

令和 6 年 6 月 25 日現在

機関番号：82111

研究種目：基盤研究(C) (一般)

研究期間：2021～2023

課題番号：21K05608

研究課題名(和文) 宿主ゲノムとウイルスゲノムを活用した病徴予測モデルの開発

研究課題名(英文) Predicting symptom severity in viroid-infected plants using the viroid genome sequence

研究代表者

松下 陽介 (Matsushita, Yosuke)

国立研究開発法人農業・食品産業技術総合研究機構・植物防疫研究部門・チーム長

研究者番号：00414665

交付決定額(研究期間全体)：(直接経費) 3,200,000円

研究成果の概要(和文)：ウイルスは分離株と宿主植物種の組み合わせにより、無病徴から強毒まで様々な症状を示す。我々は教師なし機械学習を用いて、トマトにおけるジャガイモやせいもウイルスによる病徴の程度を予測するアルゴリズムを開発した。このアルゴリズムはサイレンシング機構を模倣し、ウイルスと宿主植物のゲノム配列を用い、ウイルスの短い断片の宿主植物ゲノムへのアラインメント、アラインメントカバレッジの計算、UMAPとDBSCANを用いたカバレッジに基づくウイルスのクラスタリングを実行した。接種実験により、このアルゴリズムの有効性が確認できた。このアルゴリズムは様々なウイルスと植物の組み合わせに適用可能である。

研究成果の学術的意義や社会的意義

本研究で開発したアルゴリズムは、データベースで公開されているウイルスのゲノム情報と宿主植物であるトマトのゲノム情報を利用し、ウイルスが感染した時に生じる短い塩基をコンピュータ計算で予測し、学習データを利用してウイルスの病原性を予測できる。将来的には、本アルゴリズムを利用して、トマト以外の様々な重要作物とウイルスの組み合わせに対し、ウイルスが感染した時の病徴レベルを予測することが可能となることが期待される。

研究成果の概要(英文)：Viroids induce symptoms of varying severity, ranging from latent to severe, based on the combination of viroid isolates and host plant species. Here, we developed an algorithm using unsupervised machine learning to predict the severity of disease symptoms caused by potato spindle tuber viroid (PSTVd) in tomato. This algorithm, mimicking the RNA-silencing mechanism thought to be linked to viroid pathogenicity, requires only the genome sequences of the viroids and host plants. It involves three steps: alignment of synthetic short sequences of the viroids to the host plant genome, calculation of the alignment coverage, and clustering of the viroids based on coverage using UMAP and DBSCAN. Validation through inoculation experiments confirmed the effectiveness of the algorithm in predicting the severity of disease symptoms induced by viroids. As the algorithm only requires the genome sequence data, it may be applied to any viroid and plant combination.

研究分野：植物病理学

キーワード：ウイルス 病徴予測 ゲノム ジャガイモやせいもウイルス トマト RNAサイレンシング アルゴリズム

科研費による研究は、研究者の自覚と責任において実施するものです。そのため、研究の実施や研究成果の公表等については、国の要請等に基づくものではなく、その研究成果に関する見解や責任は、研究者個人に帰属します。

### 1. 研究開始当初の背景

ウイルスは 250-400 塩基の一本鎖環状 RNA のみからなる植物病原体である。ウイルスはタンパク質をコードせず、宿主植物の酵素を利用して自身のゲノムを複製し、宿主植物の維管束を通じて全身に感染する。ウイルスに感染した植物は、矮化や着果不良などの症状を示し、深刻な農作物被害をもたらす。

代表的なウイルスであるジャガイモやせいもウイルス (potato spindle tuber viroid; PSTVd) は、宿主範囲が広く、農作物に多いナス科やキク科をはじめとして、ヒユ科、キキョウ科、ナデシコ科などにも感染する。PSTVd は、異なる宿主植物に感染することで容易に変異し、これまでに 160 以上の変異体が見つかった (Matsushita et al. 2015)。これらの PSTVd はトマトにおける発病レベルに応じて、弱毒や強毒のように病徴の程度で分類される (図 1)。これまでにウイルスのゲノム構成とその発病レベルの関係性に着目した研究が行われているが、両者の関係性を説明できるものはなかった。さらに、異なる宿主植物種に同じウイルスを感染させたときに、病徴が宿主植物種ごとに異なることが報告されている。これらのことから、ウイルスの病原性は、それ自身のゲノム構成だけでなく、宿主植物のゲノムにも依存していることが示唆される。

