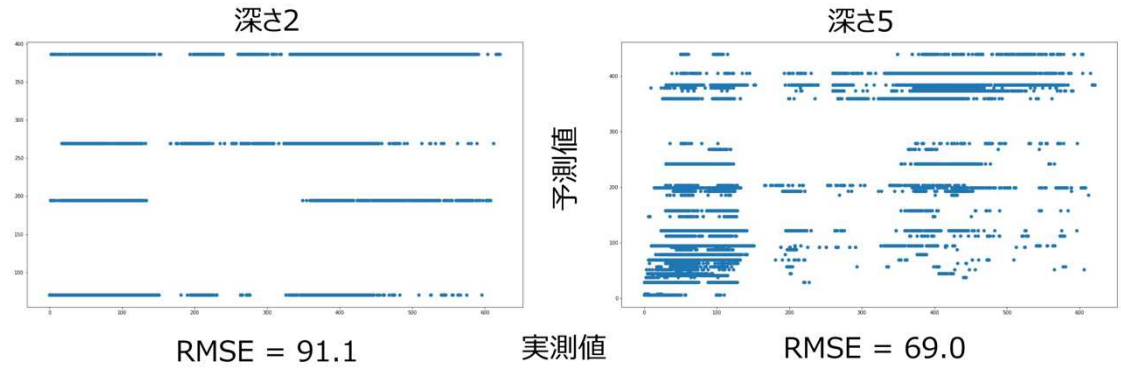


図 2. 各 MLVA 座位の差異の有無を特徴量として勾配ブースティング回帰木を利用したモデル

横軸は SNP 数の実測値、縦軸は SNP 数の予測値を示す。

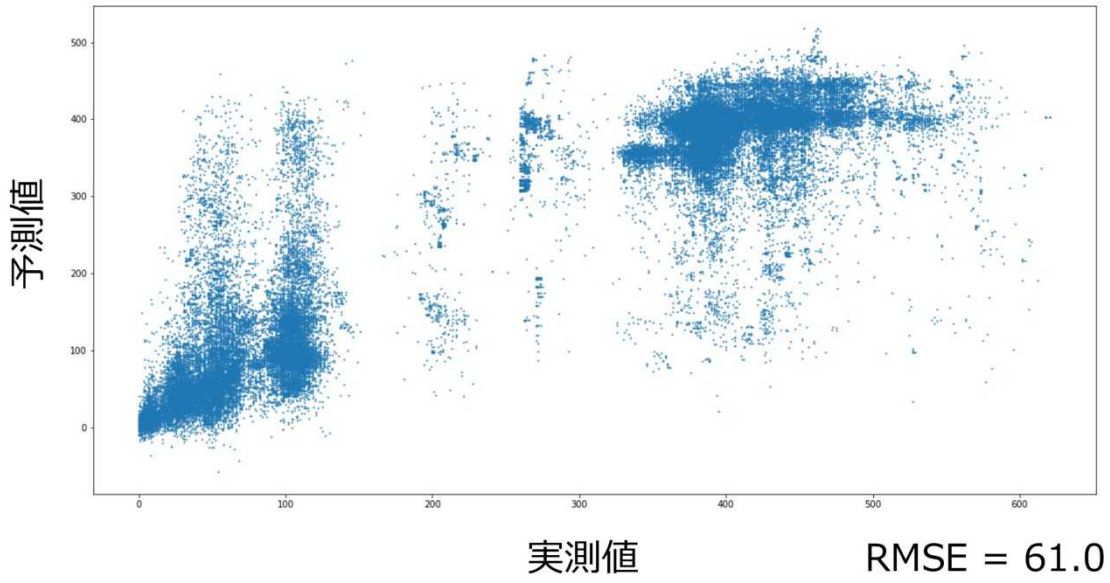


図 3. 各 MLVA 座位を特徴量として線形回帰を利用したモデル

横軸は SNP 数の実測値、縦軸は SNP 数の予測値を示す。

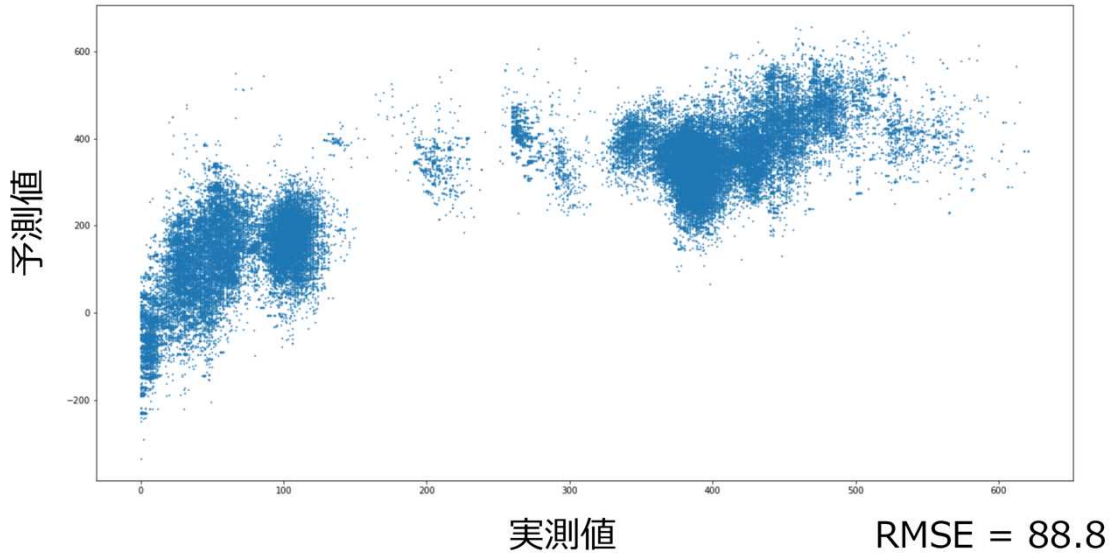


