Vol.2020-BIO-64 No.3
2020/12/7

注意機構付き LSTM によるウイルスとヒトタンパク質間相互作用の予測

築山翔 [1,a)]　Md. Mehedi Hasan[1,b)]　藤井聡 [1,c)]　倉田博之 [1,d)]

概要：ウイルスは宿主のタンパク質と相互作用することにより、細胞周期や細胞死の制御、自身の遺伝物質の宿主への核内輸送といった処理を行い、増殖を促進すると同時に自身のライフサイクルを達成する。そのため、ヒトとウイルス間のタンパク質相互作用を特定することはウイルスの感染メカニズムを理解し、薬物ターゲットの発見に必須である。また、実験的な手法によるタンパク質間相互作用の特定は多くの時間と労力を要するため、全てのタンパク質の組み合わせに関して実行することは難しい。そこで、本研究では深層ニューラルネットワークを利用した計算論的アプローチにより、ヒトとウイルスのアミノ酸配列からの相互作用の予測を行なった。可変長データであるアミノ酸配列を扱うための LSTM ユニットを用い，相互作用するアミノ酸配列を予測するための注意機構を込みこんだ。我々のモデルは AUC>0.97 の精度でタンパク質間相互作用の予測を行った。相互作用するアミノ酸配列部位とタンパク質結合ドメインの関係を解析する。

キーワード：ウイルス-ヒトタンパク質間相互作用、LSTM、attention weight、結合ドメイン・モチーフ

Prediction of virus-human protein interactions by LSTM with attention mechanism

SHO TSUKIYAMA[1,a)]　MD. MEHEDI HASAN[2]
SATOSHI FUJII[3]　HIROYUKI KURATA

Received: November 11, 2020, Accepted: November 11, 2020

Abstract: Viruses interact with host proteins to achieve their life cycle and to promote proliferation through processes such as the control of the cell cycle and cell death and nuclear transport of their genetic material into the host's nucleus. Therefore, identifying protein interactions between human and viral proteins is essential for understanding the mechanisms of viral infections and finding drug targets. Since the experimental identification of protein-protein interactions is time-consuming and labor-intensive, it is difficult to implement all human and viral protein combinations. In the present study, we tried to predict the interactions from amino acid sequences by a computational approach based on deep neural networks. The deep neural network model is constructed of the LSTM units to process the variable-length amino acid sequences and of the attention mechanism to identify local amino acid sequences related to the interaction. Our model predicted protein-protein interactions with AUC of >0.97. We analyze the relationship between interacting amino acid sequence sites and protein-binding domains.

Keywords: Virus-Human protein-protein interactions, LSTM, attention weight, binding domain and motif

1. はじめに

パンデミックをもたらしている新型コロナウイルスの現状からも分かるように、ウイルス感染は健康に害を及ぼす大きな原因の一つであると言える。ウイルスは宿主の機能を乗っ取り、利用することによって自身のライフサイクルを達成し、増殖している。この目的を達成する過程で、ウイルスは宿主のタンパク質と相互作用することで、細胞周期

1 九州工業大学
Kyushu Institute of Technology, Iizuka, Fukuoka 820-0067, Japan
a) tsukiyama.sho675@mail.kyutech.jp
b)mehedicuhk@gmail.com
c) sfujii@bio.kyutech.ac.jp
d) kurata@bio.kyutech.ac.jp

やアポトーシスの制御、自身の遺伝物質の宿主核内への輸送といった処理を行う[1][2]。そのため、ウイルスの感染メカニズムの理解と薬物ターゲットの発見において、ウイルスとヒトの間のタンパク質相互作用を特定することは重要である。しかし、種内タンパク質相互作用に比べて、種間タンパク質相互作用は特定されているものが少ない。

相互作用の特定において、酵母ツーハイブリッド（Y2H）や質量分析法を利用した実験的な方法は広く使用されているが、多くの時間と労力を要するという問題点がある。そのため、全てのタンパク質の組み合わせについて実験的な方法を適用することは難しい。そこで、本研究では深層ニューラルネットワークを利用した計算論的アプローチにより、ヒトとウイルスのアミノ酸配列からタンパク質間相互作用の予測を行なった。

