4ZC-04

JHU CSSE リポジトリデータと SARS-CoV-2 ゲノムデータの統合による新型コロナウイルスの Clade 別危険度分析

村上 唯希† 岡田 龍太郎† 中西 崇文† 北川 高嗣‡
武蔵野大学 データサイエンス学部 データサイエンス学科†
筑波大学大学院 システム情報系 コンピュータサイエンス専攻‡

1. はじめに

2020 年は，新型コロナウイルス感染症の流行により，未曾有の危機を招いた．感染者数と死亡者数が増え続ける中，未知なる新型コロナウイルスの感染に関する性質をいち早く見つけることは，世界において重要な課題となっている．

一方で，新型コロナウイルスに関するデータを積極的に公開しようとする動きが広まっている．代表的なものとしては， GISAID[1]では，各国が感染者から採取した新型コロナウイルスのゲノムデータを収集，公開している．ゲノム解析の結果と感染者数などの統計データを統合することで，変化し続けるゲノムの特徴と感染状況の関係を導く手がかりになると考えられる．

本稿では，JHU CSSE(the Center for Systems Science and Engineering at Johns Hopkins University)リポジトリデータ[2]と SARS-CoV-2 ゲノムデータ[1]を対象として，これらのデータの統合による新型コロナウイルスの Clade 別危険度分析について示す．本分析は，新型コロナウイルスにおける感染者数および死亡者数の統計データである JHU CSSE リポジトリデータと SARS-CoV-2 のゲノム解析結果である SARS-CoV-2 ゲノムデータを対象として，エピデミックおよびパンデミックのウイルスデータのオープンデータを管理公開している GISAID[1]が定義するゲノム情報の類似度によって分けられたウイルスの株である Clade ごと，国ごとの死亡率との相関を危険度として示す．本研究によって，ゲノム解析の成果を数値化し，新たなインサイトを見出すことが可能となる．

2. 関連研究

本節では，ゲノムデータを用いた関連研究を述べる．

Matsuda ら[3]は，ゲノム情報の変異の近さからグループ分けを行い，どの国から侵入してきたのか，国ごとにどういった変異の傾向があるかを示している．Matsuda らの研究[3]によって，国ごと，地域ごとにゲノムデータは大きく変化していることは解明しているが，ゲノムが変化した結果，感染状況にどれだけ影響があるかを示すことはできない．本分析では，このようなゲノムの変化から現れる死亡率の違いを相関により導出している．

3. 対象データ

本節では，本研究で使用しているデータについて述べる．3.1 節では SARS-CoV-2 ゲノムデータについて述べ，3.2 節では JHU CSSE リポジトリデータについて述べる．

3.1 SARS-CoV-2 ゲノムデータ(GISAID)

各国が感染者から採取した新型コロナウイルスの分析結果のデータで，採取した日付，分析を行った研究所の名前，1 節で説明した Clade など 27 項目がまとめられている．現在，Clade は 8 種類が定義されており，国ごとに流行っているものが異なる．

3.2 JHU CSSE リポジトリデータ

WHO や各国の情報機関が公開しているデータソースから，各国の感染者数や死亡者数など 12 項目をジョンズホプキンス大学システム科学工学センター[4]がまとめ，GitHub 上で公開しているリポジトリデータである．

4. データ統合による新型コロナウイルスの Clade 別危険度分析

本節では，本分析で行った新型コロナウイルスの Clade 別危険度分析手法について示す．4.1 節では，分析の概要を述べる．4.2 節では，3 節で述べた 2 つのデータの統合について述べる．4.3 節では，Clade 別危険度の算出について述べる．

4.1 分析の概要

本節では，分析手法の概要について述べる．分析手法の全体像は図 1 に示す．本研究の目的は，3 節で説明した SARS-CoV-2 ゲノムデータと JHU CSSE リポジトリデータを統合し，分析を行うことで，それぞれのデータだけでは見つからないインサイトを発見することである．本分析

Integration of JHU CSSE Repository Data and SARS-CoV-2 Genome Data for Risk Analysis by Clade
Yuuki Murakami†, Ryoutaro Okada†, Takafumi Nakanishi†, Kitagawa Takashi‡
†Musashino University, Department of Data Science
‡Graduate School of Systems and Information Engineering, University of Tsukuba

