Vol.2022-BIO-69 No.19
2022/3/11

Attention 機構ベースの深層学習による ヒト-ウイルスタンパク質間相互作用予測手法の提案

築山翔 [1,a)]　倉田博之 [1,b)]

概要： 現在の SARS-CoV-2 の感染状況からもわかるように、ウイルスは出現から急速に感染が拡大する。ウイルスは細胞に侵入し、宿主の機能を利用することで、自身の複製を増殖させる。それらのプロセスにおいて、ヒト-ウイルスタンパク質間の相互作用(human-virus protein-protein interaction; HV-PPI)は重要な役割を担う。そのため、ウイルスの感染メカニズムの理解と抗ウイルス薬の開発において、HV-PPI を特定することは重要である。本研究では、文脈情報を表現するための手法である word2vec を利用したアミノ酸配列の符号化に加えて、Interactive proteome-based encoding method (IPEM) と名付けた新しい符号化手法を提案した。この方法では、PPI ネットワークとマルチプルアライメントを用いることで、予測対象となるタンパク質と相互作用するタンパク質の配列情報を符号化する。さらに、注意機構と CNN に基づくニューラルネットワークを構築し、HV-PPI の予測を行った。その結果、我々のモデルは最先端の手法よりも高い精度で HV-PPI を予測した。

キーワード：深層学習、注意機構、タンパク質相互作用予測、ウイルス感染症、SARS-CoV-2

1. はじめに

ウイルス感染症は世界の深刻な健康問題の一つである。SARS-CoV-2 は驚異的な速さで感染が拡大し、世界的なパンデミックを引き起こしている。世界保健機関(WHO)によると 2021 年 12 月までに世界の累計感染者数は 2 億 8000 万人、累計死者数は 500 万人を超えている[1]。ウイルスは、細胞表面の受容体に結合したり、膜融合やエンドサイトーシスなどを誘発したりすることで、宿主の細胞に侵入する。また、宿主の機能を制御することにより、自身を複製するのに適した環境を構築し、増殖する。そのような処理において、宿主とウイルスタンパク質間相互作用は重要な役割を担う。そのため、ウイルスの感染メカニズムの理解と抗ウイルス薬の開発において、ヒト-ウイルスタンパク質間の相互作用(human-virus protein-protein interaction; HV-PPI)を特定することは重要である。HV-PPI を特定するために、Yeast two-hybrid 法や質量分析法などの実験的手法が広く使用されている。しかし、実験的な方法では多くの労力とコストがかかるため、全てのタンパク質の組み合わせについて実行するのは難しい。このような実験的方法の問題を補完する為に、PPI を予測するための様々な計算論的手法が開発されている。特に、構造情報を用いた予測手法では高い性能で相互作用を予測することが可能となっている。しかし、構造情報を明らかにすることは容易ではなく、また、構造情報が明らかになっているタンパク質の数は限られる為、新規ウイルスや変異株に対して即座に適用することは困難である。このような背景から、本研究では Attention 機構ベースの深層学習を利用することで、アミノ酸配列のみから HV-PPI を予測することに取り組んだ。

2. 方法

2.1 アミノ酸のエンコーディング

我々は、word2vec に基づくアミノ酸配列の符号化に加えて、interactive proteome-based encoding method (IPEM)と名付けた新規符号化手法により特徴行列を生成した。

2.1.1 word2vec によるアミノ酸配列の符号化

word2vec は単語の分散表現を生成する手法である。図 1 に示すように、本研究ではアミノ酸配列を連続した 4-mer で表し、word2vec モデルにより、その 4 量体の文脈情報をベクトルに畳み込んだ。UniProtKB/Swiss-Prot データベース[2]のタンパク質のアミノ酸配列を取得した後、標準アミノ酸以外のアミノ酸で構成された配列や長さが 9000 残基以上の配列を除外し、閾値を 0.9 として CD-HIT を適用することで word2vec の学習データを作成した。word2vec モデルの学習では、周辺 4-mer の幅を 5、学習回数 100 とした。配列の符号化では、学習した word2vec モデルにより、配列