ウイルスが感染した宿主植物細胞の中にはウイルス由来の 21-24 塩基の small RNA (viroid-sRNA) が多量に蓄積されていることが明らかとなっている。これらの viroid-sRNA が宿主植物の一部の遺伝子の mRNA に相補的に結合し、その mRNA を切断することでサイレンシングを誘発することが指摘されている (Adkar-Purushothama et al. 2019)。

そこで、我々はウイルスが植物に感染するときに起きる RNA サイレncing機構を鑑みて、PSTVd の病徴レベルの評価を試みた。具体的に、329 種の PSTVd の変異体のゲノムとトマトゲノムの相同配列を網羅的に調べた。次に、トマトゲノム上にあるそれらの相同配列の位置に基づいて PSTVd を分類した。続けて、我々は先行研究で調べられている弱毒・強毒の情報を、クラスタリング結果に書き加えると、ウイルスが病原性の強弱に応じて分類されていることがわかった。このことは、PSTVd のゲノム配列と宿主植物のゲノム配列がわかれば、その宿主植物に対するウイルスの病原性を予測できることを示唆している。これを受け、本研究ではこの分類アルゴリズムの最適化を行い、さらに接種試験などによる検証を重ね、ウイルスの病徴レベルを予測するアルゴリズムの開発に取り組んだ。



強毒 弱毒 非感染  
図1 ウィルスに感染したトマトの症状

### 2. 研究の目的

本研究ではトマトに対する PSTVd の発病の程度 (弱毒・強毒) を予測するアルゴリズムを開発することを目的としている。

予備試験では、329 変異体の PSTVd を 13 クラスタに分類できた。各クラスタに含まれている PSTVd を先行研究で報告された病徴レベルに照合すると、各クラスタが PSTVd の病原性の強弱に応じて分類された。しかし、先行研究の結果は、異なる基準で評価されているため、不正確である可能性があった。そこで、各クラスタから 2-3 の PSTVd 変異体を選び、基準を統一させて接種試験を行い、各クラスタを再評価することとした。

ここでは予備実験で構築したアルゴリズムの最適化を行う。本アルゴリズムは、PSTVd の RNA を断片化、断片をトマトゲノムにマッピング、マッピング位置をクラスタリングするという 3 ステップからなる。この中で、断片化のときの長さを 18-22 塩基に変更させたり、マッピングアルゴリズムを複数変更させたり、さらにクラスタリングアルゴリズムを k-means、DBSCAN、ワード法などに変更させて、分類結果がどのように変化するのかを評価する。最後に接種試験の結果を反映させ、新しい PSTVd のゲノムを代入すると、その病徴レベルを予測できる枠組みを構築する。

### 3. 研究の方法

#### (1) トマトにおける PSTVd の接種試験

接種実験では、307 株の PSTVd 分離株のうち 33 株を無作為に選択した。トマト (*Solanum*

*lycopicum* 'Rutgers') における症状の強弱を評価した。接種 2 か月後に発病の程度を調査した。発病の程度「弱毒、中間、強毒」は PSTVd 分離株 DDBJ accession No.KR611355 および No.AY518939 の病徴を基準とした。草丈が対照個体とほぼ同じ高さか、対照個体よりわずかに低い場合、症状は弱毒 (mild) とみなした。対照個体に比べて著しく発育不良であった場合、典型的なウイルス誘発症状 (発育不良、葉の巻込み、葉の曲がりなど) を示した場合、あるいは途中で枯死した場合は、症状を強毒 (severe) とした。軽症と重症の中間の症状は「中間 (moderate)」とした。

#### (2) PSTVd およびトマトのゲノム配列

合計 307 の PSTVd 分離株を用い、ゲノム配列は GenBank から入手した。植物はトマト (*S. lycopersicum* 'Rutgers') を用い、ゲノム配列は対応するアノテーション ITAG4.0 とともに、Sol Genomics Network (Hosmani et al. 2019) からダウンロードして利用した。

