

パンデミックに対してレジリエントな社会・技術基盤の構築
2022 年度採択研究代表者

2022 年度
年次報告書

清水 秀幸

東京医科歯科大学 M&D データ科学センター
教授

微生物学と AI の統合知による次世代の感染症制圧

研究成果の概要

1 年次目の研究項目は「研究項目 1. 病原体ゲノム配列情報から治療薬を迅速に探索し構造最適化を行う人工知能の開発」については微生物 CPI データを学習する高精度人工知能の開発と大規模機能アッセイデータを学習する人工知能の開発、「研究項目 2. 抗菌活性ペプチドを高精度で予測する人工知能の確立」については必要となるデータのダウンロード、人工知能学習前の整形であった。

さきがけ採択以前から取り組んできたアミノ酸配列のみから阻害剤を探索する AI である LIGHTHOUSE が査読付き学術誌に掲載された (iScience 2022)。これを受けてスタートしたさきがけ研究においては、LIGHTHOUSE を改良させ次のパンデミックに備える技術基盤の 1 つとしたい。LIGHTHOUSE は哺乳類データを学習しているため、微生物に特化した chemical protein interaction (CPI) データを学習する必要がある。そのため、数千種類の微生物の STITCH データベースに登録されている微生物の公開データ (細菌・ウイルスがほとんど) をダウンロードし整形後、UniProt および PubChem データベースと有機的に情報を紐付けるところまで完了している。

2 つ目のプロジェクトとして掲げている抗菌活性ペプチドの高精度予測については、公開されている ADAM、APD、CAMP、LAMP の 4 つのデータベースからダウンロードする。これらはいずれも実験的に検証された AMP やそのペプチド配列について収録しているデータベースであり、重複を除くと約 11,000 ペプチド配列が何らかの抗菌活性がある配列として登録されている。人工知能の構築にはこれら「ポジティブデータ」だけではなく AMP 活性がないペプチド配列、すなわち「ネガティブデータ」も必要になるが、これについては UniProt に登録されている細胞質タンパクのうち antimicrobial, antibiotic, antiviral, antifungal といった感染症との関連を示唆するアノテーション付けがされていない配列をダウンロードし重複を除いてネガティブデータを自前で構築した。