アミノ酸配列はタンパク質ごとに長さが異なるため、予測や分類から独立した前処理により固定長にエンコーディングされることが多い。しかし、予測や分類と独立したエンコーディングを行なった場合、アミノ酸の並びに関する特徴量が抜け落ちる可能性がある。そこで、本研究では、可変長データの処理が可能な LSTM ユニットを利用し、アミノ酸配列に関連した可変長の特徴行列から深層ニューラルネットワークを用いてエンコーディングと予測の両方を行なった。また、相互作用の予測に加えて、相互作用に関連した局所的なアミノ酸配列を特定するために、attention weight を深層ニューラルネットワークに組み込んだ。本論文では相互作用の予測に関する方法と結果について主に述べるが、attention weight を利用した今後の解析について最後に示す。

2. 方法

2.1 データセットの生成

相互作用のデータを Host-Pathogen Interaction Database 3.0 (HPIDB 3.0)からダウンロードした。HPIDB 3.0 は様々な宿主と病原体の間のタンパク質相互作用がアミノ酸配列情報と共に登録されている。ダウンロードしたデータからウイルスとヒトの間のタンパク質相互作用を抽出した。その後、抽出した相互作用データから、以下の処理により絞り込みを行なった。初めに、MIscore が 0.3 および 0.4 以上ある相互作用データを抽出した。MIscore は、IntAct と VirHostNet からの宿主と病原体の間のタンパク質相互作用における HPIDB に付与されている信頼度である[3]。MIscore の絞り込みの後、identity の閾値を 0.9 として CD-HIT を利用することで、冗長な相互作用を排除した。最後に、標準アミノ酸以外のアミノ酸を持つタンパク質、50 残基より短いタンパク質、および 3000 残基より長いタンパク質を含む相互作用を排除した。これらの処理により絞り込まれた相互作用データを陽性データとした。MIscore が 0.3 と 0.4 の場合のヒトとウイルスにおけるタンパク質の種類数を表 1 にまとめる。

表 1　ウイルスとヒトのタンパク質種類数

Miscore	ウイルスのタンパク質種類数	ヒトのタンパク質種類数
0.3	1153	7153
0.4	629	1775

次に陰性データの生成を行なった。我々の知る限り、ある根拠に基づいた非相互作用のデータベースは存在しない。そのため、相互作用が確認されていないタンパク質の組み合わせを陰性データとする必要がある。本研究においても陽性データにおけるタンパク質を参照して陰性データの生成を行なった。この目的において、相互作用が確認されていないタンパク質の組み合わせの中から無作為に抽出し、陰性データとするランダムサンプリング法はよく使われる方法である。しかし、この方法では相互作用するデータが誤って含まれる可能性が高いという問題点が過去の研究で報告されている[4]。そこで本研究では、Dissimilarity-based negative sampling 法を利用して陰性データの生成を行なった[4]。Dissimilarity-based negative sampling 法ではウイルスタンパク質の配列類似度に基づき、非相互作用である可能性の高いデータを生成する。以下に Dissimilarity-based negative sampling 法を利用した陰性データの生成アルゴリズムについて述べる（アルゴリズム 1 参照）。初めに、ウイルスのタンパク質の全ての組み合わせについて BLOSUM30 を使った Needleman-Wunsch アルゴリズムによりグローバルアライメントを適応し、配列類似度を計算した。その後、多くのウイルスとの間で配列類似度の低くなったウイルスを外れ値と見なし、陰性データ生成のためのウイルスタンパク質から除外した。本研究では、以下に示す閾値Tsより低い類似度を 9 割以上のウイルスとの間で示したウイルスを外れ値とみなした。

