4J1-5in2

タンパク質二次構造予測を行う深層学習モデルの Saliency による可視化

Visualization of Deep Neural Network Learned Protein Secondary Structure Prediction with Saliency

河野圭祐 *1 Keisuke Kawano　小出智士 *1 Satoshi Koide　今村千絵 *1 Chie Imamura　田所幸浩 *1 Yukihiro Tadokoro

*1 豊田中央研究所
Toyota Central R&D Labs. Inc.

タンパク質の立体構造の局所構造である二次構造をアミノ酸配列から予測する研究が多数行われており，中でも深層学習による予測モデルは一般に高精度である．一方で，学習済みの深層学習による予測モデルの内部はブラックボックスであるため，生物学的に妥当な特徴に着目して予測しているかは不明であり，予測結果を保証できない．本研究ではタンパク質二次構造予測モデルを Saliency を用いて可視化し，生物学的な知見と比較する．その結果，一部の二次構造の予測について，予測モデルが生物学的にも妥当な特徴に注目して予測を行っていることを明らかにした．

1. 目的

酵素はタンパク質であり，アミノ酸のつながったポリペプチド鎖が折りたたまれて三次元立体構造をとる．酵素の構造は機能や特性に大きく影響するが，主な解析手法である X 線などによる立体構造解析には時間がかかり，全ての構造を解明することは困難である．そのため，比較的容易に取得可能なアミノ酸配列から立体構造を予測する技術が必要である．

機械学習によって，アミノ酸配列情報からタンパク質の局所的な構造 (二次構造) を予測する研究は数多く行われている [1, 2, 3, 4, 5, 6, 7]．二次構造は α-helix や β-sheet などの 8 種類に分類することができる [8]．二次構造予測はアミノ酸の配列などの情報から，二次構造を予測する分類問題 (系列ラベリング) として定義される．

近年，Deep Neural Network(DNN) を用いた二次構造予測モデルが提案され，高精度な予測ができることが報告されている [1, 2, 4, 6]．一方で，DNN は数千から数百万という多数のパラメータが関与する非線形な関数であり，その内部はブラックボックスであるため，生物学的に妥当な特徴に着目して予測しているかは不明であり，未知のタンパク質に対する予測結果を保証できない．

本研究では，二次構造を予測する DNN が妥当な予測モデルを獲得しているかを検証する．本研究では Saliency[9] を用いて DNN の内部状態を可視化し，予測時にアミノ酸配列のどのような特徴に着目したかを明らかにする．そしてその可視化結果を生物学的な知見と比較することで，二次構造のいくつかの種類の予測について，DNN が生物学的にも妥当な部分に着目していることを示す．例えば，ある位置の二次構造ラベルを予測する際，付近にどのようなアミノ酸が存在するかを一切考慮していない場合は DNN が過学習している可能性が高いといえる．

2. タンパク質二次構造

タンパク質全体の立体構造は多様だが，数個から 20 個程度のアミノ酸によって構成される局所的な立体構造は二次構造とよばれ，数種類に分類することができる．二次構造には，いくつかの分類方法が存在しているが，本研究では Sander et al.[8] の 8 種類の分類 (Q8) を用いる．表 1 に二次構造の分類および略称を示す．

連絡先: 河野圭祐，豊田中央研究所，愛知県長久手市横道 41-1，kawano@mosk.tytlabs.co.jp

表 1: 二次構造の分類

名称	略称
irregular	L
β-bridge	B
β-strand	E
3_{10}-helix	G
π-helix	I
α-helix	H
bend	S
β-turn	T

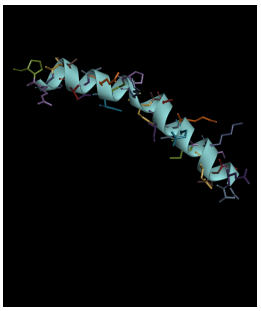

図 1: α-helix

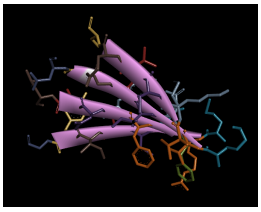

図 2: β-strand

代表的な二次構造である α-helix および β-strand をそれぞれ図 1 および図 2 に示す．α-helix は平均 3.6 残基が 1 周期の右巻きのらせん構造である [10]．このらせん構造において，全てのアミノ酸が 4 残基離れたアミノ酸と水素結合を形成することで，エネルギー的に安定な構造を保っている．一方で β-strand は直線上にアミノ酸が連なっている．

3. 二次構造予測に用いた DNN の構造と予測精度

タンパク質二次構造予測問題では，アミノ酸配列などの特徴量を入力として，二次構造のラベルを予測する．二次構造予測は系列ラベリング問題であり，アミノ酸配列のそれぞれの箇所

