G-118

機械学習によるウイルスゲノムの分類システムの構築

Establishment of a classification system of viral genome sequences by machine learning

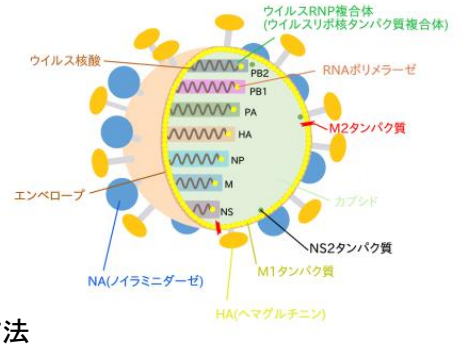
一李　建嬉　　鈴木　義紀　　溝脇　大智
Ichii Konhi,　Suzuki Yoshiki,　Mizowaki Taichi

1. 抄録

世界で猛威を振るう，インフルエンザ，麻疹，エボラ出血熱，デング熱などの原因ウイルスについて，核酸を構成している塩基配列（A，T，G，C）からゲノム情報を分類するシステムを、機械学習を用いて構築しようと試みた。

2. 研究の背景と目的

米国生物工学情報センターが公共の塩基配列データベース（GenBank）を提供している。その膨大なデータと機械学習を活用し，未知なるウイルスを発見するための分類システムの確立が最終目標である。本研究ではインフルエンザの３つの遺伝子，次にインフルエンザの 8 つの遺伝子，さらに複数のウイルスの遺伝子を混在させた塩基配列データを用い，ウイルスゲノムの分類に取組んだ。



3. 方法

3.1　データの収集と整備

GenBank からウイルスゲノムデータをダウンロード（Genbank 形式から FASTA 形式に変換）⇒ 教師用データの作成⇒A，T，G，C，混合塩基および不明な塩基の数値化　⇒ データ長の統一（個々のデータで長さが異なる）⇒ バイナリ形式に保存した。

3.2　分類器の動作

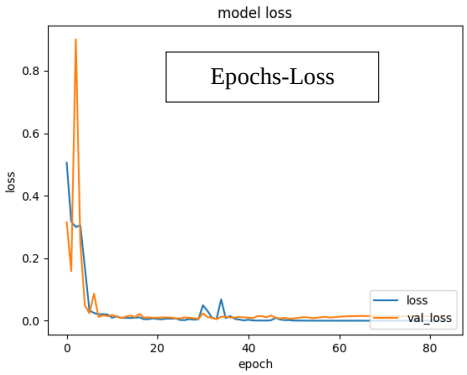
ウイルスゲノムデータの塩基配列を入力 ⇒ 塩基配列をベクトル表現に変換（Embedding 方式を採用）⇒ リカレントニューラルネットワーク (LSTM) で特徴を抽出 ⇒ Softmax 関数で活性化後に出力した

3.3　分類器の実装

Keras(TensorFlow backend)，BioPython，Numpy を使って Python に分類器を実装。すべてのプログラムは CPU：Corei7-8700K，Memory：16G，GPU：GTX1080Ti を搭載した自作 PC を用いた。処理の高速化を図るため、ミニバッチ処理を施した。

3. 結果

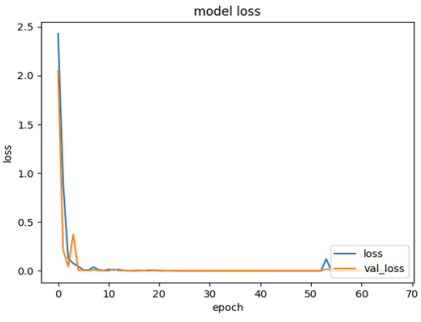
(1)第 1 期：
インフルエンザウイルスの 8 種の遺伝子のうち３つ（PA，PB1，PB2）の分類に成功(正答率：約 99.6%)。



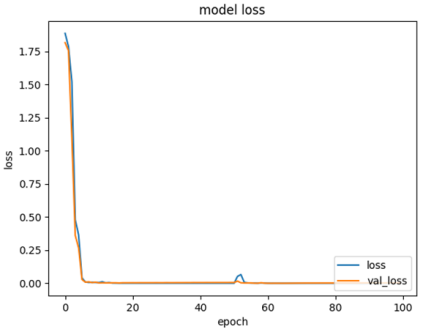
(2)第 2 期：
はしかを起こす麻疹ウイルスのＬ遺伝子を加え、4 遺伝子の分類に成功(正答率：99.7%)。