$$Ts = fq_i - 1.5 \times ir_i$$

ここで、fq_iはi番目のウイルスに関する類似ベクトルの第一四分位数、ir_iを四分位範囲とする。次に、各ウイルスの類似ベクトルに関して、最大値が 1、最小値が 0 となるように正規化を行なった後、最大値 1 から正規化後の類似度を引くことで、距離を算出した。あるウイルスタンパク質V_iとの距離が閾値Tより大きいウイルスタンパク質V_jと相互作用するヒトタンパク質を$H(V_j)$とするとき、ウイルスタンパク質V_iとヒトタンパク質$H(V_j)$の間の相互作用が確認されていない場合、このタンパク質ペアを陰性データの候補とした。本研究では、先行研究に基づき、閾値Tの値を 0.8 とした[4]。陰性データに含まれる各ウイルスのデータ数をなるべく合わせるため、陰性データの中でn個より多くのヒトタンパク質との間でペアを持つウイルスタンパク質については、そのヒトタンパク質の中からn個をランダムに選

び出し、ウイルスとヒトタンパク質の組み合わせを陰性データに含めた。一方、ペア数がn以下であるウイルスタンパク質については、全てのヒトタンパク質との組み合わせを陰性データに含めた。MIscore が 0.3 のデータにおいては n の値を 23 と 120、MIscore が 0.4 のデータにおいては n の値を 4 と 206 とすることで、陽性データと陰性データの数がほぼ等しい均衡データと陰性データが陽性データの数の約 5 倍である不均衡データを生成した。サンプリングによるバイアスを減らすために 3 つの異なるデータセットを用意した。条件ごとのデータ数を表 2 に示す。各データにおけるラベルはスカラー値とし、陽性データは 1、陰性データは 0 とした。

アルゴリズム 1　Dissimilarity-based negative sampling

for $V_i \in V$
　for $V_j \in V$
　　$BitScore(V_i, V_j) = Global\ Align(V_i, V_j)$
$[V', BitScore'] = RemoveOutliers(V, BitScore)$
for $V_i \in V'$
　for $V_j \in V$
　　$NormBitScore(V_i, V_j) = \frac{BitScore'(V_i, V_j) - \min(BitScore'(V_i, *))}{\max(BitScore'(V_i, *)) - \min(BitScore'(V_i, *))}$
　　$Distance(V_i, V_j) = 1 - NormBitScore(V_i, V_j)$
　　if $Distance(V_i, V_j) > T$ and $H(V_j) \notin H(V_i)$
　　$Candidate_i < -(V_i, H(V_j))$
　$Nnumber\ of\ Candidate_i \leq n$
　　$negative\ data < -Candidate_i$
　$Otherwise$
　　$negative\ data < -RandomPick(Candidate_i, n)$

表 2　各条件でのデータ数

Miscore	サンプリング数(n)	全データ数	陽性データ	陰性データ
0.3	23	53709	26997	26712
0.3	120	161508	26997	134511
0.4	4	5903	2858	3045
0.4	206	17144	2858	14286

2.4 特徴行列

我々はアミノ酸配列から深層ニューラルネットワークに入力するための特徴行列の生成を行なった。特徴行列内にアミノ酸配列の並びに関する情報を確保するため、各アミノ酸の 7 次元の特徴ベクトルを連結することで特徴行列を得た。各アミノ酸の特徴ベクトルの要素は、疎水性度、親水性度、側鎖の体積、極性、分極率、溶媒露出表面積、側鎖の実効電荷指数とした。これらの特徴は、先行研究を参考にして選択され、先行研究において使用された値を使用した[5]。

2.3 深層ニューラルネットワークの構築

深層ニューラルネットワークを 3 つのサブネットワークで構成することにより、予測に加えて特徴行列のエンコーディングを行った。2 つのネットワークは、それぞれ、ヒトとウイルスの特徴行列から相互作用の予測に必要な情報を抽出し、固定長ベクトルにエンコーディングするためのネットワークである。ここで、ヒトとウイルスにおけるネットワークは同じ構造を持つ。もう 1 つのネットワークは下流のネットワークにより生成された固定長ベクトルから相互作用の予測を行うためのネットワークである（図 1 参照）。