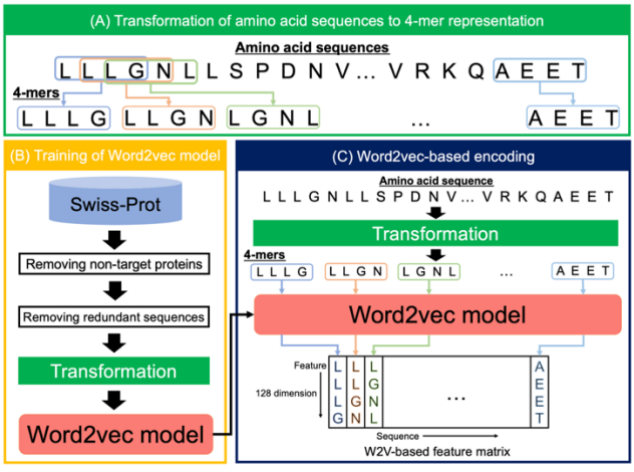

図 1 word2vec によるアミノ酸配列の符号化

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

中の連続する 4-mer をそれぞれ 128 次元の特徴ベクトルに変換し、その特徴ベクトルを配列方向に連結した。さらに、最大配列サイズが 9000 となるようにゼロパディングを施し、(8997×21)の形状を持つ特徴行列を生成した。

2.1.2 interactive proteome-based encoding method (IPEM)

本研究では、クエリタンパク質(予測対象のヒトとウイルスのタンパク質)に相互作用しうるタンパク質の配列情報から特徴行列を生成するために interactive proteome-based encoding method(IPEM)と呼ばれる新規エンコーディング方法を開発した。まず初めに、クエリタンパク質に相互作用するタンパク質、またはクエリタンパク質と相同なタンパク質と相互作用するタンパク質を特定する為に、PPI ネットワークを構築した (図 2)。その構築では、BioGRID[3]のヒトタンパク質間 PPI と HVIDB[4] の HV-PPI を使用した。各データセットから取得した PPI のうち、非標準アミノ酸を含むタンパク質や長さが 9000 残基以上のタンパク質で構成される PPI は除外され、698502 のヒトタンパク質間 PPI と 41787 の HV-PPI が残った。PPI ネットワークは無向グラフで表され、ノードはタンパク質に対応する。また、2 つのタンパク質が相互作用する場合、それらのタンパク質に対応するノードはエッジで接続される。各データセットの評価データにおける PPI を未知のものとするため、それらの PPI に対応するエッジは切断された。

クエリタンパク質に相互作用するタンパク質を特定する際、ネットワーク内にクエリタンパク質のノードが存在しない場合や存在するがノードが 5 つ以上のエッジを持たない場合は、5 つ以上のエッジを持つタンパク質の中から BLAST[5] によりクエリタンパク質と相同なタンパク質を探索し、クエリタンパク質の代わりに使用した。

符号化では、クエリタンパク質と相互作用するタンパク質または、クエリタンパク質と相同なタンパク質と相互作用するタンパク質に MAFFT[6]によるマルチプル配列アライメント（MSA）を適用することで、それらの配列をギャップを含む同じ長さの配列に変換した。アライメントされた配列において高度に保存された領域の情報を表現するために、各位置においてギャップでない残基の比率を計算し、その比率の並びにおける連続した 5 つの値（両端では連続する 3 つまたは 4 つの値）を平均することで平滑化を行った。

アライメントされた配列の符号化では、アラインメントされた配列の長さが 9000 以上の場合、配列の長さが 9000 より小さくなるまで、アラインメントされた配列から平滑化された比率が低い位置を排除した。次に、各位置のアミノ酸を one-hot ベクトルで表し、それらのベクトルを連結することで配列ごとに 2 値行列を生成した。ここで、one-hot ベクトルは、アミノ酸に対応する成分のみ 1 で、それ以外の成分は 0 であるベクトルである。また、ギャップは、すべての成分が 0 のベクトルで表現した。最後に、すべての

