Vol.2021-MPS-136 No.11
2021/12/13

Transformer及び既存BERTモデルを用いたRNA-蛋白の結合予測

木村 高幸[1,a)]　安尾 信明[2]　関嶋 政和[1]

概要：RNA 蛋白質の結合は、生体内で重要な鍵となる相互作用であるが、実験でのコスト等からコンピュータによる予測が求められている。既に多くのモデルが報告されているが、精度や評価の点で改善の余地がある。自然言語処理の分野等で Transformer が大きな成果を出しているが、RNA 蛋白質の結合予測にはまだ応用されていない。本研究では、この Transformer を利用して、RNA 蛋白質の結合予測モデルの構築を行う。

RNA-protein Binding Prediction with Transformer and the Two Existing BERT Models

Abstract: Bindings between RNA and protein are essential interactions in organisms. However, experimental approaches are relatively expensive. This is why inexpensive computational approaches are expected. A lot of models were already reported, but there is room for improvement in terms of AUROC and the evaluation methods. By the way, attention-based architecture called Transformer has been successful in many fields such as the natural language processing field, but the Transformer was not directly used to this RNA-protein binding prediction problems. In this study, we build a Transformer model to solve RNA-protein binding prediction problem especially for proteins that have no binding data to RNA.

1. 序論

1.1 RNA-蛋白質の結合予測

蛋白質と RNA は、セントラルドグマに関わる 3 つの要素（DNA、RNA、蛋白質）の内の 2 つであり、生体内での基本的な反応に関わる重要な分子である [1]。RNA の情報から蛋白質が生成される過程（翻訳）において、RNA にある種の蛋白質が結合して、この翻訳の過程を促進したり阻害することがわかっている。この制御の乱れが遺伝性の病気の原因にもなりうる [2]。従って、RNA のどの部分にどの蛋白が結合するのかという情報は生体内のこうした制御機構の理解や、それに伴う病気の原因究明や治療に重要であると考えられる。RNA に結合する蛋白質は少なくとも 1500 種類報告されている [3]。RNA と蛋白質の結合実験データを集めた公的なデータベースは既に存在するが、蛋白質の種類は 400 程度に留まる [4]。データベースで対象の蛋白あるいは RNA の結合情報が見つからない、または充分な情報がない場合でも、実験を行って結合情報を得ることもできる。実験でのアプローチには CLIP などの様々な手法があり [5]、効率などの点で改善もされてきている [6] が、時間や費用などのコストが掛かる。そこでコンピュータによる予測が求められている。コンピュータによる結合予測（2 値分類）については、既に様々なモデルが提案されている。大きく分けて、3 次元構造を用いるモデル [7], [8], [9] と用いないモデル [10], [11], [12], [13], [14], [15], [16], [17] がある。入手可能な蛋白 RNA 複合体 3 次元構造データの量はそれぞれの配列の量に比べて非常に少ない。例えば、Protein Data Bank(PDB) [18] において、RNA と蛋白質を含むエントリーの数は、2021 年 11 月の時点で約 4000 個である。ENCORE と呼ばれるデータベース [16] は約 150 個の RNA 結合蛋白質についての RNA 配列への結合情報を含むが、1 つの蛋白質について少なくとも 60,000 個の正例と同数の負例が存在する [19]。したがって、配列のみを使った結合予測のほうが有用性が高いと考えられる。配列のみを使うモデルについては、蛋白質と RNA 両方の配列を用いるもの [10], [11], [12], [13], [14], [15] と、RNA 配列のみ

[1] 東京工業大学 情報理工学院
[2] 東京工業大学 物質・情報卓越教育院
[a)] kimura.t.bf@m.titech.ac.jp

を入力とし特定の蛋白について最適化したモデル [16], [17] がある。RNA 配列のみを用いるモデルは、その対象蛋白についての RNA への結合データが一定量存在することが前提になる。したがって、上に述べた様に、データベースで結合データが見つからない場合や、実験を行なわない場合、RNA 配列のみを用いるモデルは使えない。そこで本研究では、RNA と蛋白質を入力として用い、既存データの存在を前提とせず、結合の有無を予測するモデルを考える。