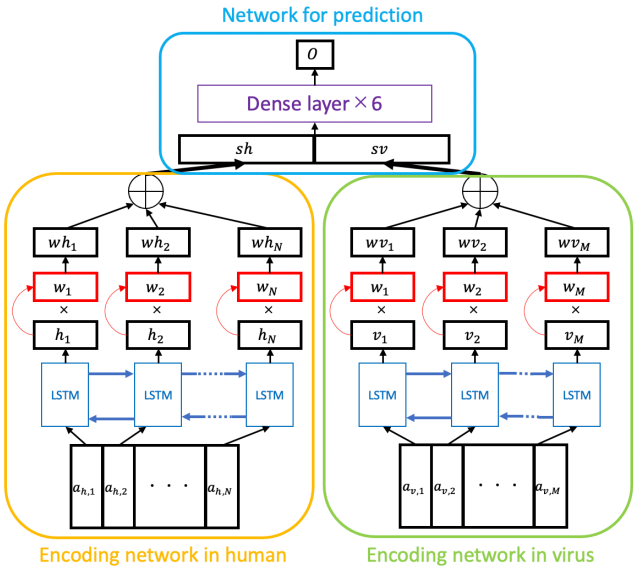

図 1　深層ニューラルネットワークの構造

特徴行列における各アミノ酸の特徴ベクトルは LSTM ユニットのそれぞれのステップに入力される。LSTM ユニットは双方向に展開され、N 末端から C 末端方向と C 末端から N 末端方向に出力を伝搬する。LSTM ユニットは各方向から 32 次元のベクトルを出力し、両方向のベクトルを水平に連結することで生成された 64 次元のベクトルを次の層に伝搬する。過学習を防ぐため、LSTM ユニットに Dropout を適用した。ここでの Dropout 率は 0.5 とした。

深層ニューラルネットが推論の際、注目したアミノ酸配列を特定するため、LSTM ユニットの出力ベクトルに 2 層の全結合層を適用することで、各ステップにおける attention weight を生成した（図 2 参照）。1 層目の全結合層からの出力は 32 次元のベクトルとした。活性化関数として ReLU 関数を使用し、出力に Dropout を適用した。ここでの Dropout 率は 0.5 とした。2 層目の全結合層はスカラー値を出力する。各ステップにおいて出力されたスカラー値を連結し、softmax 関数を適用することで attention weights を生成した。Dropout 率を 20%とし、attention weights に Dropout を適用した。各ステップにおける attention weight を LSTM の出力ベクトルと同じ 64 次元のベクトルにブロードキャストした後、LSTM の出力に掛け合わせた。

重み付けされた LSTM ユニットの各ステップにおける出力ベクトルを足し合わせることで固定長のベクトルを生成した。ヒトとウイルスにおける特徴行列から生成されたベクトルを水平に連結し、予測のためのネットワークに入力した。説明の都合上、それぞれのネットワークから生成されたベクトルを連結することによって得られたベクトルを

「中間ベクトル」と呼ぶ。中間ベクトルから相互作用の予測を行うためのネットワークは、6 層の全結合層からなる。全結合層の出力ベクトルの次元数は、ニューラルネットワークの入力側から、64、32、16、8、4、1 とし、活性化関数は Sigmoid 関数を使用した。最後の層以外の 5 層の全結合層からの出力ベクトルには Dropout と Batch normalization を適用した。ここでの Dropout 率は 0.5 とした。