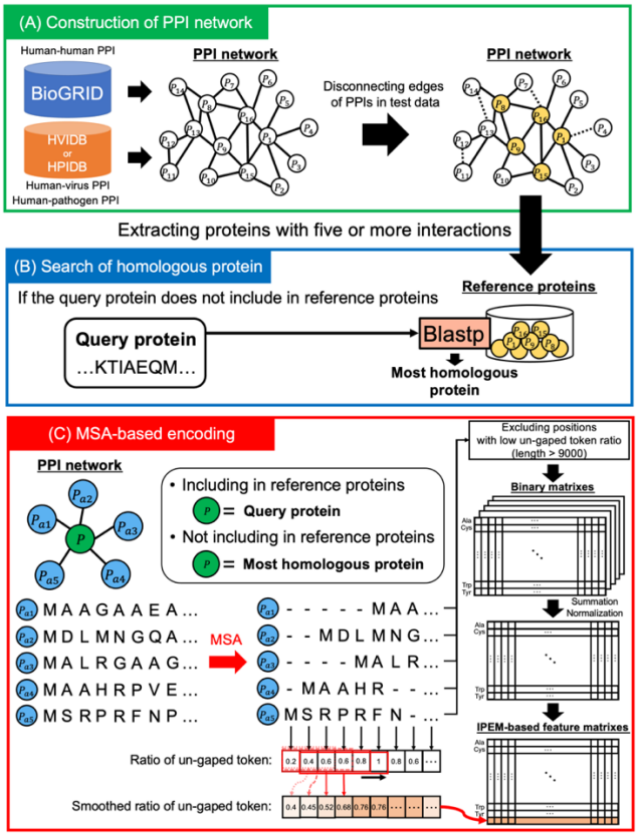

図 2 IPEM における処理フロー

配列における行列を要素ごとに合計した後、各位置で正規化した。

これらの処理により生成された行列と平滑化した比率を各位置で連結した後、最大配列サイズが 9000 となるようにゼロパディングを施すことで、9000×21 の形状を持つ特徴行列を生成した。

2.2 深層学習モデルの構築

本研究で使用された深層学習モデルのネットワークは、1 次元畳み込み層、max-pooling 層、multi-head attention layer、position-wise feed-forward network により構成される。これらのアーキテクチャの説明とそれらを組み合わせることで構築された提案方法のネットワーク構造についての説明を以下に記す。

2.2.1 1 次元畳み込み層

畳み込み層ではフィルターを用いることで、入力特徴の様々な特徴パターンを捉えることができるアーキテクチャである。長さ n、チャネル s の入力行列 X から特徴抽出を考える場合、1-D 畳み込み層では入力行列 X の特定の領域にサイズwのf枚のフィルターを適用し、ストライド t でそれらのフィルタをスライドすることで長さ$(n-w)/t+1$、特徴次元がfの行列 C に変換される。ここで、k 番目のフィルターM_kで生成された特徴ベクトル内の i 番目の位置の値は以下のように計算される。

$$C_{i,k} = \sum_{j=1}^{w}\sum_{l=1}^{s} M_{k,j,l} X_{i+j-1,l}$$

Vol.2022-BIO-69 No.19
2022/3/11

ただし、 $(1 \le i \le (n-w)/t+1, 1 \le k \le f)$

2.2.2 Max-pooling 層と Global max-pooling 層

頑健性の向上などを目的として、Pooling 層は畳み込み層と組み合わせて使用されることが多い。本研究では Max-pooling 層と Global max-pooling 層をネットワークに組み込んだ。Max-pooling 層では、畳み込み層の出力における、あるチャネルの特定の領域(カーネル内)における最大値をサンプリングし、その領域をずらしながら処理を繰り返し行うことで出力を生成する。一方、Global max-pooling 層では、畳み込み層の出力における、特定のチャネルの中で最も高い値をサンプリングすることで出力ベクトルを生成する。

2.2.3 Multi-head attention layer

Multi-head attention layer は Attention mechanism におけるプロセスを head と呼ばれる単位ごとに並列に実行するネットワークである。この Attention mechanism では、情報を与える特徴ベクトルx_iにより、入力された特徴ベクトルy_i^{pre}を更新し、出力ベクトルy_iを生成する。具体的には、以下の式のように、y_i^{pre}とx_iに異なる 3 つの重みを適用することで、Query、Key、Value と呼ばれる特徴表現を生成する。

$$q(y_i^{pre}) = y_i^{pre} W^Q$$
$$k(x_i) = x_i W^K$$
$$v(x_i) = x_i W^V$$

ここで、W^Q、W^K、W^Vはそれぞれ、Query、Key、Value を生成するための重みであり、$q(\cdot)$, $k(\cdot)$, $v(\cdot)$は Query、Key、Value を計算するための変換を表す。次に、x_jが更新後のベクトルであるy_iの生成に影響を与える度合いである attention weight $\alpha_{i,j}$を Key と Query のドット積を以下のように計算することで算出する。

$$\alpha_{i,j} = softmax\left(Mask\left(\frac{q(y_i^{pre}) k(x_j)^T}{\sqrt{d_{key}}} \right) \right)$$

