アテンションネットワークによるT細胞受容体とペプチド結合の予測・その機能の解釈性に関する研究

小山 恭平
大阪大学 大学院生命機能研究科 計算生物学研究室

1. はじめに

蛋白質は、遺伝子をもとに体内で生成された1次元のアミノ酸残基の配列が3次元に折りたたまれた立体構造であり、多様な種類の蛋白質が生体内で機能している。蛋白質の機能とは、蛋白質間の相互作用や、低分子との相互作用によって発揮される細胞内での各種の役割のことである。1次元のアミノ酸配列が蛋白質の構造や機能を規定している。

そこで本研究では、機械学習手法、特にアテンションレイヤーを用いたニューラルネットワーク (NN) の解釈性に関する手法を取り入れることで、蛋白質の1次構造からのリガンド結合予測の結果を解釈し、特定の機能に寄与している部分構造を発見することを目的とする。蛋白質のリガンド結合が予測できることで、生体機能の理解や疾患メカニズムの理解が進み、薬剤の開発などの応用が期待される。

2. T 細胞受容体とペプチドの結合予測

2.1 T細胞受容体とペプチド

T-Cell Receptor (TCR) は抗原受容体であり、多くはα(TCRα) とβ(TCRβ) によって構成されている蛋白質である。TCR は抗体 (BCR) と同様に Complementarity determining region (CDR) 領域における配列の多様性が知られている。TCR の CDR は主要組織適合性遺伝子複合 (Major Histocompatibility Complex) 上のペプチドと結合し反応することで、生体内での免疫反応が進む。TCR のリガンドである。これらに対する免疫反応は、すなわち疾患の治療に繋がる。

2.2 結合予測タスク

本研究では、TCR の CDR とペプチド間の結合予測を扱う。TCR とペプチドの結合予測は、最終的に結合の有無である確率を出力する二値の分類問題である。予測モデルのアウトプットとして「結合する」と「結合しない」をラベルとして予測する。インプットとしては、CDR 配列とペプチドの2つの配列を入力値として取る。

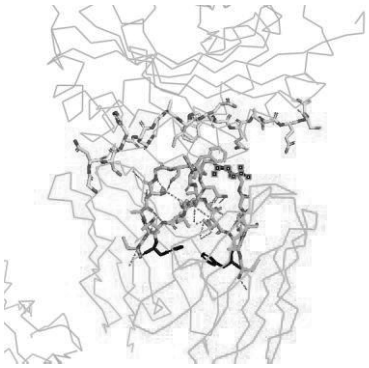

図1：Protein Data Bank (PDB) に登録されたペプチドと TCR が結合した複合体の例 (PDB ID:1BNA(1))

2.3 課題とチャレンジ ・新規性

1次元のアミノ酸配列は、シークエンシング技術の進歩により取得は容易になってきているが、3次元構造はアミノ酸配列に比べて取得にコストと時間がかかる。アミノ酸配列は文字列で表現されるため、近年の機械学習分野における自然言語処理の進展が、バイオインフォマティクス分野、1次元のアミノ酸配列への情報処理に大きく貢献している。

一方、T -Cell は抗体と同様に免疫細胞であり、その CDR は外部の物質に反応しなければならない。そのため、CDR・TCR の配列パターンは膨大で、最大で 10 の 61 乗になるとも言われており、どのペプチド配列がどの CDR に結合するのかを予測することは難しい。また、予測が難しいタスクであるが故に、ニューラルネットワークモデルの解釈可能性は必要不可欠である。

しかし、機械学習における予測結果は一般的に解釈が難しい。また、蛋白質の機能に寄与しているアミノ酸残基情報はコンボリューショナル・ニューラルネットワーク (CNN) によって抽出されてきたが、CNN は短い距離の情報に注目してしまうため、共起的な複数部分構造を加味した予測や、遠く離れた残基間におけるインタラクションが機能予測に反映されず、解釈の困難さと予測精度の向上にも限界がある。1次構造からの蛋白質・蛋白質間の相互作用に関する予測についても、同様の理由で、予測結果

の解釈は自明ではなかった。

そこで本研究では、TCR とペプチドの結合予測に、アテンションレイヤーを用いたモデルを応用し、アテンションレイヤーを可視化することで、予測モデルを通して結合原理の理解と解釈を試みる。

2.4 関連研究

トランスフォーマーモデル (2) が提案されてから、アテンションレイヤーは自然言語処理を代表的なタスクとして多くのモデルに応用されてきた。蛋白質への自然言語処理モデルの応用は、蛋白質の構造予測や機能予測、及び相互作用予測などが存在している (3)(4)。アテンションレイヤーは近年の自然言語処理において最も強力な NN のひとつであり、文字列や文章内全体の情報に基づいて NN の重みを変更することで、意味のある情報を文字列から抽出することができる。TCR 結合予測の 既存研究は、Springer.et.al による ERGO-II(5) などがある。