1.2 既存の予測モデルと本研究での提案

蛋白質と RNA の配列のみを入力とし、結合の有無を予測する既存モデルで使われている手法は、Convolutional Neural Network [11]、サポートベクターマシンやランダムフォレスト [10]、ナイーブベイズ [15]、Broad Learning System[13]、Feature Selection Ensemble[12]、Stacked Autoencoder[14] など多岐に渡る。RPI369 など主に使われるベンチマーク [10] における既存モデルの AUROC は 0.95 以上に達しており、高精度のモデルが報告されている。しかし、高品質の負例や鎖の長さでフィルタリングを行い最近提案された新しいベンチマーク RPI1446[11] に対しては、報告されている結果で約 0.90 にとどまっており改善の余地が残る。また、RPI369 などよく使われるベンチマークではデータサイズが充分とは言えず、モデルの比較には使えても精度の評価には不十分である可能性がある。本研究では、自然言語処理の分野で大きな貢献を示し、分子プロパティ予測などの分野でも応用され始めている Transformer[20] を用いて精度のさらなる改善を試みる。Transformer を使った RNA 蛋白の結合予測は既に存在するが（BERT-RBP[21]）、入力データは RNA の配列のみであり、結合データのない蛋白質に対する評価は行っていないため本研究とは異なる。RPI369 などのベンチマークにおいては、蛋白質や RNA の重複が見られるため、クロスバリデーションを行う際に、結合データを持たない蛋白質の評価にはなっていない。そこで、本研究では、既存モデルとの比較だけでなく、結合データを持たない蛋白質における RNA への高精度な結合予測モデルの実現を目指す。今まで Transformer などアテンションベースのみのアーキテクチャーが使われてこなかった理由は 2 つ考えられる。まず、Transformer が比較的新しいこと、そして、計算リソースの問題である。RNA 蛋白の結合予測においては、例えば蛋白質の長さ（残基の数）が、文章の長さに相当するが、RNA 結合蛋白質には蛋白の長さが 4000 近くになるものも存在する。これは文章の長さで 4000 語に相当する。自然言語では存在しない非常に長い文章を扱うことになるためメモリ上限の問題と訓練速度低下の問題が発生する。メモリ使用量を抑えるために、層の数を減らすなど小さなモデルにすること、また少ないエポック数で最適化できる工夫が求められる。そこで本研究では、蛋白質と RNA それぞれの BERT モデル [22] を pretrain することと、当研究グループで作成した RNA 蛋白質間の統計ポテンシャル [23] によるアテンション強化 [24] を行うことで、メモリや訓練時間などの問題を克服し、高精度な予測モデルの構築を目指す。

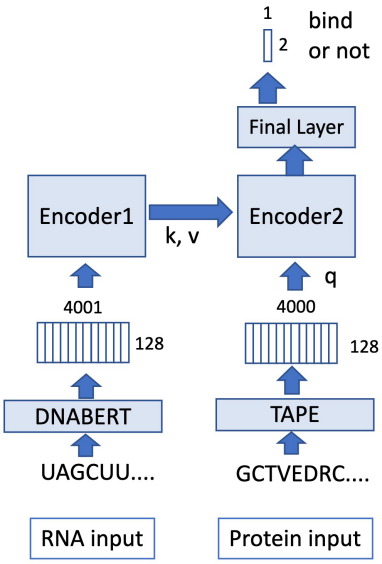

図 1 モデルの全体図
Fig. 1 The Overview of the Model

2. 手法

2.1 モデル全体の構成

本モデルの全体図を**図 1** に示す。RNA の配列を DNABERT に通し、その出力を RNA の表現ベクトルとして用いる。同様に蛋白質の配列を入力として、TAPE の出力を蛋白質の表現ベクトルとして用いる。DNABERT[25]、TAPE[26] はともに BERT を使った既存モデルである。TAPE については pretrain されたモデルをダウンロードして用い、DNABERT については、1-mer で訓練されたモデルがなかったため、本研究で使用するベンチマーク全てを用いて独自に pretrain を行った。つまり DNABERT、TAPE はともに pretrain されたものを用い、本研究での最適化（fine tuning）対象とはしない。DNABERT の出力である RNA の表現ベクトルは、Encoder1 に入り、アテンションなどの計算を行った後に key 及び value として、Encoder2 に入る。Encoder2 では、蛋白のセルフアテンションと蛋白と RNA のクロスアテンションの計算が主に行われる。その後 final layer を通り、2 次元のベクトルが最終出力となる。Encoder2 の詳細は**図 2** を参照のこと。