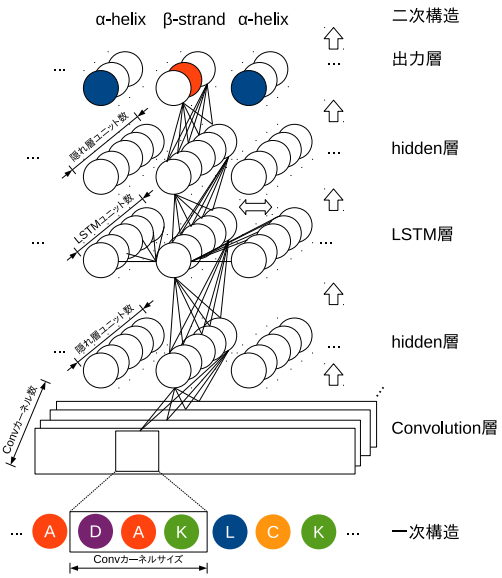

図 3: タンパク質二次構造予測のための DNN の概略図

について二次構造が予測される．学習及びテストに用いるデータセットとして，CB513[1] データセットを用いた．このデータセットには学習データとして 5534 配列，テストデータとして 514 配列が含まれる．アミノ酸配列の長さはタンパク質によって異なるが，このデータセットでは全ての配列が長さ 700 に統一される．長さを統一する際に，長さが足りない配列は配列の末尾から 700 まで NoSeq という記号によって埋められる．

このデータセットでは入力として，アミノ酸配列中のそれぞれのアミノ酸に対して，以下の $c = 46$ 次元の特徴量が定義されている．

- アミノ酸の onehot 表現 (20 種類のアミノ酸と未知またはその他のアミノ酸を表すワイルドカード，NoSeq に対応する 22 次元ベクトル)．
- C 末端もしくは N 末端かどうか (アミノ酸配列の両端に当たるかどうか，0 または 1 の値を取る 2 次元ベクトル)．
- 配列プロファイル [1] (類似するアミノ酸配列において，それぞれの位置にどのアミノ酸または NoSeq が多いかという 22 次元のベクトル)．

本研究では [11, 12] を参考にして，図 3 のような DNN を構成した．図に示すように，DNN は Convolution 層と bi-directional LSTM 層 [13] をもつ．bi-directional LSTM を用いることで，アミノ酸配列の両端から入力することができる．Convolution 層のカーネルサイズ等のパラメータを表 2 に示す．これらのパラメータは交差検証によって決定された．

予測モデルは学習データで学習した後，テストデータにおける各アミノ酸に対する二次構造の種類 [8] を予測し，予測精度を求められる．今回可視化に用いた DNN の予測精度は 68%であり，これは現状の state-of-the-art(68.3%[3]) に近い性能である．Confusion Matrix および二次構造の分布を図 4 に示す．なお，分布のヒストグラムにおいて，NoSeq の項は省略した．この図より与えられたデータセットにおいて，二次構造の数には大きな偏りがあり，極端に数が少ない二次構造ラベルの予測が失敗する傾向があることがわかる．特に 3_{10}-helix (G) および π-helix (I) は α-helix (H) と誤って予測されることが多い．

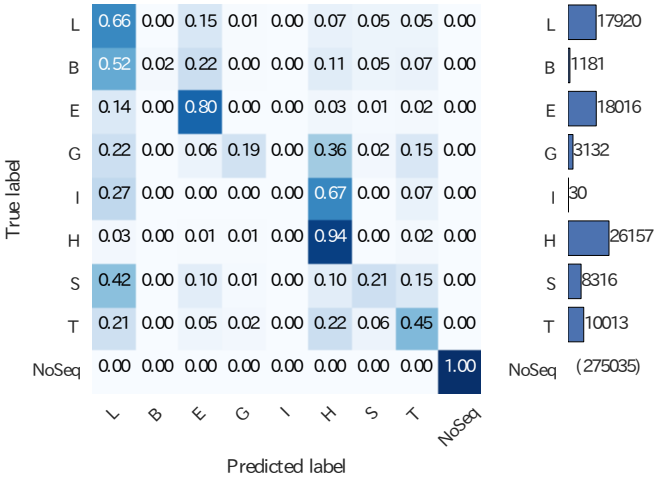

図 4: テストデータに対する Confusion Matrix．分布のヒストグラムのうち NoSeq の項目はグラフを表示していない．

このような誤り傾向に対する予測精度の向上は重要な課題だが，本論文では可視化の問題の取り扱いが主要な課題であるため，これ以上は取り扱わない．

4. Saliency による DNN の可視化

DNN の出力層には判別対象のクラス数と同数のニューロンがあり，ある入力ベクトル $x_0 \in \mathbb{R}^c$ が与えられたとき，出力層にある各ニューロン $l = \{\mathrm{L}, \mathrm{B}, \mathrm{E}, \mathrm{G}, \mathrm{I}, \mathrm{H}, \mathrm{S}, \mathrm{T}, \mathrm{NoSeq}\}$ の出力値 S_l で最も大きい値を求める．そして，そのニューロンに対応したラベル $\mathrm{argmax}_l S_l$ が予測結果として選択される．このとき，Saliency は入力 x に関する偏微分の値として式 1 のように定義される．

