様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

# 科学研究費助成事業　　研究成果報告書

令和　５ 年　 ６ 月　 ３ 日現在

機関番号：３４４１７
研究種目：若手研究
研究期間：2020～2022
課題番号：２０Ｋ１８９８１

研究課題名（和文）DNA鑑定実務に資する人工知能によるアーチファクト自動判定ツールの開発

研究課題名（英文）Development of an artificial intelligence-based automatic artifact identification tool for DNA profiling

研究代表者
眞鍋　翔（MANABE, Sho）

関西医科大学・医学部・助教

研究者番号：００７９４６６１

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）：　本研究では、DNA鑑定で検出されるアーチファクトを機械学習の一手法であるランダムフォレスト法により識別できるか検討した。使用するデータは、単一個人のDNA試料（350例）および2人から4人までのDNA混合試料（180例）から得られた全ピーク（43,158個）とした。全ピークの3/4を訓練データ、1/4をテストデータに振り分け、Pythonのライブラリscikit-learnを用いて機械学習を実施したところ、訓練データとテストデータの正解率はそれぞれ100%、約98.9%と非常に高い値が得られた。また、両データの正解率にほぼ差はなく、明らかな過学習は観察されなかった。

研究成果の学術的意義や社会的意義
DNA鑑定で扱われる試料は、複数人のDNAが混合した試料や量的に極めて少ない試料が多く、ヒトDNA由来のシグナルとアーチファクトを人の手で識別するのは困難である。本研究を通して、AIと既存のソフトウェアを組み合わせることで、人の手を介さなくても高精度でアーチファクトを判定できるようになった。もちろん、AIで100%正しく判定できるわけではないので、専門家によるレビューは欠かせないが、本研究成果は客観性が強く求められるDNA鑑定実務に大きく貢献できると考えられる。さらに、客観性の高いDNA鑑定が普及し、DNA鑑定の証拠能力が上がれば、犯罪立証だけでなく犯罪抑止にもつながるものと期待される。

研究成果の概要（英文）：In this study, we examined whether artifacts observed in forensic DNA testing can be identified by the random forest, which is one of the methods of machine learning. The data used were peaks (n = 43,158) obtained from DNA samples of a single-source profiles (n = 350) and mixed DNA samples from two to four individuals (n = 180). Three-fourths of all peaks were assigned to training data and one-fourth to test data, and machine learning was performed using the library "scikit-learn" for the Python programming language. The accuracy of both training and test data was 100% and approximately 98.9%, respectively. There was almost no difference in the accuracy for both data, and no obvious overfitting was observed.

研究分野：法遺伝学

キーワード：DNA鑑定　法医学　人工知能　アーチファクト　混合試料

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

法医学分野における DNA 鑑定では、犯罪現場に残された DNA が誰のものであるかを正しく判断することが求められる。しかし、扱う DNA は他の科学実験とは大きく異なり、微量かつ変性している場合が多い。また、犯罪現場に残された DNA は単一個人に由来するとは限らず、事件の被害者や犯人などの DNA が混合しているもの（混合試料）が多い。特に、極微量の DNA が混入しているものを検査した場合、その DNA に由来する本物のシグナル（アレル）と、検査システムの都合上検出されてしまうアーチファクトを人の手で鑑別するのは困難である。アレルとアーチファクトを正しく鑑別できないと、誰の DNA であるかを正しく判断できなくなる可能性がある。

近年ではこの問題を克服すべく、複雑な数理モデルを駆使して混合試料の解析を行う continuous model と呼ばれる手法（Taylor D: Forensic Sci Int Genet 2013）が考案され、私もこの手法に基づくソフトウェア“Kongoh”を開発した（Manabe S: PLoS ONE 2017）。これにより、誰の DNA であるかを推定するのに使われる尤度比を厳密に求めることが可能となり、一部のアーチファクトについては、適切な数理モデルを立てることで、人の判断で鑑別する必要がなくなった。しかし最近では、DNA 型の検査で用いられているマイクロサテライト（別名 short tandem repeat: STR）の検査試薬やキャピラリー電気泳動装置の改良が急速に進み、検出感度が大きく向上したため、従来の DNA 型検査に比べて検出されるアーチファクトの種類が膨大になった。このため、あらゆるアーチファクトを処理するための適切な数理モデルの構築には限界があり、私が開発したソフトウェアにおいても実装には至っていない。