2.2 アテンション

i 番目のヘッドへの入力を H_i とすると、まず Q_i、K_i、V_i が以下の式で計算される。

$$Q_i = W_i^Q H_i, K_i = W_i^K H_i, V_i = W_i^V H_i \quad (1)$$

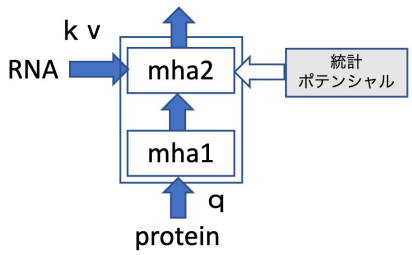

図 2 Encoder2 の構成（1 層分）
Fig. 2 The Structure of Encoder2 (1 layer)

アテンション A は次の式で計算される [20]。

$$A^{(i)} = softmax(\frac{Q_i K_i^T}{\sqrt{d_k}})V_i \quad (2)$$

d_k は key の配列の長さである。

2.3 クロスアテンション

RNA の表現、蛋白質の表現ベクトルは、それぞれセルフアテンションブロックを経るが、最終的にはクロスアテンションにより合流する（図 2）。その際、RNA 側を key 及び query とし、蛋白質側を value とした。RNA と蛋白質の役割を入れ替えたネットワークも同時に使用するいわゆる co-attention network [27] もあるが、今回はメモリの使用量を抑えるため co-attention network ではない形で進める。アテンション強化は、Maziarka らのモデル [24] を参考に以下のように計算する。

$$A^{(i)} = (\lambda_{pi} S_{pi} + \lambda_{hb} S_{hb} + \lambda_a softmax(\frac{Q_i K_i^T}{\sqrt{d_k}}))V_i \quad (3)$$

で計算される。S_{pi}、S_{hb} はそれぞれ、配列のペアごとに予め作成した π相互作用、水素結合の統計ポテンシャルの表（4000×4001）、λ_{pi}、λ_{hb}、λ_{a} は、π相互作用、水素結合、アテンションの重みを示すスカラー値である。アテンション強化に用いる統計ポテンシャルは、当研究グループで計算した水素結合とπ相互作用の値の 2 種類の表であり、当グループで最適化した残基塩基間のスカラー値である。水素結合については 80 個、π相互作用は 36 個の数値からなる。mha1 は主に蛋白表現のセルフアテンション計算を行うブロックで、mha2 は、主に RNA と蛋白表現のクロスアテンション計算を行う。図 2 は 1 層での構造であり、複数の層からなる場合はこの構造が繰り返される。また、図 2 は、layer normalization 層や、feed forward network 層を便宜上省略した簡略図であることに注意されたい。

2.4 TAPE の Pretrain

TAPE については、 pretrain 済のものをダウンロード (https://github.com/songlab-cal/tape) した。TAPE 及び DNABERT については 12 層と 12 個のアテンションヘッドで構成される BERT-base モデルと呼ばれる構成のモデルを使用した。TAPE は、Pfam[28] 内の約 3100 万個の蛋白質ドメインの配列を使って pretrain されている。ここでの pretrain とは、一部の配列を隠して残りの配列から推測することでラベル無しでモデルを最適化する Masked Language Model アプローチである。

2.5 DNABERT の Pretrain

DNABERT は、RNA の配列データを集めた上で pretrain を行った。統計ポテンシャルによるアテンション強化のために 1-mer モデルを作成する必要があると考えたためである。k-mer とは、連続する k 個の塩基を 1 語とするモデルである。本研究で用いる統計ポテンシャルは、残基塩基間の数値なので、1-mer を使用した。訓練データは NPInter[29]、RPI369、RPI488、RPI1807、RPI2241 [10] の 7570 個の RNA 配列を用いた。