#### (3) PSTVd 感染トマトの small RNA-seq の解析

接種実験の結果に基づき、無作為に選択した 3 株 (No.MG450357、LC388854、LC388852) および弱毒および強毒の症状を示した 4 株 (No.FR851463、DD220185、JX280944、X58388) の PSTVd 分離株をそれぞれ用いて、vd-sRNA の発現プロファイルを調べるために small RNA-seq 解析を行った。各サンプルの全 RNA を接種後 2 ヶ月以上経過した最上部の葉から抽出した。21~24nt の長さのリードのみを、CircSeqAlignTk (Sun et al. 2022) を用いて PSTVd 分離株のゲノム RNA にマッピングし、vd-sRNA の発現プロファイルをまとめた。

#### (4) ウィロイドによる発病程度の予測アルゴリズム

宿主植物におけるウィロイドの病徴の強毒程度を予測するためアルゴリズムは、次の 3 つのステップからなる。宿主植物におけるウィロイド誘発症状の予測アルゴリズムは、アライメント、アライメント要約、クラスタリングの 3 つのステップからなる (図 2)。このアルゴリズムは、複数のウィロイド分離株のゲノム RNA 配列と宿主植物のゲノム配列を入力として必要とし、各入力 PSTVd 分離株のクラスタ識別を出力として生成する。

アライメントを作成するために、まず各 PSTVd について、PSTVd 分離株の円形ゲノム RNA 配列上の開始位置をランダムに選択し、そこから 21、22、23、24 nt を 10,000 回切り出した。次に、これらの短い配列をトマトの参照配列にアライメントした。

各アライメント領域におけるアライメントカバレッジ (すなわちリード数) については、各行がアライメント領域、各列が PSTVd 分離株、各セルが対応する分離株の対応するアライメント領域のアライメントカバレッジを表す行列にまとめた。アライメントの要約により、r 行 (すなわち領域) と 307 列 (すなわち PSTVd 分離株) を持つ行列が得られた。行列の各列は、対応する PSTVd 分離株の特徴ベクトル (r 次元) として認識された。

特徴ベクトルの次元は、n 個の近傍をパラメータとする UMAP を用いて 2 次元に縮小された。次に、2 次元座標系において、パラメータ eps (1 つのクラスターとみなすための 2 つのサンプル間の最大距離) および minPts (ある点をコア点とみなすためのクラスター内の最小サンプル数) を有する DBSCAN を用いて、307 の PSTVd 分離株をクラスタリングした。

最適なクラスタリング・パラメーター (例えば、d、l、v1、v2、n、eps、minPts) を同定するため、接種実験に用いた 37 の PSTVd 分離株のうち 20 をランダムに選択し、接種実験の結果を教師付き標識として用いた。中等度の症状の重篤度は、軽症や重症の症状の重篤度と比較して曖昧な場合があるため、弱毒および強毒の症状を示すとタグ付けされた PSTVd 分離株のみを教師付きラベルとして用いて最適なパラメータを求めた。最適なパラメータを決定した後、残りの 17 の PSTVd 分離株を用いてクラスタリング結果を評価した。さらに、学習用分離株と検証用分離株のランダムサンプリングがクラスタリング性能に与える影響を打ち消すため、学習用分離株と検証用分離株の組み合わせを変えながら、学習と検証のプロセスを 100 回繰り返した。

#### 5) 検証指標の算出 (F1-score)

F1 スコアは、精度と想起を組み合わせた指標である。この 2 つのメトリクスは調和平均として計算される:  $F1\text{-score} = 2 / [(1/\text{precision}) + (1/\text{recall})]$ 。精度は  $TP / (TP + FP)$  として計算され、一方想起は  $TP / (TP + FN)$  として計算され、ここで TP, FP, FN はそれぞれ真陽性 (重度に正しく予測されたもの)、偽陽性 (軽度で誤って重度と予測されたもの)、偽陰性 (重度に誤って軽度と予測されたもの) を表す。F1 スコアは、精度と想起の両方を同時に最大化するように努め、モデルの性能の包括的な尺度を提供する。