$$(\text{Saliency}) = \max_c \left| \frac{\partial S_l}{\partial x}\Big|_{x_0} \right| \tag{1}$$

Saliency は一種の感度解析の結果を表しているといえる．例えば，α-helix に対応する出力層のニューロンに対して Saliency を求める場合を考える．このとき Saliency は α-helix に対応する出力層のニューロンを発火させるために，局所的に入力のどの部分を変化させればよいかを可視化する．例えば，Saliency のうちアミノ酸配列のある位置の値が大きな値であることは，その位置の入力を変化させることで，出力に大きな影響があることを意味する．

Saliency を用いると，ある位置 j の二次構造ラベル l_j を予測する際に，周辺にあるどの位置のアミノ酸 (特徴量) が大きく寄与しているかを知ることができる．例えば十分離れた位置

表 2: DNN のパラメータ

パラメータ名	値
LSTM ノード数	800
LSTM 層数	1
Conv カーネルサイズ	17
Conv カーネル数	128
Dropout	あり
LSTM の種類	bi-directional

など，予測と関係しない位置ではその影響が限りなくゼロに近づく傾向があると期待される．

そして，これらの結果と生物学的にすでに知られている知見が一致すれば，学習された DNN は生物学的に妥当な特徴をうまくとらえているとみなすことができる．

本研究では，特に予測の精度が高い α-helix および β-strand について，それらが 9 個以上連続している部分に着目して Saliency によって可視化する．α-helix は平均 3.6 残基が 1 周期の右巻きのらせん構造である [10]．よって，アミノ酸配列のある位置の二次構造が α-helix であるかどうかを予測する際には，3，4 残基離れた位置にどのようなアミノ酸が存在するかが重要であると考えられる．一方で，β-strand はアミノ酸が直鎖状に連なった構造をしており，予測の際には，単に距離が近いアミノ酸との関係が重要であると考えられる．DNN が正しい予測モデルを獲得している場合には，α-helix を予測する際には β-strand を予測する際と比較して，3，4 残基離れた位置の Saliency の値が高くなることが予想される．

Saliency は各出力のニューロンに対して，各入力の各特徴量に対応する値を求める手法である．このため，今回のデータに対して，Saliency を求めると，N(配列数) $\times$ 9(ラベル数) $\times$ 42(特徴量数) $\times$ L(配列長) 個の値が求まる．本研究では，これらの Saliency の値から以下の条件をすべて満たすものを抜き出し，統計量を求める．

1. α-helix または β-strand が 9 個以上連続している部分に対応するもの．
2. 二次構造の予測結果が正しいもの．
3. 出力ニューロンが α-helix または β-strand に対応するもの．

評価に用いたアミノ酸配列 514 配列の中に，α-helix が 9 個以上連続している部分，かつ，二次構造予測が正しいものは計 8807 箇所，β-strand が 9 個以上連続している部分，かつ，二次構造予測が正しいものは計 3566 箇所あった．今回の可視化では，α-helix または，β-strand をとった位置からの相対位置での Saliency を求める必要がある．よって，それぞれの Saliency の値に対して，α-helix または β-strand をとった位置 j からの距離が同じものを集めて平均を求める．

5. 可視化結果

Saliency による可視化の結果を図 5 に示す．β-strand の予測では配列上の距離が遠くなるにつれて，徐々に Saliency の値が減少している．一方で，α-helix の予測については，1 残基目つまり隣のアミノ酸から 3 残基目まで Saliency の値が減少していない．これは α-helix の 1 周期が 3.6 残基であることと一致する．

以上から，可視化した二次構造予測の DNN では，α-helix および β-strand の予測において，生物学的にも妥当な部分に着目して予測していることが示唆された．

一方で，今回の可視化では，α-helix または β-strand が 9 個以上連続している部分のみを対象にした．これは，3.6 残基を 1 周期とするという α-helix の性質が Saliency による可視化で容易に確認できるだろうと予め予想できたからである．DNN が妥当な予測モデルを獲得しているかをより正しく評価するためには，α-helix 以外の部分についても可視化を行い，生物学的な知見と比較する必要がある．