2.6 Fine Tuning

NPInter、RPI369、RPI488、RPI1807、RPI2241 の各ベンチマークについて、5 フォールドのクロスバリデーションを行う。TAPE の pretrain 以外の全ての計算は東京工業大学のスーパーコンピューター TSUBAME3.0 を利用した。TAPE 及び DNABERT からの出力は固定し、その後のネットワークのパラメータのみを最適化対象とした（図 1）。また、TAPE 及び DNABERT からの出力は基本的なベクトルが 768 次元だが、メモリ節約のため 128 次元に落とした。損失関数は negative log likelihood、最適化アルゴリズムは Adam を使った。アテンション強化に用いる統計ポテンシャルは、負の値になるほど安定であるから、− 1 を掛けてから softmax に通した。バッチサイズが 1 より大きくなるとメモリエラーになったため、gradient accumulation を使いパラメータ更新の頻度を下げることで精度の向上に努めた。また、メモリ使用量を抑えるため、各ブロック（Encoder1 は RNA のセルフアテンションで 1 ブロック、Encoder2 は蛋白質のセルフアテンションとクロスアテンションの合計 2 ブロック）での層は 3 層、アテンションヘッドの数は 4 に設定した。

謝辞 本研究を進めるに当たり、早稲田大学の浜田道昭教授と山田啓介氏から貴重なアドバイスを頂いた。お礼を申し上げたい。

参考文献

[1] Hentze, M. W., Castello, A., Schwarzl, T. and Preiss, T.: A brave new world of RNA-binding proteins, *Nature reviews Molecular cell biology*, Vol. 19, No. 5, pp. 327–341 (2018).
[2] Gebauer, F., Schwarzl, T., Valcárcel, J. and Hentze, M. W.: RNA-binding proteins in human genetic disease, *Nature Reviews Genetics*, Vol. 22, No. 3, pp. 185–198 (2021).
[3] Gerstberger, S., Hafner, M. and Tuschl, T.: A census of

human RNA-binding proteins, *Nature Reviews Genetics*, Vol. 15, No. 12, pp. 829–845 (2014).

[4] Berglund, A.-C., Sjölund, E., Östlund, G. and Sonnhammer, E. L.: InParanoid 6: eukaryotic ortholog clusters with inparalogs, *Nucleic acids research*, Vol. 36, No. suppl_1, pp. D263–D266 (2007).

[5] Jensen, K. B. and Darnell, R. B.: CLIP: crosslinking and immunoprecipitation of in vivo RNA targets of RNA-binding proteins, *RNA-Protein Interaction Protocols*, Springer, pp. 85–98 (2008).

[6] Van Nostrand, E. L., Pratt, G. A., Shishkin, A. A., Gelboin-Burkhart, C., Fang, M. Y., Sundararaman, B., Blue, S. M., Nguyen, T. B., Surka, C., Elkins, K. et al.: Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP), *Nature methods*, Vol. 13, No. 6, pp. 508–514 (2016).

[7] Suresh, V., Liu, L., Adjeroh, D. and Zhou, X.: RPI-Pred: Predicting ncRNA-protein interaction using sequence and structural information, *Nucleic Acids Research*, Vol. 43, No. 3, pp. 1370–1379 (online), DOI: 10.1093/nar/gkv020 (2015).

[8] Peng, C., Han, S., Zhang, H. and Li, Y.: Rpiter: A hierarchical deep learning framework for ncRNA-protein interaction prediction, *International Journal of Molecular Sciences*, Vol. 20, No. 5 (online), DOI: 10.3390/ijms20051070 (2019).

[9] Pan, X., Rijnbeek, P., Yan, J. and Shen, H. B.: Prediction of RNA-protein sequence and structure binding preferences using deep convolutional and recurrent neural networks, *BMC Genomics*, Vol. 19, No. 1, pp. 1–11 (online), DOI: 10.1186/s12864-018-4889-1 (2018).

[10] Muppirala, U. K., Honavar, V. G. and Dobbs, D.: Predicting RNA-Protein Interactions Using Only Sequence Information, *BMC Bioinformatics*, Vol. 12, No. 1 (online), DOI: 10.1186/1471-2105-12-489 (2011).

[11] Zhang, S. W., Zhang, X. X., Fan, X. N. and Li, W. N.: LPI-CNNCP: Prediction of lncRNA-protein interactions by using convolutional neural network with the copy-padding trick, *Analytical Biochemistry*, Vol. 601, No. April, p. 113767 (online), DOI: 10.1016/j.ab.2020.113767 (2020).

[12] Wang, L., Yan, X., Liu, M. L., Song, K. J., Sun, X. F. and Pan, W. W.: Prediction of RNA-protein interactions by combining deep convolutional neural network with feature selection ensemble method, *Journal of Theoretical Biology*, Vol. 461, pp. 230–238 (online), DOI: 10.1016/j.jtbi.2018.10.029 (2019).