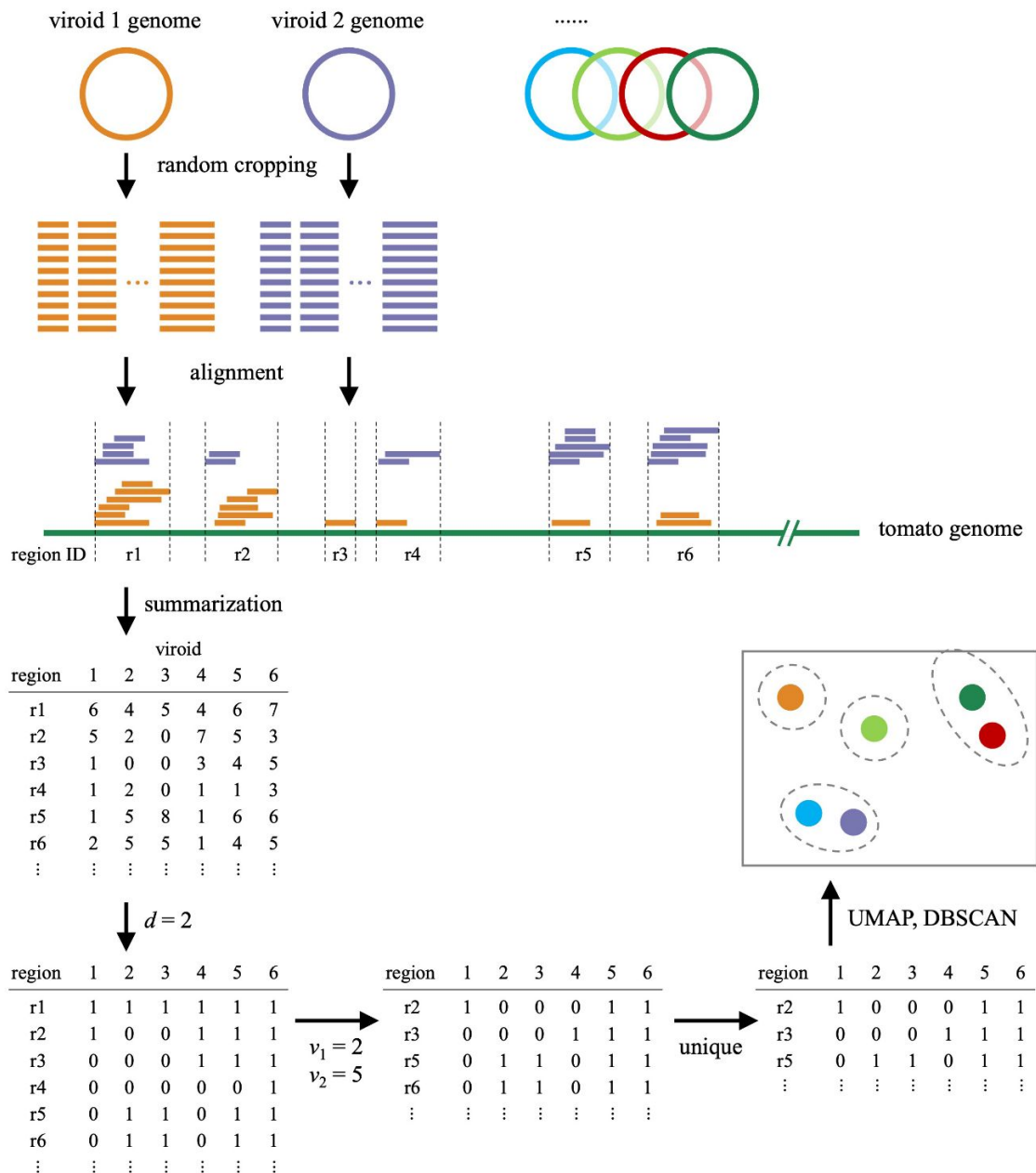


図2 ウイロイドによる病徴の程度を予測するアルゴリズムの概要図

#### 4. 研究成果

##### (1) PSTdV を接種したトマトの病徴の多様性

無作為に選んだ 33 株の PSTdV をトマトに接種した結果、15 株および 11 株がそれぞれ弱毒および強毒の症状を引き起こすと判定された。また、弱毒と強毒の中間の症状を引き起こした 7 株は中間と判定された。

##### (2) PSTdV 感染トマト植物における vd-sRNA の発現プロファイル

vd-sRNA の発現プロファイルを探索するため、軽症および重症の症状を誘発する PSTdV 分離株から無作為に選んだ 3 株 (MG450357, LC388854, LC388852) および 4 株 (FR851463, DD220185, JX280944, X58388) について、それぞれ small RNA-seq 解析を行った。これらのプロファイルに共通する特徴として、21 nt と 22 nt のアラインメントリードが優勢で、23 nt と 24 nt のアラインメントリードはわずかであった。さらに、+鎖と-鎖で対称性のない複数のホットスポットが観察された。