また、最近になって人工知能（AI）が法医学分野でも着目されるようになり、AI にアーチファクトやノイズなどのシグナルを判断させる試みがなされるようになった（Taylor D: Forensic Sci Int Genet 2016）。しかし、予備検討に留まっているのが現状であり、アーチファクトの判定を AI が行うツールは未だ開発されていない。

２．研究の目的

本研究では、人の主観的判断に依らない DNA 鑑定を可能にすることを目的に、AI によりアーチファクトを自動的にどの程度判断できるかを検討した。

３．研究の方法

（1）AI に判断させるべきアーチファクトの選別

まず、6 種類のアーチファクト（back stutter: BS、forward stutter: FS、double-back stutter: DS、minus 2-nt stutter: M2S、pull-up: PU、その他: Other）のうち、AI に判断させるべきものを選別するために、各アーチファクトの数理モデルを人為的に構築できるか否かを検討した。数理モデルの構築が難しいアーチファクトについては、AI の判断に依存することとした。一方で、構築できたモデルについては、既存のソフトウェア“Kongoh”に実装することで処理できるようにした。“Kongoh”の開発には、R 言語を用いた。

（2）ノイズを排除するための閾値の設定

DNA 型の情報をキャピラリー電気泳動装置で検出する現行の検査法では、アレルやアーチファクトよりもシグナルの高さが小さいノイズが検出される。AI の判定精度を高めるために、ノイズを排除するための適切な閾値を設定した。

DNA を全く含まない溶液として、Low-TE buffer を 11 例用意し、現在の DNA 鑑定実務で使用されている GlobalFiler™ PCR Amplification Kit（Thermo Fisher Scientific）を用いて、DNA 型の検査を行った。検出されたノイズの高さを正規分布に近似し、平均＋10SD の高さを閾値とした。

（3）機械学習の手法について

本研究では、機械学習の手法としてランダムフォレスト法を採用した。ランダムフォレスト法は他の方法と比較して、過学習（手元のデータに対する予測結果は良い精度だが、未知のデータに対する予測結果の精度が悪い状態）を起こしにくい点、特徴量（分類の手掛かりとなる変数）を標準化・正規化してスケールを揃える必要がない点が優れている。

（4）機械学習のためのデータの準備

単一個人に由来する DNA 試料（350 例）および 2 人から 4 人までの混合試料（180 例）を実験的に作製した。その際に、DNA 量および混合比率については、DNA 鑑定実務で遭遇し得る様々な条件のものを準備した。次に、GlobalFiler™ PCR Amplification Kit を用いて、21 座位の STR および 3 座位の性別判定用マーカーの DNA 型検査を行った。得られたシグナルのうち、(2) で設定した閾値以上となったシグナル（43,158 個）について、既知の DNA 型情報を基にアレルおよび 6 種類のアーチファクトに分類した。続いて、機械学習の特徴量として、質的変数（座位の情報など）と量的変数（各シグナルの高さの比率や合計値など）を含む 47 種類を挙げた。

(5) ランダムフォレスト法による機械学習

まず、全シグナル(43,158 個)の 3／4 を訓練データ(32,368 個)、1／4 をテストデータ(10,790 個)に振り分けた。次に、訓練データを用いて、特徴量選択(分類に有効な特徴量の選別)、交差検証(過学習がないことの確認)、パラメータのグリッドサーチ(ランダムフォレスト法の条件検討)を行った。続いて、これらの結果を基に、訓練データを用いて再学習を行った。最後に、再学習の結果を基に、テストデータについてアーチファクトの判定を行い、得られた結果を評価した。一連の解析には、Python のライブラリ scikit-learn を用いた。