[13] Fan, X. N. and Zhang, S. W.: LPI-BLS: Predicting lncRNA–protein interactions with a broad learning system-based stacked ensemble classifier, *Neurocomputing*, Vol. 370, pp. 88–93 (online), DOI: 10.1016/j.neucom.2019.08.084 (2019).

[14] Pan, X., Fan, Y. X., Yan, J. and Shen, H. B.: IPMiner: Hidden ncRNA-protein interaction sequential pattern mining with stacked autoencoder for accurate computational prediction, *BMC Genomics*, Vol. 17, No. 1, pp. 1–14 (online), DOI: 10.1186/s12864-016-2931-8 (2016).

[15] Wang, Y., Chen, X., Liu, Z. P., Huang, Q., Wang, Y., Xu, D., Zhang, X. S., Chen, R. and Chen, L.: De novo prediction of RNA-protein interactions from sequence information, *Molecular BioSystems*, Vol. 9, No. 1, pp. 133–142 (online), DOI: 10.1039/c2mb25292a (2013).

[16] Consortium, E. P. et al.: An integrated encyclopedia of DNA elements in the human genome, *Nature*, Vol. 489, No. 7414, p. 57 (2012).

[17] Song, J., Tian, S., Yu, L., Xing, Y., Yang, Q., Duan, X. and Dai, Q.: AC-Caps: Attention Based Capsule Network for Predicting RBP Binding Sites of LncRNA, *Interdisciplinary Sciences: Computational Life Sciences*, Vol. 12, No. 4, pp. 414–423 (online), DOI: 10.1007/s12539-020-00379-3 (2020).

[18] Berman, H., Henrick, K., Nakamura, H. and Markley, J. L.: The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data, *Nucleic acids research*, Vol. 35, No. suppl_1, pp. D301–D303 (2007).

[19] Pan, X., Fang, Y., Li, X., Yang, Y. and Shen, H.-B.: RBPsuite: RNA-protein binding sites prediction suite based on deep learning, *BMC genomics*, Vol. 21, No. 1, pp. 1–8 (2020).

[20] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. and Polosukhin, I.: Attention is all you need, *Advances in neural information processing systems*, pp. 5998–6008 (2017).

[21] Yamada, K. and Hamada, M.: Prediction of RNA-protein interactions using a nucleotide language model, *bioRxiv*, p. 2021.04.27.441365 (online), available from ⟨https://doi.org/10.1101/2021.04.27.441365⟩ (2021).

[22] Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding, *arXiv preprint arXiv:1810.04805* (2018).

[23] Kimura, T., Yasuo, N., Sekijima, M. and Lustig, B.: Statistical potentials for RNA-protein interactions optimized by CMA-ES, *Journal of molecular graphics & modelling*, Vol. 110, p. 108044.

[24] Maziarka, Ł., Danel, T., Mucha, S., Rataj, K., Tabor, J. and Jastrzabski, S.: Molecule attention transformer, *arXiv preprint arXiv:2002.08264* (2020).

[25] Ji, Y., Zhou, Z., Liu, H. and Davuluri, R. V.: DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome, *Bioinformatics*, Vol. 37, No. 15, pp. 2112–2120 (2021).

[26] Rao, R., Bhattacharya, N., Thomas, N., Duan, Y., Chen, X., Canny, J., Abbeel, P. and Song, Y. S.: Evaluating protein transfer learning with TAPE, *Advances in neural information processing systems*, Vol. 32, p. 9689 (2019).

[27] Cheng, Y., Wang, R., Pan, Z., Feng, R. and Zhang, Y.: Look, listen, and attend: Co-attention network for self-supervised audio-visual representation learning, *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 3884–3892 (2020).

[28] Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G. A., Sonnhammer, E. L., Tosatto, S. C., Paladin, L., Raj, S., Richardson, L. J. et al.: Pfam: The protein families database in 2021, *Nucleic Acids Research*, Vol. 49, No. D1, pp. D412–D419 (2021).

[29] Teng, X., Chen, X., Xue, H., Tang, Y., Zhang, P., Kang, Q., Hao, Y., Chen, R., Zhao, Y. and He, S.: NPInter v4. 0: an integrated database of ncRNA interactions, *Nucleic acids research*, Vol. 48, No. D1, pp. D160–D165 (2020).