図 4. 各 MLVA 座位を特徴量として勾配ブースティング回帰木を利用したモデル

横軸は SNP 数の実測値、縦軸は SNP 数の予測値を示す。

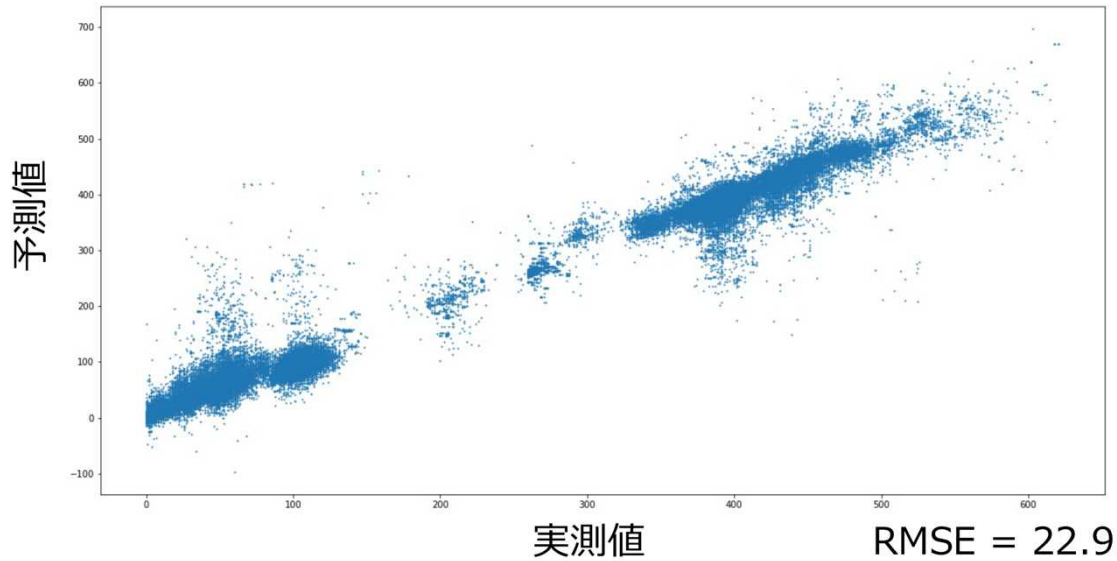


図 5. 連続値予測の機械学習モデルの予測結果

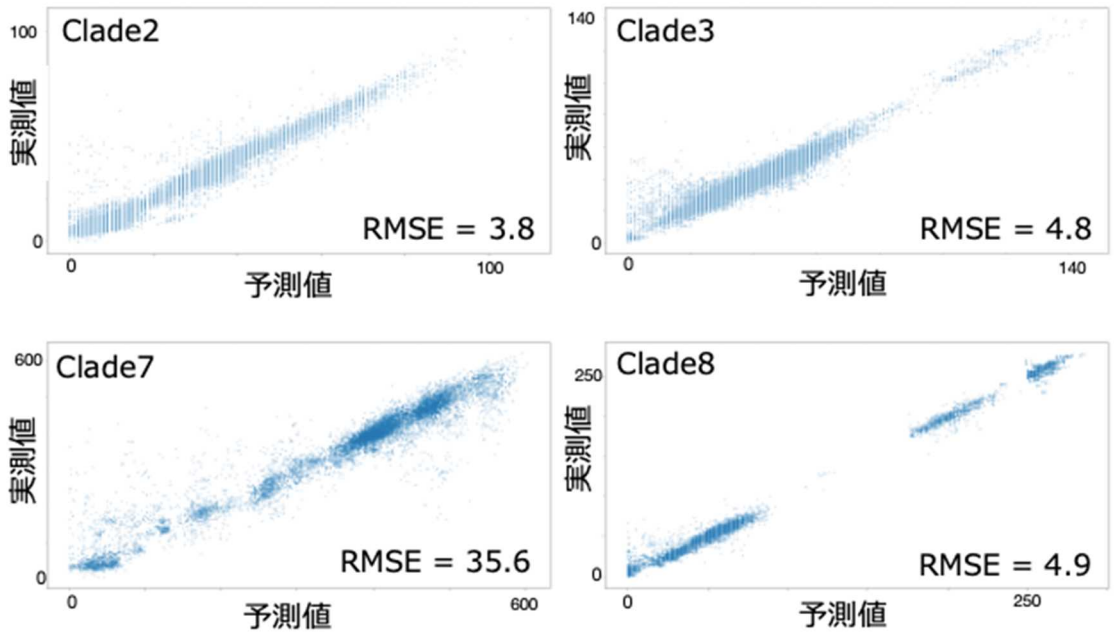


図 6. 連続値予測の機械学習モデルの予測結果（閾値 10 SNP の混同行列）

Clade2

実測値＼予測値	≦10	>10	
≦10	4,489	965	82.3%
>10	554	13,545	
	89.0%		

Clade3

実測値＼予測値	≦10	>10	
≦10	657	407	61.6%
>10	34	26,574	
	95.1%		

Clade7

実測値＼予測値	≦10	>10	
≦10	0	82	0%
>10	3	14,837	
	0%		

Clade8

実測値＼予測値	≦10	>10	
≦10	1,207	342	77.9%
>10	1	17,904	
	99.9%		

赤:Recall (再現率) , 青:Precision (適合率)