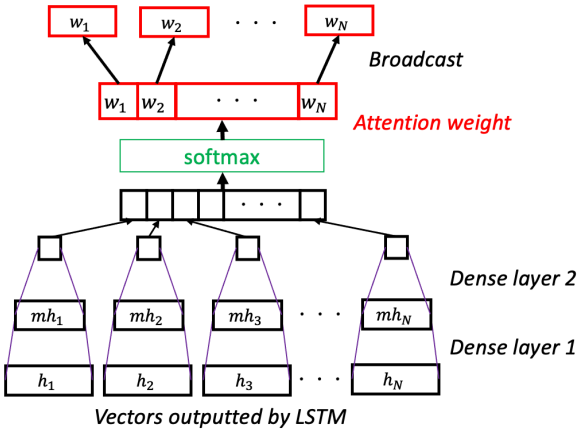

図 2　attention weights の生成

2.4 深層ニューラルネットワークの学習

データセットは 8:2 の割合で、学習時に使用する訓練データと検証時に使用する検証用データにランダムに分割された。本研究ではミニバッチ学習を適用し、バッチサイズを 256 とした。また、ミニバッチの間でのラベルの比率がほぼ等しくなるように各ミニバッチを決定した。学習率は、0.001 とし、最適化関数として RAdam optimizer を使用した[6]。

様々な分類・予測の問題において、ラベル数の割合が均一でない不均衡データを使用する場面がある。しかし、機械学習を用いたモデルの学習時に不均衡データを使用した場合、どちらかのラベルのみを出力するモデルとなる可能性がある。そこで、不均衡データの学習時には under-sampling、over-sampling、cost-sensitive learning といった手法が利用される。under-sampling は、数の多いラベルにおけるデータ数を減らす方法であり、over-sampling は逆に数の少ないラベルにおけるデータ数を増やす方法である。また、cost-sensitive learning では損失関数に重み付けを行う方法である。under-sampling では、重要な特徴を持つデータのサンプリングを適切に行う必要があり、特徴パターンを減少させてしまう可能性がある。また、over-sampling では重複したデータを使った学習により過学習を引き起こしてしまう可能性がある。そこで、本研究では Cui らによって報告された重み付けを用いて、cost-sensitive learning を行なった[7]。使用した損失関数を以下に示す。

$$CE(p,y) = \frac{1-\beta^{n_y}}{1-\beta}\{(y \times logx + (1-y) \times \log(1-x))\}$$

重み付けを適用する前の損失関数として binary cross entropy loss function を使用した。ここで、yは正解ラベル、xはモデルにより予測された相互作用する確率、n_yはミニバッチ 内のラベルがyであるデータ数、βはハイパーパラメータである。先行研究より、βを 0.9999 とした[7]。

過学習を防ぐため、35 エポック連続して検証用データの損失誤差が最小値を更新しない場合、学習を終了した。2.5 で述べる中間ベクトルの可視化では、検証用データの損失誤差が最小値となったエポックのモデルを解析に使用した。

2.5 中間ベクトルの可視化

2.3 において述べたように、我々の深層ニューラルネットワークは 3 つのネットワークから構成される。そのうち 2 つのネットワークはヒトとウイルスそれぞれの特徴行列からエンコーディングを行い、固定長ベクトルを生成するためのネットワークである。この 2 つのネットワークが、どの程度相互作用に関連した特徴を抽出することができたかを調べるため、中間ベクトルの可視化を行なった。中間ベクトルについては 2.3 を参照する。それぞれの条件において、訓練された深層ニューラルネットワークのモデルを使用して、全データ（訓練用データと検証用データ）の相互作用の予測を行い、予測時の中間ベクトルを取得した。中間ベクトルの可視化を行うための次元削減のアルゴリズムとして t-SNE を使用した[8]。T-SNE についての簡単な説明を以下に述べる。T-SNE では、高次元データ空間の点x_iとx_jの間の距離を以下の同時確率分布関数を用いて表す。

$$p_{ij} = \frac{p_{i|j} + p_{j|i}}{2}$$

$$p_{j|i} = \frac{\exp(-\|x_i - x_j\|^2/2\sigma_i^2)}{\sum_{k \neq i}\exp(-\|x_i - x_k\|^2/2\sigma_i^2)}$$

ここで$p_{i|i}$は 0 とし、σ_i^2は以下の式を満たすように決定される。

$$Perp(P_i) = 2^{H(P_i)}$$

$$H(P_i) = -\sum_j p_{j|i}\log_2 p_{j|i}$$

$Perp(P_i)$は Perplexity と呼ばれるハイパーパラメータである。次元削減後の低次元データ空間の点y_iとy_jの間の距離を以下の同時確率分布関数を用いて表す。

$$q_{ij} = (1 + \|y_i - y_j\|^2)^{-1} \Big/ \sum_{k \neq i}(1 + \|y_i - y_j\|^2)^{-1}$$