ここで、d_{key}は Key の次元を表す。また、$softmax(\cdot)$と$Mask(\cdot)$は Softmax 関数と Masking 処理を表す。Masking 処理では可変長データの次元を揃えるために付与されたパディングの影響を無視するために，パディング位置の要素をマイナス無限大に設定し，ソフトマックスを適用するとゼロになるようにする。次に、Key と Query の関係に応じて Value から選択的に情報を抽出するために、以下のように attention weight を使用して Value の重み付け和を計算する。

$$Attention_i = \left(\sum_{j=1}^{n} \alpha_{i,j} v(x_j) \right)$$

ここで、$Attention_i$はy_i^{pre}を更新するための重み付き和を表す。最終的に、算出された重み付き和に重みを適用した後、更新前の特長ベクトルy_i^{pre}と足し合わせることにより出力ベクトルy_iを生成する。

2.2.4 Position-wise feed forward network (FFN)

Position-wise feed forward network (FFN)は以下のように 2 つの全結合層と ReLU 関数で構成される。

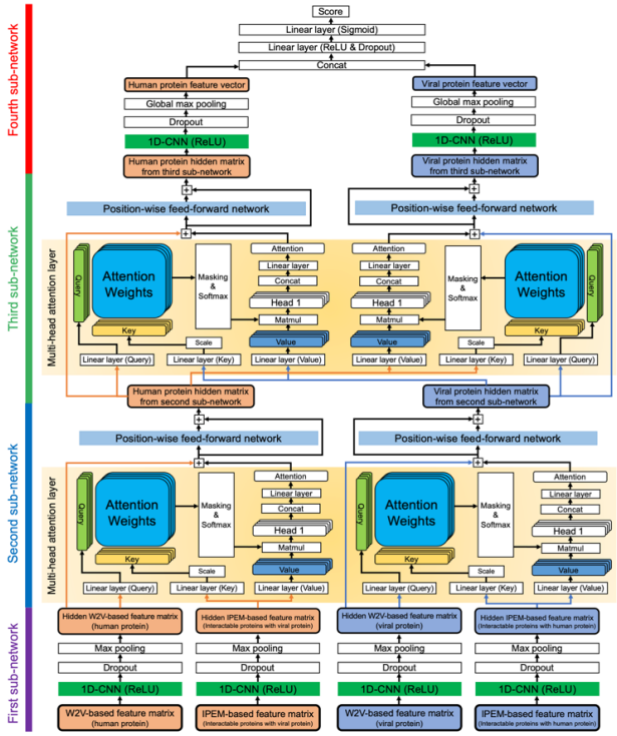

図 3 深層学習モデルのネットワーク構造

$$y = max(0, xW_1 + b_1)W_2 + b_2$$

ここで、x, W_1, b_1, W_2, b_2, yはそれぞれ、入力されたベクトル、1 層目の全結合層の重み、1 層目の全結合層のバイアス、2 層目の全結合層の重み、2 層目の全結合層のバイアス、及び出力ベクトルを示す。

2.2.5 提案手法のネットワーク構造

図 3 に示すように、本研究で開発した深層学習モデルは 4 つのサブネットワークで構成される。最初のサブネットワークでは、ヒトおよびウイルスタンパク質における word2vec に基づく特徴行列と IPEM に基づく特徴行列に 1 次元畳み込み層と max-pooling 層を適用する。先行研究の知見から[7]、1 次元の畳み込み層からの出力を 0.5 の比率で dropout した後、max-pooling 層を適用した。1 次元畳み込み層では、フィルター数、フィルターサイズ、ストライドをそれぞれ、100、20、10 に設定した。max-pooling 層ではカーネルのサイズとストライドをそれぞれ 3 と 1 に設定し、入力行列と出力行列の次元が同じになるようにゼロパディングを施した。この変換は、その後の処理において 2 つの利点を生む。1 つ目は、入力する特徴行列における複数のアミノ酸と 4-mer レベルの特徴ベクトルから、予測に重要な配列パターンを捉えることで、ペプチドレベルの特徴行列を生成することである。これにより後続のサブネットワークにおいて、モチーフやドメインなどに関する特徴間の依存関係を捉えることが可能となる。2 つ目は、入力する特徴行列における配列方向の次元数を減らすことにより、後続のサブネットワークにおける計算コストを大幅に