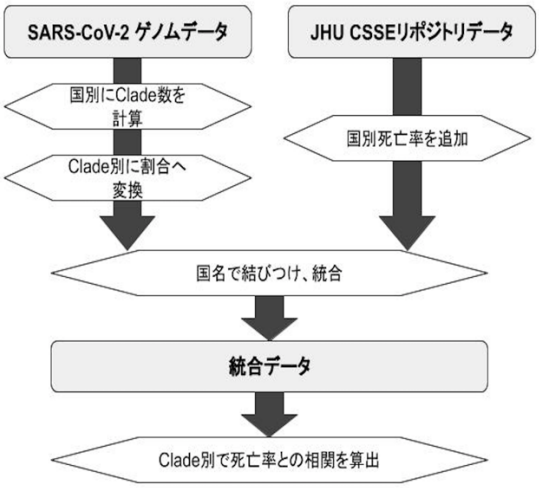

図 1：分析手法の全体像

では，各国の死亡率を危険度とし，Clade 別に相関を算出することで，Clade の中で死亡率に大きく関わっているものを探し出すことができる．

4.2 データ統合

本節では，データ統合の手法について述べる．SARS-CoV-2 ゲノムデータからは，国ごとの Clade 数をカウントし，各国で分析した Clade 数のうち，それぞれの Clade の割合を算出した．JHU CSSE リポジトリデータからは，国ごとの死亡者数と感染者数から死亡率を算出した．これらの 2 つのデータを統合することにより，Clade 別危険度の算出を試みた．なお，2 つのデータで国名の表記が異なる国については，名寄せした．

4.3 Clade 別危険度分析

本節では，Clade 別危険度分析の手法について述べる．4.1 節で述べたように，各国の死亡率を危険度としているため，死亡率と正の相関が強い Clade は危険度が高いと言い換えることができる．したがって，国ごとの Clade 別割合と死亡率から相関を算出した．ただし，Clade 合計数が少ない国を用いた結果では正確性に欠けることから，Clade 合計数が 100 以下の国を除外した．また，Clade 別危険度の変化を観察するため，10 月から 12 月の 3 ヶ月間，データの更新をした上で，それぞれの相関の計算を行った．

5. 分析結果

本節では，本分析の算出結果を示す．表 1 に Clade 別の死亡率との相関の示す．表 1 は，2020 年 10 月 11 日時点のデータセット，2020 年 11 月 10 日時点のデータセット，および 2020 年 12 月 14 日時点のデータセットでの相関係数を示している．

ここで，表 1 より，2020 年 10 月 11 日時点と

表 1：Clade 別の死亡率との相関係数

Clade 名	相関係数		
	2020/10/11	2020/11/10	2020/12/14
S	-0.046	-0.046	0.009
L	0.017	0.017	0.041
V	-0.180	-0.180	-0.149
O	-0.303	-0.303	-0.238
G	0.274	0.273	0.261
GR	0.018	0.018	-0.018
GH	-0.057	-0.057	-0.108
GV	-	-	0.026

2020 年 11 月 10 日時点の相関を比較して変化が乏しい理由として，GISAID[1]の新しい Clade づけに関する公開が遅れていたことが原因である．表 1 より，G 種と死亡率について強い正の相関があることがわかる．G 種が流行っている国は死亡率が高くなる傾向があり，極めて危険だということである．12 月に入り GV 種が現れているが，それほど強い相関は見られない．今後，引き続き新たな Clade 種が現れることも考えられるため，継続的に分析が必要である．

6. おわりに

本稿では，ゲノム解析から得られたデータと，各国の感染者数，死亡者数のデータを統合し，危険度を相関という数値を算出する方法を示した．本研究により，感染拡大の危険性が高い地域を見つけることが可能となり，感染症対策を決める指針の一つになると考えている．

今後の課題としては，感染拡大の危険性が高い地域の死亡率から，Clade 推定の実現が挙げられる．

参考文献

[1] GISAID, https://www.gisaid.org/
[2] COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University, https://github.com/CSSEGISandData/COVID-19
[3] Tomoko Matsuda, Hikoyu Suzuki, Norichika Ogata1, Phylogenetic analyses of the severe acute respiratory syndrome coronavirus 2 reflected the several routes of introduction to Taiwan, the United States, and Japan, arXiv preprint arXiv:2002.08802, 2020.
[4] he Center for Systems Science and Engineering (CSSE) at JHU, https://systems.jhu.edu/