カルバック・ライブラー情報量を用いた以下の損失関数を定義し、確率的勾配降下方により最適化することで次元削減後の点を決定する。

$$C = \sum_i KL(P||Q) = \sum_i \sum_j p_{ji} log \frac{p_{ji}}{q_{ji}}$$

本研究では、Perplexity の値を 50 とし、更新時の学習率を

200 とした。勾配ノルムが1.0×10^{-7}より小さくなった場合は最適化を打ち切り、最適化の最大繰り返し回数を 5000 とした。

3. 結果と考察

深層ニューラルネットワークを評価するための指標として、sensitivity、specificity、accuracy、Matthews correlation coefficient（MCC）、Area Under Curve（AUC）、F1 score を使用した。陰性と陽性を分けるためのカットオフ値は、以下の式で与えられる Youden index が最大となる値とした。

$$Youden\ index = Sensitivity + Specificity - 1$$

条件ごとに、3 つのデータセットにおける上記の 6 つの指標とカットオフ値、学習エポック数を平均化した。それらの結果を表 3 に示す。全ての条件において AUC が 0.97 を超える精度が得られた。また、不均衡データにおける MCC、F1 score の値は 0.85 以上となっており、少数派データである陽性データを正しく特定できていることがわかる。この結果より、不均衡データを使用した深層ニューラルネットワークの学習において cost-sensitive learning は有用であると考えられる。

表 3　各条件における予測結果

Miscore	0.3		0.4	
サンプリング数(n)	23	120	4	206
学習エポック数	204	135	186	117
カットオフ値	0.42	0.41	0.55	0.48
Sensitivity	0.950	0.950	0.949	0.922
Specificity	0.969	0.971	0.980	0.990
Accuracy	0.960	0.970	0.965	0.979
MCC	0.920	0.896	0.931	0.923
AUC	0.992	0.991	0.994	0.976
F1 score	0.961	0.904	0.964	0.935

次に予測時に生成された中間ベクトルの可視化を行なった。それぞれの条件における結果を図 3 に示す。どの条件においても、陽性データと陰性データは、ほぼ分離しており、深層ニューラルネットワークの高い精度での予測を反映していることがわかる。一部のデータセット において、陰性データと陽性データの間に偽陽性や偽陰性が位置していることが観測できる。これらのデータは、陽性データと陰性データの間で同じような中間ベクトルが生成されたことにより、正しく予測を行うことができなかったものであると推測される。また、陰性データ内でも複数のクラスターが現れた。この現象は、陰性データが陽性データ内のタンパク質から構成されたことにより、複数の類似した非相互作用パターンが陰性データ内に含まれたからであると考えられる。