図 7. 連続値予測の機械学習モデルの予測結果（閾値 20 SNP の混同行列）

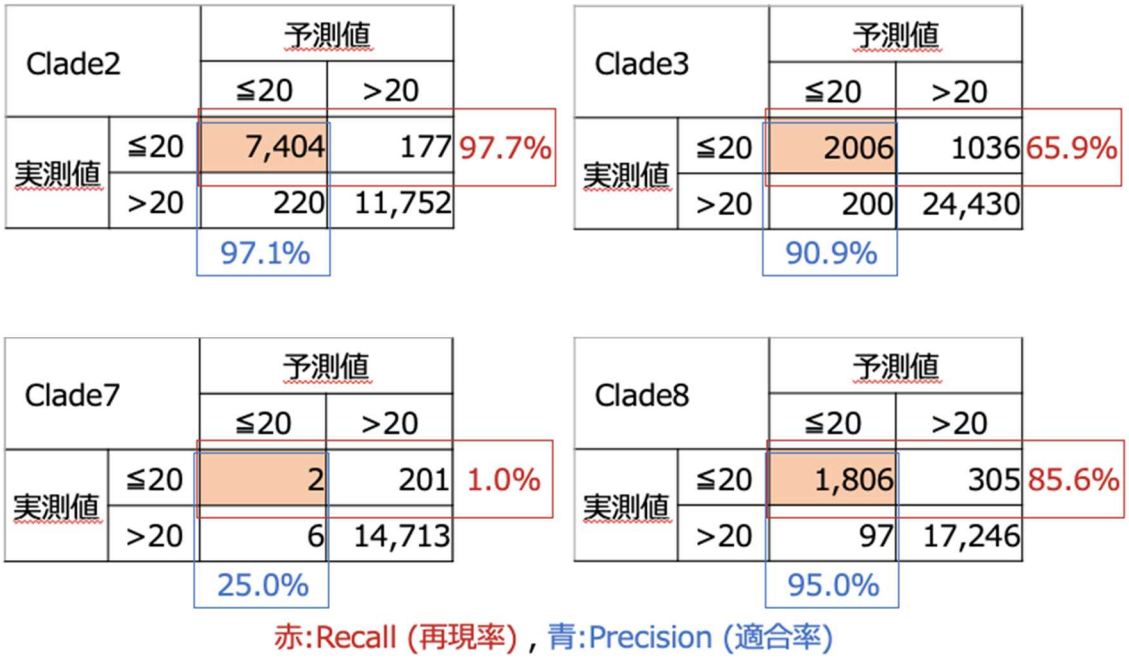


図 8. カテゴリ予測の機械学習モデルの予測結果（閾値 10 SNP の混同行列）