ERGO-II も TCR と Peptide の結合を予測しているが、モデルの解釈性には踏み込んで はいないと言える。

3. 手法概要

3.1 Attention Layer モデル

アテンションレイヤーに変更を加えた、クロスアテンションレイヤーは、特に 2 つのアミノ酸配列を入力値として取り、一方の配列全体の情報に基づいて他方の配列全体の情報から意味のある情報を抽出することを可能にしており、相互作用予測用の NN である。特に、結合予測に用いられたクロスアテンションレイヤーに解釈性を加えることで、配列を跨いだ相互作用の理由を統計的に解釈することが可能になった。

クロスの場合のアテンションレイヤーは、下記式で規定される。 Q, K 及び V は、各配列データを行列化したものである。配列長 L、隠れ層次元 E として、L x E サイズで各配列を行列で表現する。Q=K=V の場合に、L_1=L_2 となり、セルフアテンションレイヤーとなる。d はスケーリングファクターである。 クロスにする場合は、K=V であり、Q は異なるデータを入力する。この時、*Softmax* 関数が、行列 Q が入力されたときの V への重みを定義し、和が 1 になるように重みを配分している。この $Softmax\ \frac{QK^T}{d}$ がアテンションであり、可視化に用いた。

$$Attention(Q,K,V) = Softmax\left(\frac{QK^T}{d}\right)V,$$

$$w\ ere\ Q\ L\quad E\ K\ L\quad E\ V\ L\quad E$$

3.2 言語・配列としての CDR ・ペプチド

データの表現に関して、CDR 配列とペプチドは、20 種類のアミノ酸配列で表現される。アテンションレイヤーは、CDR とペプチドのパターンを学習し、結合を予測できる。本研究では、予測モデルへの入力値として、配列のみを利用し、遺伝子座の情報や MHC の情報といった副次的な情報は利用しない。

3.3 ニューラルネットワークモデルの詳細

本研究で用いた予測モデルの概要を図 2 に示す。ペプチドと TCR α・β の配列をそれぞれ別々にエンベディングレイヤーとトランスフォーマーで 処理していき、クロスアテンションレイヤーにて相互にトランスフォーマーモデルに入力し、最終的に 出力層にて平均をとり、連結し、マルチレイヤーパーセプトロン (MLP) レイヤーにて 1 つの予測値を結合確立として アウトプットする。ロス関数はクロスエントロピーを用いた。

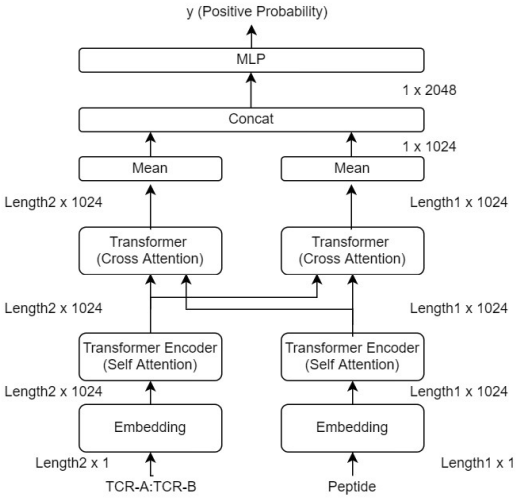

図 2：本研究で用いたモデルの概要図

図 2 における相互作用のクロスアテンションレイヤーを可視化することで、配列を跨いだ相互作用の理由を解釈することが可能になる。クロスアテンションレイヤーはトランスフォーマーモデルを改変したものであり、ペプチドをインプットにして TCR の各残基に対して比重を割り振る形で、TCR の重要な部位を学習する。

4. 実験

配列データのトレーニングのスコアをベンチマークの既存研究と比較することで手法の妥当性を確認した。その上でアテンションレイヤーから得られる重要部位を立体構造に対応させ、構造情報とアテンションの関係性を見出した 。

4.1 配列データと結合データ

前章で述べた通り、データセットは TCR とペプチドのペアに対してバイナリのラベルが付与されているものを既存研究から引用し、ベンチマークデータセットとして既存研究のモデルと比較した。さらには、汎用性の高いアテンションレイヤーの獲得を目指し、データセットをオープンデータレポジトリの VDJDB から取得し、データを拡張したモデルでトレーニングした。McPAS, VDJdb_without10x データセット (5)(6)(7) は既存研究で用いられているデータで、ベンチマークデータセットである。一方で、VDJdb_with10x データをアテンションレイヤートレーニング用として用いた。

バイナリ分類問題であり、モデルのトレーニングのためには、ネガティブラベルのデータが必要であるが、TCR とペプチドの反応データはポジティブラベルのデータが多いため、本研究では、既存研究に倣ってランダムな TCR とペプチドのペアの組み合わせを生成し、ネガティブラベルを付与することで、正負の比率を調節した。

全データを統合すると、インタラクションのデータ数は 199222 件あり、ユニークな TCRα と TCRβ のデータ数はそれぞれ 17954 件と 19162 件であった。また、ペプチドのユニークな個数は 478 件であった。ネガティブとポジティブのラベル数はそれぞれ 166096 件と 33126 件であった。