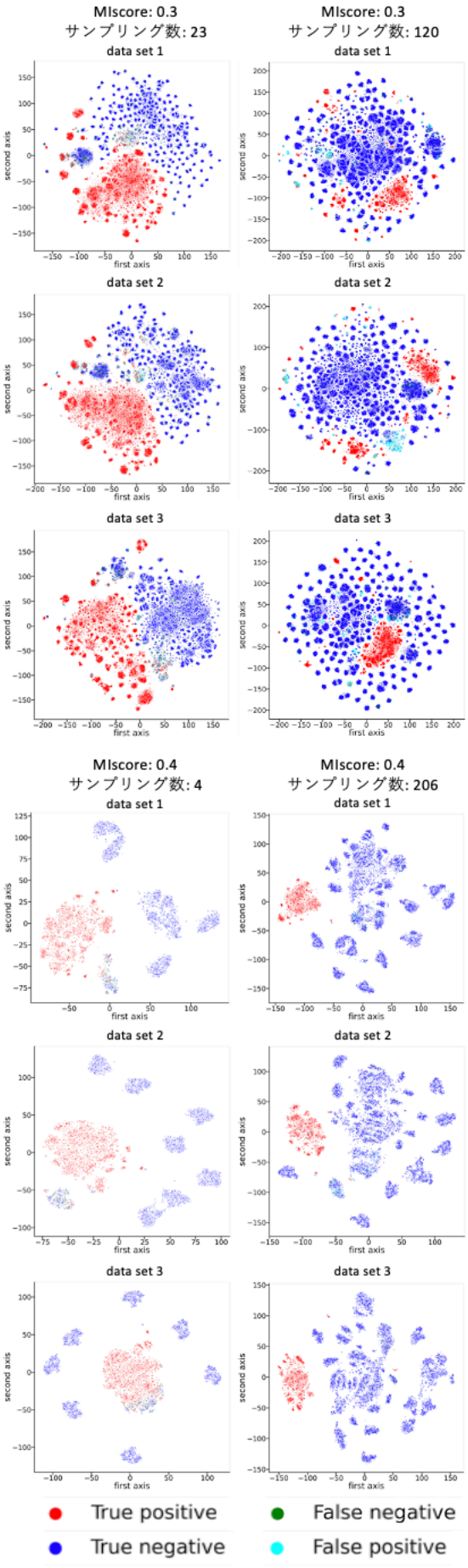

図 3　t-SNE による可視化

4. 今後の解析

今後の解析では、深層ニューラルネットワークにおいて推論時に生成された attention weights を用いて、以下の 2 つの解析を行う。1 つ目の解析では、相互作用する 2 つのタンパク質におけるアミノ酸の構造上の距離と attention weights の関係について調べる。2 つ目の解析では、既知の結合ドメインもしくは結合モチーフを持つ相互作用する 2 つのタンパク質について、その結合ドメイン・モチーフ内の attention weights を調べる。

5. おわりに

本研究では深層ニューラルネットワークを使用し、ウイルスとヒトタンパク質の相互作用の予測を行なった。さらに、エンコーディングのためのネットワークから生成された中間ベクトルを可視化した。我々の深層ニューラルネットワークのモデルは、いずれの条件においても 90%以上の精度で予測を行なった。また、中間ベクトルの可視化において、次元削減後の点は相互作用と非相互作用の間で分離した。これらの結果より、我々のニューラルネットワークが高い精度で予測を行うことができたのは、エンコーディング のためのネットワークにおいて、相互作用に関係する重要な特徴を各アミノ酸配列から抽出することができたからであると考えられる。今後の課題はネットワーク内に組み込まれた attention weights を調べ、深層ニューラルネットワークの予測における根拠を明らかにすることで、薬物ターゲットとなる局所的な配列部位の特定を行ことである。

参考文献

[1] Dyer, M. D. et al.. The landscape of human proteins interacting with viruses and other pathogens. PLOS pathogens, 2008, vol. 4 (2), p.e32-e32.
[2] Yang, S. et al.. Understanding human-virus protein-protein interactions using a human protein complex-based analysis framework. mSystems, 2019, vol. 4 (2), p.e00303-18.
[3] Liu-Wei, W. et al.. Prediction of novel virus-host interactions by integrating clinical symptoms and protein sequences. bioRxiv, 2020, DOI: https://doi.org/10.1101/2020.04.22.055095.
[4] Eid, F.E. et al.. DeNovo: virus-host sequence-based protein-protein interaction prediction. Bioinformatics, 2016, vol. 32 (8), p. 1144-1150.
[5] Lian, X et al.. Machine-learning-based predictor of human-bacteria protein-protein interactions by incorporating comprehensive host-network properties. Journal of proteome, 2019, vol. 18 (5), p. 2195-2205.
[6] Liu, L. et al.. On the variance of the adaptive learning rate and beyond. International Conference on Leaning Representations 2020.
[7] Cui, Y. et al.. Class-balanced loss based on effective number of samples. Conference on Computer vision and pattern recognition, 2019.
[8] Maaten, L. V. D. et al.. Visualizing data using t-SNE. Journal of Machine Learning Research, 2008, vol. 9 (86), p. 2579-2605.