４．研究成果

(1) 4 種類のアーチファクトは数理モデルの構築が可能であった

6 種類のアーチファクトのうち、BS、FS、DS、M2S の 4 種類については、数理モデルの構築が可能であった。これらの数理モデルにより、アレルかアーチファクトかを無理に判断する必要はなく、確率的に扱うことが可能となった(アレルの確率が 70%、BS の確率が 30%など)。そこで、これらの数理モデルをソフトウェア“Kongoh”に実装し、公開した。

(2) ノイズを排除するための閾値を実験的に設定した

ノイズを排除するための最適な閾値は、50 RFU(relative fluorescence unit、シグナルの高さの単位)であることが明らかとなった。よって、以降の機械学習では、50 RFU 以上のシグナルを対象とすることにした。

(3) 過学習のない安定したランダムフォレスト法の条件を構築できた

特徴量選択の結果、量的変数を中心に 17 種類が採用された。交差検証の結果、データの分割の仕方によらず正解率(accuracy)は約 98.9%で安定していたことから、大きな過学習は発生していないと判断できた。また、パラメータのグリッドサーチにより、ランダムフォレスト法の条件を客観的に設定することができた。

(4) シグナルの判定精度は 98.9%であった

(3) の結果を基に訓練データの再学習を行った結果、正解率は 100%となった。そして、再学習の結果を基に、テストデータについてアーチファクトの判定をコンピュータに行わせたところ、正解率は約 98.9%と非常に高い結果となった。

(5) 実務で問題となり得る誤判定は 0.065%に留まった

図 1 は、テストデータ 10,790 個における判定結果の詳細であり、アレルおよび 6 種類のアーチファクトについて、縦軸に実際の答え、横軸に予測結果を示している。対角線上の数字が比較的大きく、正しく判定できた例数が多いことが読み取れる。一方で誤判定も認められており、殆どはアレル、BS、FS、DS、M2S の 5 つの中で生じていた。これらのアーチファクトは、いずれも数理モデルを構築することができたため、本研究で改良したソフトウェア“Kongoh”で確率的に扱えば、実務上問題とはならないと考えられた。一方、PU と Other については、AI(機械学習)の判断に依存することになるが、誤判定はわずか 0.065%に留まった。

以上より、AI の判定結果を基に PU と Other を除去した上で、ソフトウェア“Kongoh”での解析を実施することで、アーチファクトを適切に処理できると結論づけられた。

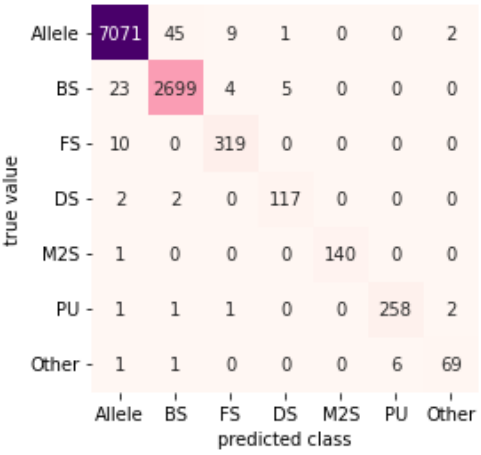


図 1　テストデータの判定結果

(6) 専門家によるレビューは欠かせない

本研究では、人の主観的判断に依らない DNA 鑑定を目指した。前述の研究成果より、AI とソフトウェア“Kongoh”を組み合わせることで、人の手を介さなくても高精度でアーチファクトを判定できるようになった。しかし、AI で 100%正しく判定できるわけではなかったので、AI やソフトウェアの結果に完全に依存するのではなく、専門家が結果をレビューすることは極めて重要であると考えられた。