4.2 蛋白質の立体（三次元）構造データ

蛋白質の立体構造データは、Protein Data Bank (PDB, https://www.wwpdb.org) から取得した。TCR とペプチドに関与するデータは 49 件あり、そのうち、TCR- ペプチドペアの重複を除くと 38 件が残った。これら 38 件についてアテンションレイヤーの可視化を行った。

5. 結果

5.1 結果

ベンチマークデータセットに対する既存研究モデルと本研究のモデルの比較を表 1 に示す。本研究のスコアは、配列のみからの結合予測では両方のベンチマークデータについて既存研究を上回った。しかし、既存研究のベストスコアの設定である配列 V,J 遺伝子 MHC 種類を用いた場合の McPAS データのスコアには及ばなかった。

モデル名	データセット名	データ種類	ROCAUC スコア
本研究モデル	McPAS	配列のみ	0.933
本研究モデル	VDJdb_ without10x	配列のみ	0.925
ERGO-II, LSTM	McPAS	配列のみ	0.855
ERGO-II, LSTM	VDJdb_ without10x	配列のみ	0.800
ERGO-II, LSTM	McPAS	配列＋ V,J 遺伝子 MHC 種類	**0.939**
ERGO-II, LSTM	VDJdb_ without10x	配列＋ V,J 遺伝子 MHC 種類	0.849

表 1： ベンチマークデータセットに対する スコア

5.2 アテンション解釈

立体構造での距離行列と、各ヘッドにおけるアテンションレイヤーの可視化例を図 3、4 に示す。アテンションレイヤーの値が大きい TCR 残基において、立体構造内で相互作用があることが確認できた 。

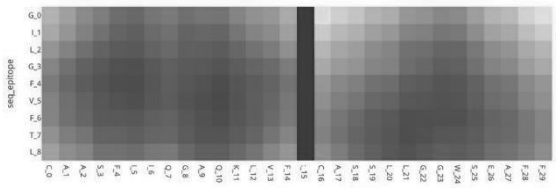

図 3：立体構造での距離行列。横軸に CDR 配列、縦軸にペプチド配列として距離を可視化した。

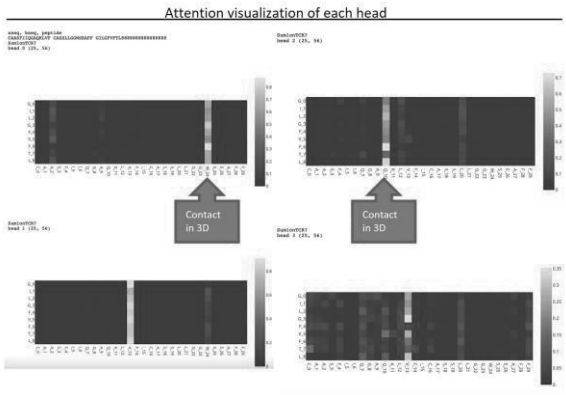

図 4：各ヘッドにおけるアテンションレイヤーの可視化。特に上段 2 つのヘッドについて、アテンションレイヤーの値が大きい TCR 残基において立体構造内で結合があることが確認できた。上の例では、ペプチド配列が入力されたときに、各 TCR 配列が重みを割り振られている。よって、x 軸方向に和をとると 1 になる。

6. 考察・まとめ

本研究では、アテンションレイヤーで、蛋白質の１次構造からの結合機能の予測を解釈した。実験結果から、残基のアテンションが大きいことが立体構造内の相互作用を通して蛋白質機能において重要な役割を果たすことに対応していることが示唆された。

アテンションは配列情報のみから手に入れられたものであるため、タンパクの立体構造を入手せずとも、TCR と CDR の結合にかかわる残基を判別できる可能性がある。

参考文献

(1) CRYSTAL STRUCTURE OF COMPLEX BETWEEN D10 TCR AND PMHC I-AK/CA, (1999)

(2) Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).

(3) Jumper, John, et al. "Highly accurate protein structure prediction with AlphaFold." Nature 596.7873 (2021): 583-589.

(4) Jurtz, Vanessa Isabell, et al. "An introduction to deep learning on biological sequence data: examples and solutions." Bioinformatics 33.22 (2017): 3685-3690.

(5) Springer, I., Tickotsky, N. and Louzoun, Y., 2021. Contribution of t cell receptor alpha and beta cdr3, mhc typing, v and j genes to peptide binding prediction. Frontiers in immunology, 12.

(6) Tickotsky N, Sagiv T, Prilusky J, Shifrut E, Friedman N (2017). McPAS-TCR: A manually-curated catalogue of pathology-associated T cell receptor sequences. Bioinformatics 33:2924-2929

(7) Bagaev, Dmitry V., et al. "VDJdb in 2019: database extension, new analysis infrastructure and a T-cell receptor motif compendium." Nucleic Acids Research 48.D1 (2020): D1057-D1062.