弱毒症状を引き起こした 3 つの分離株 (MG450357, LC388854, LC388852) は、同様の発現プロファイルを示した：フォワードストランドには 5 つのホットスポット (P1, P3, P4, P5, P8) があり、P5 が最も高いカバーレッジを示した。さらに、-鎖には 3 つのホットスポット (M3, M4, M6) があり、中でも P4 のカバー率が最も高かった。

強毒症状を引き起こした 4 株のうち、DD220185, JX280944 および X58388 は、弱毒の症状を引き起こした 3 株と比較的類似した発現プロファイルを示した。しかし、弱毒症状を引き起こす分離株では、+鎖上に P5 という単一のハイカバーレッジホットスポットが見られたのに対し、強毒症状を引き起こす分離株では、P5 を含む複数のハイカバーレッジホットスポットが見られた。

### 3) 接種実験による予測アルゴリズムの検証

ウイロイド感染植物における病徴の程度を予測するために、教師なし機械学習アプローチに基づく予測アルゴリズムを開発した。このアルゴリズムは、306 の PSTVd 分離株のゲノム RNA 配列とトマトのゲノム配列を入力として使用し、各入力 PSTVd 分離株のクラスタ識別を出力として割り当てる。このアルゴリズムは教師なし機械学習アプローチに基づいているが、パラメータ (d、l、v1、v2、n、r、eps、minPts) を最適化し、最終的なクラスタを注釈するために、教師ありデータを必要とする。

この目的のため、症状の重篤度がラベル化された 37 株の PSTVd 分離株のうち 20 株を無作為に選択し、提案アルゴリズムの最適パラメータを決定した。その後、残りの 17 の PSTVd 分離株の接種結果を用いてクラスタリング結果を検証した。ランダム選択がモデルの性能に影響を与える可能性を考慮し、学習と検証の単離株の組み合わせを変えながら、学習と検証のプロセスを 100 回繰り返した。検証の平均 F1 スコアは 0.85 であった。その結果、11 ケースは完全に予測可能 (F1 スコア = 1.0) であったが、残りの 89 のシミュレーションケースには 1 つか 2 つの誤分類があった。最も誤判定しやすい分離株は FR851463 で、トマト株に軽い症状を引き起こすと予測されたが、接種実験結果は強毒症状を示した。次に誤分類が多かったのは U23060 で、トマトで強毒な症状を引き起こしたにもかかわらず、「弱毒」と予測された。

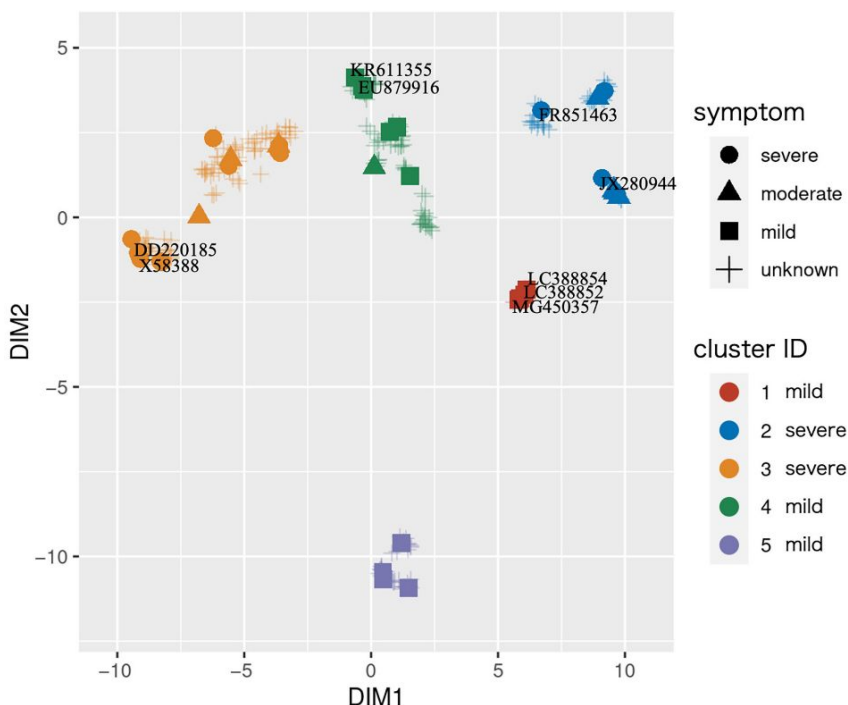
次に、F1 スコアが 1.0 の 11 のケースを検討したところ、様々な訓練と検証のサブセットの組み合わせで一貫した結果を発見した。それぞれのケースで、F1 スコア 1.0 を出すパラメータの組み合わせが 6 つあった。例えば、そのような組み合わせの 1 つは、PSTVd 分離株の特徴ベクトルの次元 (すなわち、r) を 273 として指定し、5 つのクラスタをもたらし (図 3)、クラスタ 1、4、5 の PSTVd はトマトで弱毒の症状を引き起こすと予測され、クラスタ 2 と 3 の PSTVd は強毒の症状を引き起こすと予測された。中間症状を引き起こす PSTVd は、弱毒と強毒の両方のクラスタに分布していたが、これはこれらの症状が曖昧であり、どちらの症状にも近い可能性があるためである。UMAP から得られた座標は非線形変換に基づいていることに注意することが重要である。そのため、点間の距離はデータ点間の類似性を反映しているが、座標そのものが固有の意味を持つわけではない。さらに、minPts の値 (= 2、4、6、8、10) に関係なく一貫した結果が得られたことから、minPts はクラスタリングの結果に大きな影響を与えないことがわかった。