(3)第 3 期：
インフルエンザの遺伝子(7 種)＋麻疹のＬ遺伝子の分類(正答率：約 98.2%)。デングウイルスの遺伝子 10 種類の分類(正答率：約 99.9%)，エボラウイルスの 6 種類の遺伝子の分類(正答率：約 99.9%)，ロタウィルス の 11 種類位の遺伝子(正答率：約 99.7%)などに成功した。

デングウイルス:10 種類
epoch:100
中間層数:256
Sum:32458



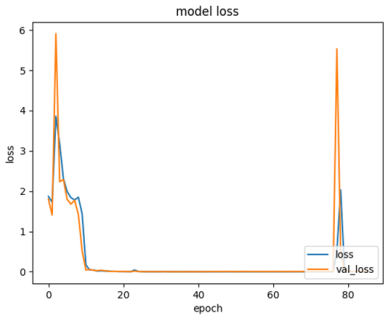
エボラウイルス:6 種類
epoch:100
中間層数:256
Sum:27811



†一般社団法人情報処理学会，IPSJ

‡ 学校法人大阪医科薬科大学　高槻高等学校

エボラウイルス:6 種類
epoch:100
中間層数:400
Sum:27811
※過学習と思われる



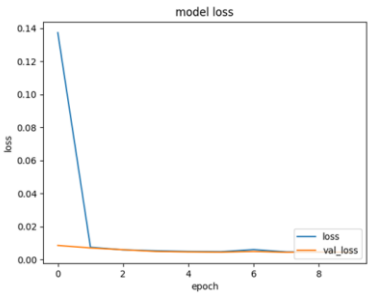
(4)第 4 期：

第 3 期まではデータ長に大きな違いがあるので、最大データ長に合わせ整備したデータセットを分類器に通す処理をしていた。新たに，データ長を 50，100，400 に切ったデータセットを作成し分類した。結果，インフルエンザの遺伝子 8 種類の分類の正答率が約 99.9%に向上し，処理にかかる時間が大幅に短縮された。

現在，“その他クラス”を設けた分類に着手している。

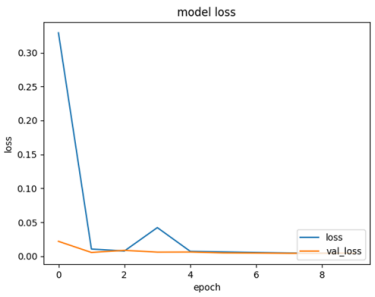
インフルエンザウイルス:8 種類
epoch:10
中間層数:256

データ長:
N=100



インフルエンザウイルス:8 種類
epoch:10
中間層数:256

データ長:
N=400



４．考察

個々のデータ長に大きな違いがある（数十から数千）ことが困難であった。データの整備およびその処理を試行錯誤で取り組み，メモリ不足などと戦った。データ長を固定値で切ってそろえることによって改善された。TestLoss 値から判断して，データ長 100 個ぐらいが良いと判断している。

教師あり学習において，ラベルに該当しないデータを「その他」として分類できるようにすることが，未知のウイルス発見の第 1 歩になると考えている。閾値に基づく分類を試みたがまだ良い結果は得られていない。閾値の設定の工夫，閾値以外の方法にも取り組みたい。

今後より多くのウイルスゲノムデータを学習させ、さらに、塩基配列に加えてアミノ酸配列の分類システムの構築も試み，制度・汎用性をより高めたい。

５．参考文献

[1] 岡崎康司，坊農秀雅（監訳）（2005）．バイオインフォマティクス　メディカル・サイエンス・インターナショナル　pp.350-362

[2] Kryukov, K., Ueda, M. T., Imanishi, T. & Nakagawa, S. (2018). Systematic survey of non-retroviral virus-like elements in eukaryotic genomes. Virus Research. In press.

[3] Gago, S., Elena, S. F., Flores, R. & Sanjuán, R. (2009). Extremely high mutation rate of a hammerhead viroid. Science 323, 1308–1308.