減少させることができることである。

2 番目と 3 番目のサブネットワークでは、multi-head attention layer により特徴抽出を行った。2 番目のサブネットワークでは、word2vec の特徴行列と IPEM の特徴行列の 1 番目のサブネットワークからの出力を multi-head attention layer に入力した。これにより、片方の予測対象のタンパク質と、もう一方のタンパク質に相互作用するタンパク質の配列情報の関係性から特徴抽出を行う。3 番目のサブネットワークでは、1 番目と 2 番目のサブネットワークから生成された 2 つのクエリタンパク質に関する特徴行列を multi-head attention layer に入力する。これにより、予測対象の 2 つのタンパク質における配列情報の関係性から特徴抽出を行う。また、multi-head attention layer 適用後に生成された特徴行列を Position-wise feed forward network に入力することで、さらなる特徴抽出を行った。Multi-head attention layer における head 数、Key、Query、Value の次元、Position-wise feed forward network における中間層の出力ベクトルの次元をそれぞれ、4、32、128 に設定した。

最後のサブネットワークでは、3 番目のサブネットワークからの出力に 1 次元の畳み込みそうと Global max-pooling 層を適用することで特徴ベクトルを生成した。さらに、それらのベクトルを連結した後、2 層の全結合層を適用することで最終的な出力を算出した。1 次元畳み込み層におけるフィルター数、フィルターのサイズ、ストライドをそれぞれ 32、5、1 に設定した。また、1 層目の全結合層における出力ベクトルの次元を 32 に設定し、そのベクトルに 0.3 の比率でドロップアウトを適用した。

2.3 モデルの学習と評価

モデルの学習ではミニバッチのサイズを 64 とし、binary cross-entropy 誤差関数により誤差を計算した。また、最適化は学習率 0.0001 として、Adam optimizer により行った。過学習を防ぐため、検証用データの Area under the ROC curve（AUC）が 20 エポック連続で更新されない場合、学習を終了した。

学習モデルの評価では精度（ACC），マシューズ相関係数（MCC），F1 スコア（F1），AUC の 4 つの指標を用いた。

3. 結果

3.1 先行研究のモデルとの比較

Eid らのデータセット[8]を使用して、過去の研究において開発された 4 つの機械学習ベースのモデル[8-11]と 2 つの深層学習ベースのモデル[12, 13]、及び本研究の提案手法の間で性能の比較を行った。モデルの構築を行うには相互作用データである陽性サンプルに加えて、陰性サンプルが必要であるが、十分な数の非相互作用が含まれているデータベースは存在しない。そのため、陰性サンプルの生成を行う必要がある。過去の研究におけるデータセットの構築データは相互作用することが確認されていないタンパク質のペアからランダムにサンプリングする方法が用いられた[14]。しかし、近年の研究[15]において、この方法では誤った陰性データが多く含まれることが指摘された。Eid らのデータセットには、Dissimilarity-based negative sampling 法と呼ばれる方法により配列情報に基づいて生成された信頼性の高い陰性サンプルが含まれている。このような信頼性の高い陰性データが含まれている Eid らのデータセットは過去の多くの HV-PPI 予測の研究においてベンチマークデータセットとして使用された。このデータセットには 5020 の陽性サンプルと 4734 の陰性サンプルで構成されるトレーニングデータと 425 の陽性サンプルと 425 の陰性サンプルで構成されるテストデータが含まれる。我々は、9000 残基以上の長さを持つタンパク質で構成されるサンプルをトレーニングデータから除外した。その後、トレーニングデータを使用して 5 分割交差検証を行うことでモデルの学習を行い、テストデータにより学習されたモデルの評価を行った。表 1 に示すように、全ての指標において本研究の提案手法は過去の手法より高い性能を示した。提案手法が、このような高い性能で予測を行うことができたのは Attention ベースの深層学習モデルによって、複数の配列間の依存関係を捉えることができたためであると考える。過去の研究における深層学習ベースのモデルでは、配列情報から特徴を抽出するためのネットワークがヒトタンパク質とウイルスタンパク質で独立しており、タンパク質特異的に情報の抽出が行われていた。我々の提案手法では深層学習ネットワーク内に multi-head attention layer を組み込むことで、複数の配列間の関係性を考慮することができ、相互作用を予測するために重要な情報を選択的に抽出することが可能となったのではないかと考える。

Model name	Author (Year)	ACC	MCC	AUC	F1
Denovo (SVM)	Eid et al.[8] (2015)	0.819	NA	NA	NA
SVM	Alguwaizani et al.[9] (2018)	0.865	0.729	0.926	NA
SVM	Zhou et al.[10] (2018)	0.845	0.692	0.897	NA
RF+Doc2vec	Yang et al.[11] (2020)	0.932	0.866	0.981	0.931
DeepViral (Sequence)	Liu-Wei et al. [12] (2021)	0.931	0.865	0.960	0.929
DeepViral (Joint)		0.939	0.881	0.976	0.937
CNN	Yang et al. [13] (2021)	0.941	NA	NA	0.939
Our proposal	-	**0.954**	**0.909**	**0.987**	**0.953**