さらに、vd-sRNA の発現プロファイルとクラスタリング結果を比較すると、vd-sRNA の発現プロファイルが類似している場合、2 つの分離株はクラスタリング空間内でより近くにグループ化することがわかった。

ウイロイドとその宿主植物のゲノム配列のみを用いて、ウイロイドが宿主植物に引き起こす病徴レベルを予測するアルゴリズムを発表した。このアルゴリズムの検証は有望な結果を示した: PSTVd 分離株によって誘発される病徴を 11 のケースで正しく予測し、100 のシミュレーションにおいて誤分類は 1 または 2 件、示しただけであった。このアルゴリズムは、トマトにおいて PSTVd 分離株によって生じる病徴の強弱の程度を予測する上で高い有効性を示した。ウイロイドが alternative splicing やメチル化修飾に影響を与えるメカニズムの理解が進むにつれて、これらの複雑な相互作用を組み込んだアルゴリズムの性能のさらなる向上が期待される。

**図 3 PSTVd 分離株のクラスタリング結果**

DIM1 と DIM2 は、一様多様体近似と射影 (UMAP) と投影 (UMAP) から計算された 2 次元を表し、色はノイズを含むアプリケーションの密度ベース空間クラスタリング (DBSCAN) で決定されたクラスタ識別を表す。



5. 主な発表論文等

〔雑誌論文〕 計1件（うち査読付論文 1件/うち国際共著 0件/うちオープンアクセス 1件）

1. 著者名 Sun Jianqiang, Yosuke Matsushita	4. 巻 -
2. 論文標題 Predicting Symptom Severity in PSTVd-Infected Tomato Plants Using the PSTVd Genome Sequence	5. 発行年 2024年
3. 雑誌名 Molecular Plant Pathology	6. 最初と最後の頁 -
掲載論文のDOI（デジタルオブジェクト識別子） 10.1111/mpp.13469	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 -

〔学会発表〕 計0件

〔図書〕 計0件

〔出願〕 計1件

産業財産権の名称 情報処理装置、情報処理方法、及びプログラム	発明者 孫 建強・松下 陽介	権利者 同左
産業財産権の種類、番号 特許、特開2022-139956	出願年 2022年	国内・外国の別 国内

〔取得〕 計0件

〔その他〕

-

6. 研究組織

	氏名 (ローマ字氏名) (研究者番号)	所属研究機関・部局・職 (機関番号)	備考
研究分担者	孫 建強 (Sun Jianqiang) (90838624)	国立研究開発法人農業・食品産業技術総合研究機構・農業情報研究センター・主任研究員  (82111)	

7. 科研費を使用して開催した国際研究集会

〔国際研究集会〕 計0件

8. 本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関
---------	---------