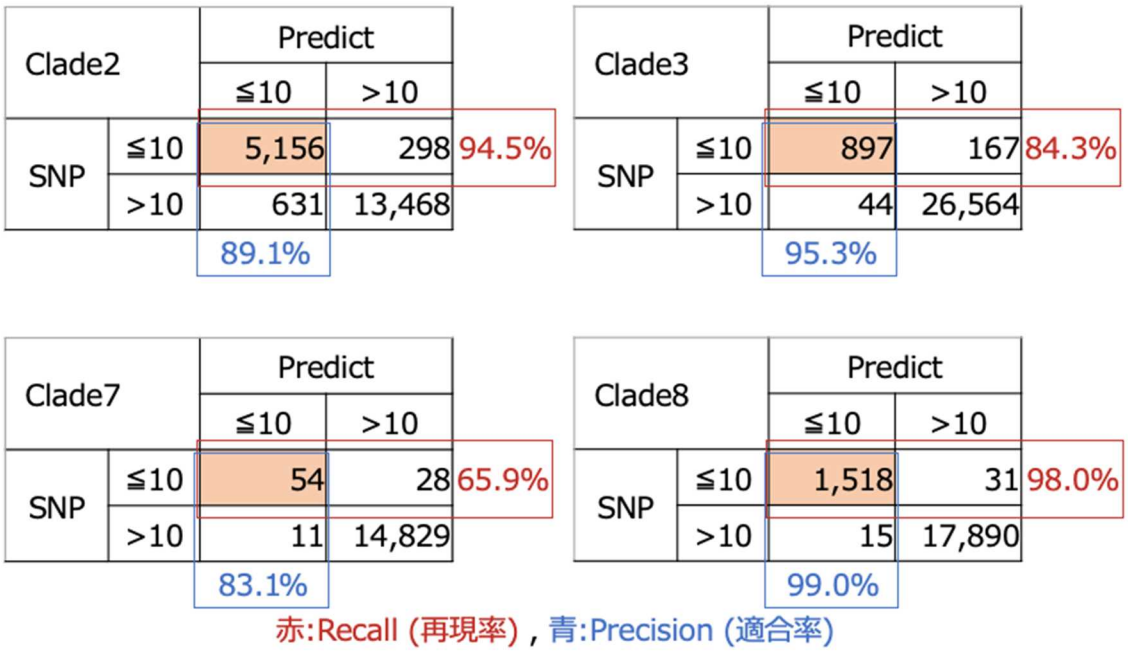


図 9. カテゴリ予測の機械学習モデルの予測結果（閾値 20 SNP の混同行列）

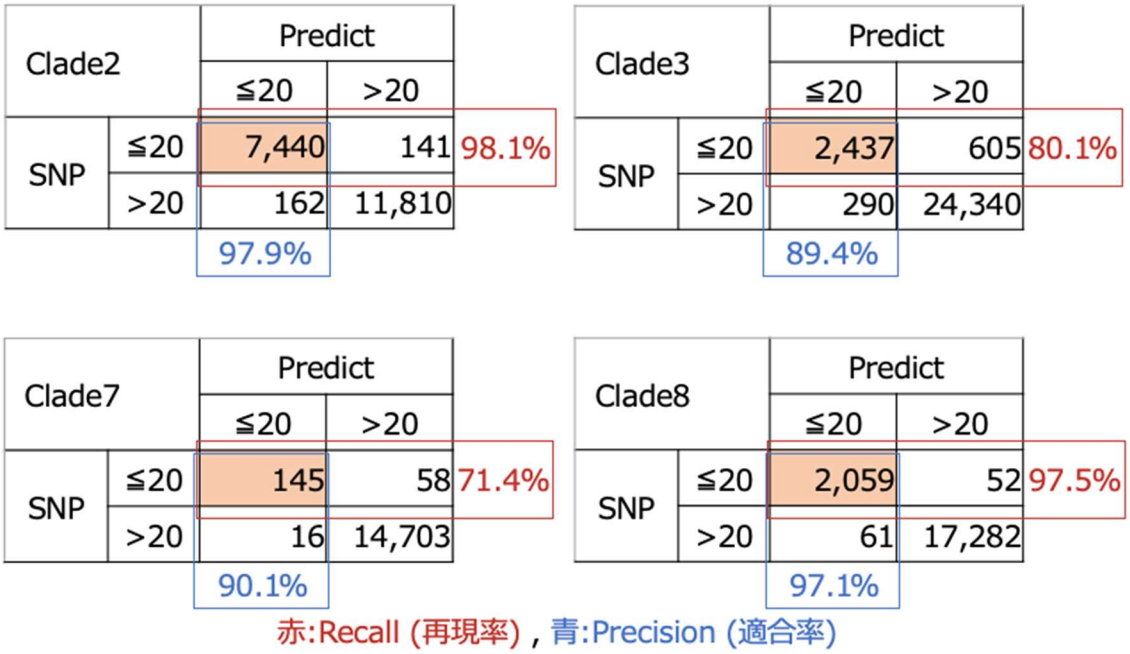
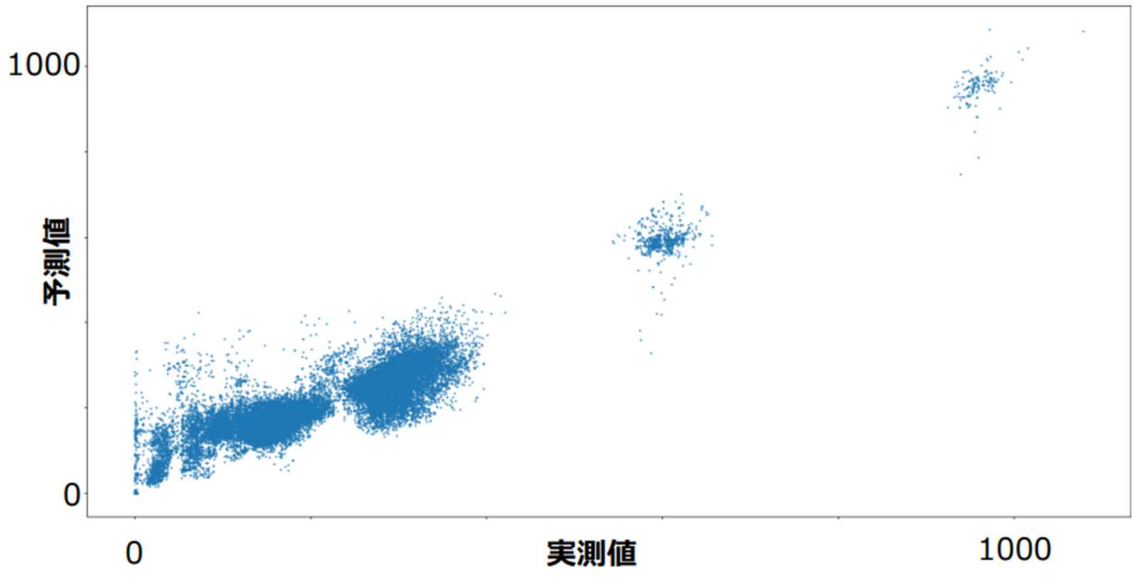