表 1 本研究の提案手法と過去の手法の性能比較
（先行研究の性能は各論文を参照した）

3.2 未知のウイルスに対する予測

提案手法が未知のウイルスや変異株における HV-PPI を高い性能で予測することができるかどうかを検証するために、

Zhou らの研究[10]において構築されたデータセットを使用した。このデータセットは以下に示すような 4 つのトレーニングデータと 2 つのテストデータにより構成される。

（トレーニングデータ）

- **TR1**: ヒトと H1N1 以外のウイルス間のサンプル
- **TR2**: ヒトとエボラウイルス以外のウイルス間のサンプル
- **TR3**: 何らかの宿主と H1N1 以外のウイルス間のサンプル
- **TR4**: 何らかの宿主とエボラウイルス以外のウイルス間のサンプル

（テストデータ）

- **TS1**: ヒトと H1N1 間のサンプル
- **TS2**: ヒトとエボラウイルス間のサンプル

Zhou らはテストデータにおけるウイルス種がトレーニングデータに含まれないようにトレーニングデータとテストデータを組み合わせることによって 4 つのデータセットを構築した。具体的には、Dataset 1、2、3、4 には、それぞれ TR1 と TS1、TR2 と TS2、TR3 と TS1、TR4 と TS2 が含まれる。モデルの構築ではトレーニングデータの 20%を学習回数を決定するためのデータとして使用し、残りのデータでモデルの最適化を行った。また、学習されたモデルをテストデータによって評価した。テストデータにおけるウイルス種を未知のウイルスとするために、IPEM のネットワーク構築後に H1N1 またはエボラウイルスのタンパク質に対応するノードをネットワークから除外した。我々はこのデータセットを用いて、Zhou らのモデルに加えて、以前の研究で我々が開発した LSTM-PHV[16]と予測性能の比較を行った。提案手法は、いずれのデータセットにおいても 0.94 以上の AUC を示し、過去の研究のモデルより高い性能で予測を行った(図 4)。このことから本研究の提案手法は高い汎化性能を持ち、未知のウイルス種の PPI の予測においても有用であることが示唆された。

3.3 SARS-CoV-2 における PPI の予測

我々は SARS-CoV-2 の HV-PPI における提案手法の予測性能を評価するため、Yang らの研究[13]において構築された SARS-CoV-2 の PPI を含むデータセットを使用した。Eid らの研究において構築されたデータセットと同様、このデータセットの陰性サンプルは Dissimilarity-based negative sampling 法によって生成された。このデータセットには、568 の陽性サンプルと 5680 の陰性サンプルが含まれている。本研究では Yang らと同様に 5 分割交差検証によりモデルのトレーニングと評価を行った。モデルの構築ではトレーニングデータの 10%を学習回数を決定するためのデータとして使用し、残りのデータでモデルの最適化を行った。このデータセットは陽性サンプルと陰性サンプル数が大きく異る不均衡データであるため、精度に加えて Area under the precision-recall curve (AUPRC)により評価を行った。図 5