５．主な発表論文等

〔雑誌論文〕　計3件（うち査読付論文　3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<br>Manabe Sho、Fujii Koji、Fukagawa Takashi、Mizuno Natsuko、Sekiguchi Kazumasa、Inoue Kana、Hashiyada Masaki、Akane Atsushi、Tamaki Keiji | ４．巻<br>52 |
| ２．論文標題<br>Evaluation of probability distribution models for stutter ratios in the typing system of GlobalFiler and 3500xL Genetic Analyzer | ５．発行年<br>2021年 |
| ３．雑誌名<br>Legal Medicine | ６．最初と最後の頁<br>101906～101906 |
| 掲載論文のDOI（デジタルオブジェクト識別子）<br>10.1016/j.legalmed.2021.101906 | 査読の有無<br>有 |
| オープンアクセス<br>オープンアクセスではない、又はオープンアクセスが困難 | 国際共著<br>－ |

| | |
|---|---|
| １．著者名<br>Manabe Sho、Fukagawa Takashi、Fujii Koji、Mizuno Natsuko、Sekiguchi Kazumasa、Akane Atsushi、Tamaki Keiji | ４．巻<br>54 |
| ２．論文標題<br>Development and validation of Kongoh ver. 3.0.1: Open-source software for DNA mixture interpretation in the GlobalFiler system based on a quantitative continuous model | ５．発行年<br>2022年 |
| ３．雑誌名<br>Legal Medicine | ６．最初と最後の頁<br>101972～101972 |
| 掲載論文のDOI（デジタルオブジェクト識別子）<br>10.1016/j.legalmed.2021.101972 | 査読の有無<br>有 |
| オープンアクセス<br>オープンアクセスではない、又はオープンアクセスが困難 | 国際共著<br>－ |

| | |
|---|---|
| １．著者名<br>Manabe Sho、Hashiyada Masaki、Akane Atsushi | ４．巻<br>－ |
| ２．論文標題<br>Internal validation for analytical and stochastic thresholds in the GlobalFiler system | ５．発行年<br>2022年 |
| ３．雑誌名<br>Japanese Journal of Forensic Science and Technology | ６．最初と最後の頁<br>－ |
| 掲載論文のDOI（デジタルオブジェクト識別子）<br>10.3408/jafst.845 | 査読の有無<br>有 |
| オープンアクセス<br>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<br>－ |

〔学会発表〕　計5件（うち招待講演　0件／うち国際学会　0件）

| |
|---|
| １．発表者名<br>眞鍋 翔、深川貴志、藤井宏治、水野なつ子、関口和正、玉木敬二、大林将弘、榎本祐子、松本智寛、橋谷田真樹、赤根 敦 |
| ２．発表標題<br>GlobalFiler kit に対応した混合試料解析ソフトウェアの検証 |
| ３．学会等名<br>第105次日本法医学会学術全国集会 |
| ４．発表年<br>2021年 |

１．発表者名
眞鍋翔、藤井宏治、深川貴志、水野なつ子、関口和正、井上花菜、赤根敦、玉木敬二

２．発表標題
混合試料解析ソフトウェアKongohのGlobalFiler対応に向けたstutter ratioのモデル作成

３．学会等名
日本法科学技術学会第26回学術集会

４．発表年
2020年

１．発表者名
眞鍋翔、深川貴志、藤井宏治、井上花菜、水野なつ子、関口和正、赤根敦、玉木敬二

２．発表標題
法医鑑識領域のDNA検査に対応したDNA混合試料解析ソフトウェアの開発

３．学会等名
日本DNA多型学会第29回学術集会

４．発表年
2020年

１．発表者名
眞鍋翔、橋谷田真樹、赤根敦

２．発表標題
GlobalFiler検査におけるanalytical thresholdとstochastic thresholdについてのinternal validation

３．学会等名
日本法科学技術学会第28回学術集会

４．発表年
2022年

１．発表者名
眞鍋翔、橋谷田真樹、赤根敦

２．発表標題
ランダムフォレスト法を用いたDNA鑑定における検出ピークの判別

３．学会等名
日本DNA多型学会第31回学術集会

４．発表年
2022年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

| | 氏名<br>（ローマ字氏名）<br>（研究者番号） | 所属研究機関・部局・職<br>（機関番号） | 備考 |
|---|---|---|---|

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 |
|---|---|