図 10. O26 についての連続値予測の機械学習モデルの予測結果



RMSE=44.8

図 11. O26 についての連続値･カテゴリ予測の機械学習モデルの予測結果（混同行列）

赤: 再現率（Recall）、青:適合率（Precision）

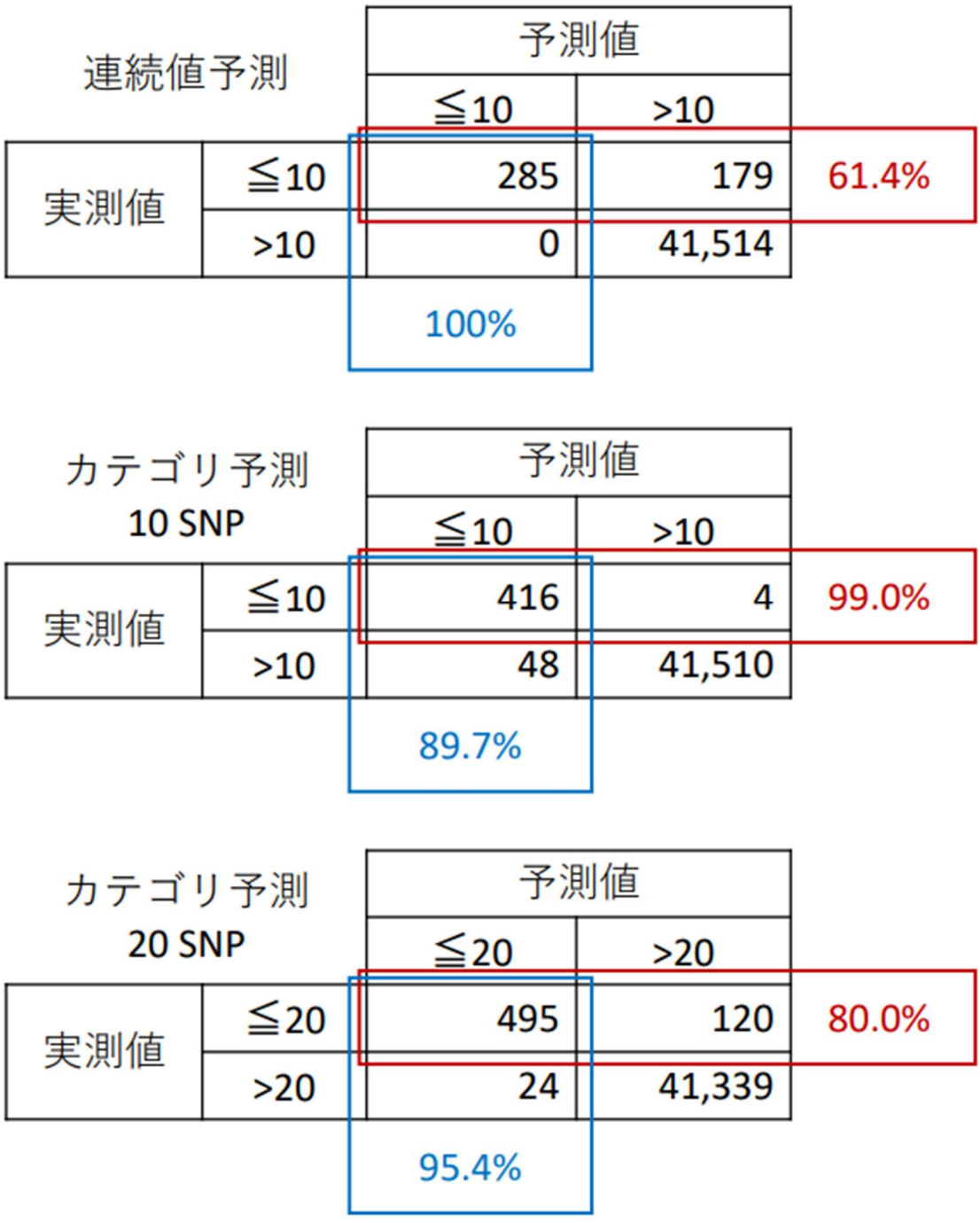
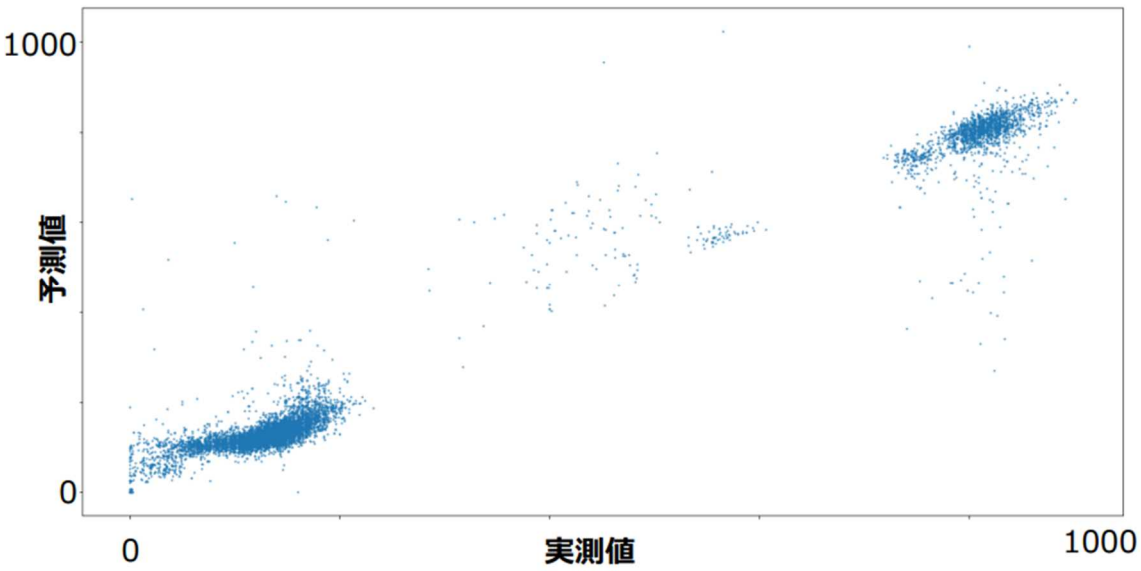