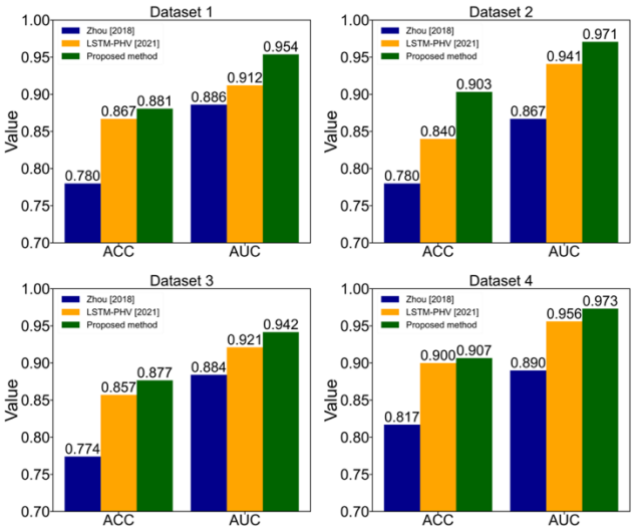

図 4 未知のウイルスにおける予測
（先行研究の性能は各論文を参照した）

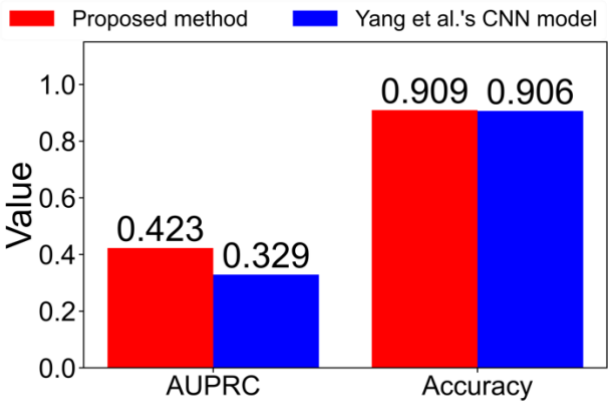

図 5 SARS-CoV-2 の PPI の予測における精度と AUPRC
（先行研究の性能は各論文を参照した）

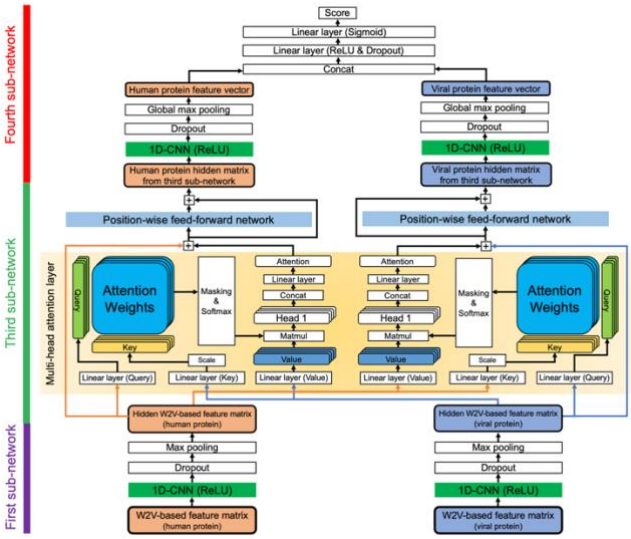

図 6 IPEM の情報を使用しないモデルのネットワーク構造

に示すように、精度は Yang らの CNN ベースの手法と大きく違いがないのに対し、AUPRC の値は提案手法が大きく上回った。このことから Yang らのモデルに比べて、提案手法は不均衡データをうまく学習することができ、多くの陰性サンプルが含まれるサンプルの中から陽性サンプルをよ

り高い性能で特定しうる。

3.4 IPEM の有用性の検証

我々は IPEM の有用性を検証するために、IPEM からの情報を使用しない深層学習モデルの構築を行った。具体的には、図 6 に示すように、IPEM の特徴行列から特徴抽出を行うための 1 番目のサブネットワークの畳み込み層と max-pooling 層、及び 2 番目のサブネットワークを元々のネットワークから取り除いた。その後、Eid らのベンチマークデータセットと Yang らの SARS-CoV-2 のデータセットを用いてモデルの学習と評価を上記の章で記した方法と同じ方法で行った。図 7 に示すように、Eid らのデータセットでは 2 つのネットワーク構造による大きな性能の違いはなかった。一方、Yang らのデータセットでは、IPEM ベースの特徴行列を使用しなかった場合に比べて、使用した場合の予測では高い性能が示された(図 8)。また、5 分割交差検証により学習された 5 つのモデルの性能指標に対して、対応のある片側 t 検定を行ったところ、IPEM を使用した場合の性能指標は使用しなかった場合に比べて有意に高かった。Yang らのデータセットは Eid らのデータセットに比較して非常に少ない陽性サンプルにより構成されている。このことから、我々は IPEM に基づく特徴行列は少ない相互作用データでの HV-PPI 予測モデルの構築において有用であることを提案する。