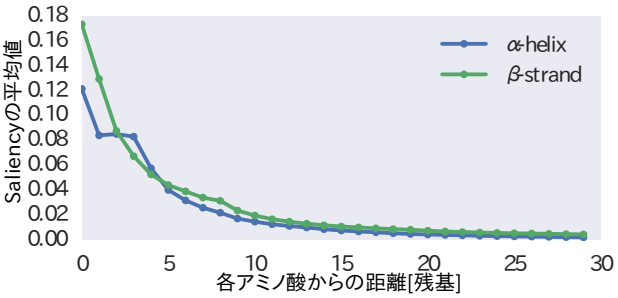

図 5: Saliency による可視化の結果

6. まとめと今後の課題

タンパク質の二次構造を予測する DNN の内部状態を Saliency を用いて可視化した．DNN による α-helix および β-strand の予測に対する Saliency を求めたところ，以下の結果が得られた．

1. β-strand の予測では配列上の距離が遠くなるにつれて，徐々に Saliency の値が減少する．
2. α-helix の予測については，1 残基目つまり隣のアミノ酸から 3 残基目まで Saliency の値が減少していない．

これらの結果は生物学的な知見と一致するため，DNN が入力された特徴量のうち，生物学的にも妥当な部分に着目して α-helix および β-strand の予測を行っていることがわかった．

今後の課題として以下がある．

- α-helix，β-strand 以外のラベルに対する Saliency による可視化と生物学的な知見との比較．
  本研究では，生物学的な知見との比較が容易な α-helix，β-strand についてのみ可視化した．DNN が生物学的に妥当な内部状態で二次構造を予測しているか評価するためには，8 種類ある二次構造のラベル全てに対して可視化を行い，生物学的な知見と比較する必要がある．
- Saliency 以外の可視化手法の適用．
  二次構造を予測する DNN を Saliency を用いて可視化すると，非常に多くの Saliency が作成されてしまう．人間が解釈するためには，それらの Saliency から統計量を求める必要があるが，どのような統計量を求めるかを設計する必要がある．DNN の可視化方法として，Saliency の他に，Activation Maximization[14] などが提案されている．これらの別の可視化手法を用いることで，統計量を陽に設計することなく知見を取り出せる可能性がある．

参考文献

[1] J. Zhou and O. G. Troyanskaya. Deep supervised and convolutional generative stochastic network for protein secondary structure prediction. In ICML, pp. 745–753, 2014.

[2] S. Wang, J. Peng, J. Ma, and J. Xu. Protein Secondary Structure Prediction Using Deep Convolutional Neural Fields. Nature Scientific Reports, 2016.

[3] Z. Wang, F. Zhao, J. Peng, and J. Xu. Protein 8-class secondary structure prediction using conditional neural fields. Proteomics, Vol. 11, No. 19, pp. 3786–3792, 2011.

[4] C. N. Magnan and P. Baldi. Sspro/accpro 5: almost perfect prediction of protein secondary structure and relative solvent accessibility using profiles, machine learning and structural similarity. Bioinformatics, Vol. 30, No. 18, pp. 2592–2597, 2014.

[5] A. Drozdetskiy, C. Cole, J. Procter, and G. J. Barton. Jpred4: a protein secondary structure prediction server. Nucleic acids research, p. gkv332, 2015.

[6] S. K. Sønderby and O. Winther. Protein Secondary Structure Prediction with Long Short Term Memory Networks. Arxiv, 2016.

[7] 椿真史, 新保仁, 松本裕治. 表現学習と深層学習を用いたタンパク質の相同性探索と構造予測. 日本人工知能学会全国大会, 2016.

[8] W. Kabsch and C. Sander. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers, Vol. 22, No. 12, pp. 2577–2637, 1983.

[9] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034, 2013.

[10] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Robers, and P. Walter. Molecular biology of the cell, 5th edition. 2007.

[11] Protein And Secondary Structure Prediction with Convolutions and Vertical-Bi-Directional RNNs. `https://github.com/alrojp/CB513/blob/master/Article/cb513_artikel.pdf`. Accessed: 2016-10-28.

[12] S. K. Sønderby, C. K. Sønderby, H. Nielsen, and O. Winther. Convolutional lstm networks for subcellular localization of proteins. In International Conference on Algorithms for Computational Biology, pp. 68–80. Springer, 2015.

[13] A. Graves, N. Jaitly, and A.-r. Mohamed. Hybrid speech recognition with deep bidirectional lstm. In Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on, pp. 273–278. IEEE, 2013.

[14] J. Yosinski, J. Clune, A. Nguyen, T. Fuchs, and H. Lipson. Understanding neural networks through deep visualization. arXiv preprint arXiv:1506.06579, 2015.