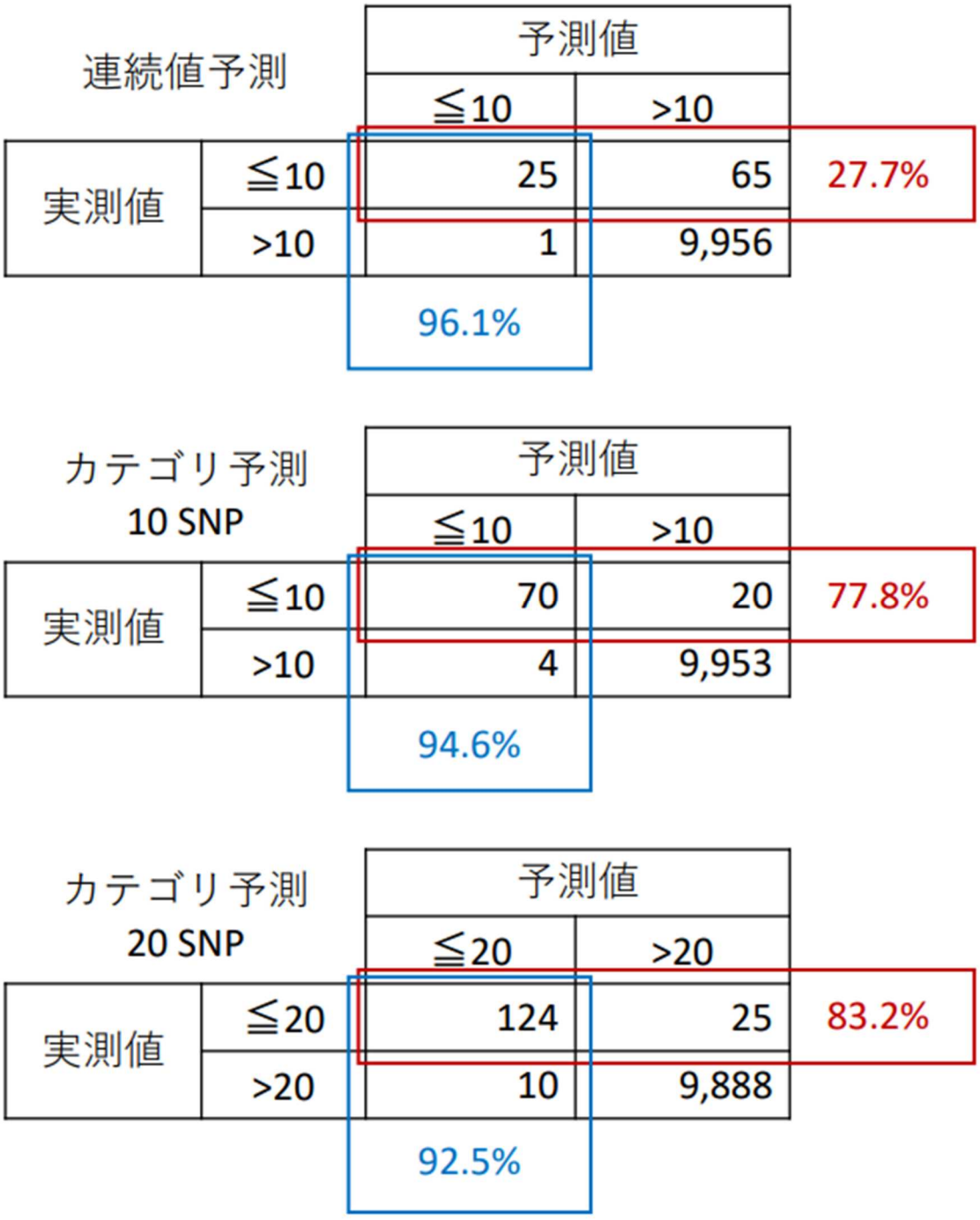
図 12. O111 についての連続値の機械学習モデルの予測結果



RMSE=38.7

図 13. O111 についての連続値･カテゴリ予測の機械学習モデルの予測結果（混同行列）

赤: 再現率（Recall）、青:適合率（Precision）



別紙４

研究成果の刊行に関する一覧表レイアウト

書籍

著者氏名	論文タイトル名	書籍全体の編集者名	書　籍　名	出版社名	出版地	出版年	ページ
なし							

雑誌

発表者氏名	論文タイトル名	発表誌名	巻号	ページ	出版年
なし					