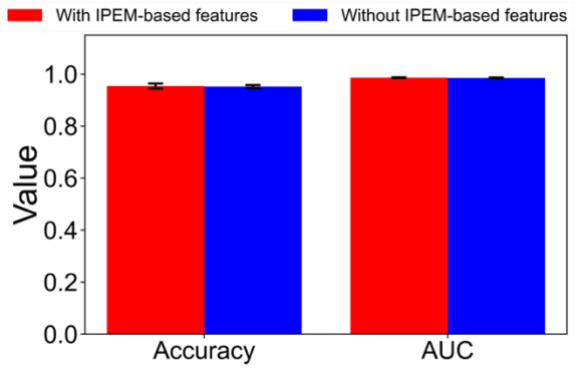

図 7 Eid らのデータセットにおける IPEM ベースの特徴行列を使用した場合(赤)と使用しなかった場合(青)の性能比較

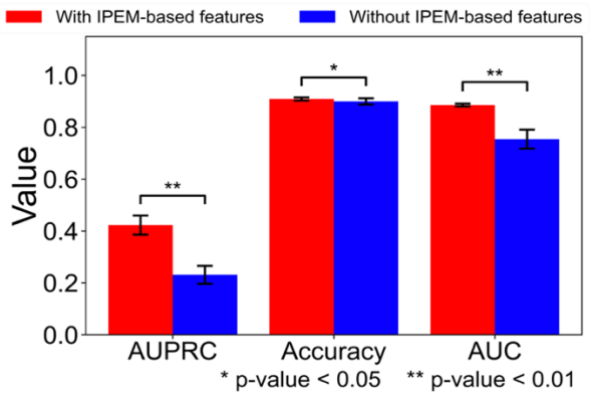

図 8 Yang らのデータセットにおける IPEM ベースの特徴行列を使用した場合(赤)と使用しなかった場合(青)の性能比較

4. おわりに

本研究では IPEM と呼ばれる新規エンコーディング手法を提案し、Attention ベースの深層学習モデルにより、HV-PPI の予測に取り組んだ。その結果、提案手法のモデルは過去の研究におけるモデルに比べて、高い予測性能を示した。このような高い性能で予測を行うことができたのは、複数の配列の依存関係を捉えることができたためであると考える。また、IPEM を使用した場合と使用しなかった場合の比較において、相互作用データ数の少ない場合に、IPEM を使用した予測は有効であることが示された。

参考文献

[1] World Health Organization et al. Coronavirus disease (covid-19) situation dashboard. https://covid19.who.int/ (December 29 2021, date last accessed).
[2] The UniProt Consortium. UniProt: the universal protein knowledgebase, Nucleic Acids Res 2017;45:D158-D169.
[3] Oughtred R, Rust J, Chang C et al. The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions, Protein science : a publication of the Protein Society 2021;30:187-200.
[4] Yang X, Lian X, Fu C et al. HVIDB: a comprehensive database for human-virus protein-protein interactions, Brief Bioinform 2021;22:832-844.
[5] Altschul SF, Gish W, Miller W et al. Basic local alignment search tool, J Mol Biol 1990;215:403-410.
[6] Katoh K, Misawa K, Kuma K et al. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform, Nucleic Acids Res 2002;30:3059-3066.
[7] Wu H, Gu X. Max-Pooling Dropout for Regularization of Convolutional Neural Networks. 2015, arXiv:1512.01400.
[8] Eid FE, ElHefnawi M, Heath LS. DeNovo: virus-host sequence-based protein-protein interaction prediction, Bioinformatics 2016;32:1144-1150.
[9] Alguwaizani S, Park B, Zhou X et al. Predicting Interactions between Virus and Host Proteins Using Repeat Patterns and Composition of Amino Acids, Journal of Healthcare Engineering 2018;2018:1391265.
[10] Zhou X, Park B, Choi D et al. A generalized approach to predicting protein-protein interactions between virus and host, BMC Genomics 2018;19:568.
[11] Yang X, Yang S, Li Q et al. Prediction of human-virus protein-protein interactions through a sequence embedding-based machine learning method, Comput Struct Biotechnol J 2020;18:153-161.
[12] Liu-Wei W, Kafkas Ş, Chen J et al. DeepViral: prediction of novel virus–host interactions from protein sequences and infectious disease phenotypes, Bioinformatics 2021;37:2722-2729.
[13] Yang X, Yang S, Lian X et al. Transfer learning via multi-scale convolutional neural layers for human–virus protein–protein interaction prediction, Bioinformatics 2021;37:4771-4778.
[14] Barman RK, Saha S, Das S. Prediction of interactions between viral and host proteins using supervised machine learning methods, PLoS One 2014;9:e112034.
[15] Dey L, Chakraborty S, Mukhopadhyay A. Machine learning techniques for sequence-based prediction of viral-host interactions between SARS-CoV-2 and human proteins, Biomed J 2020;43:438-450.
[16] Tsukiyama S, Hasan MM, Fujii S et al. LSTM-PHV: prediction of human-virus protein–protein interactions by LSTM with word2vec, Briefings in Bioinformatics 2021;22:bbab228.