Vol.2018-MPS-118 No.40
Vol.2018-BIO-54 No.40
2018/6/15

深層学習を用いたタンパク質予測立体構造モデルの評価

佐藤倫[1,a)]　石田貴士[1,b)]

概要：タンパク質の立体構造はタンパク質の機能に大きく関わり，創薬等の生命科学において重要な情報となる．実験的に立体構造を決定するのは時間的・金銭的にコストがかかるため，計算機を用いて立体構造を予測する手法が多く開発されてきた．一方で立体構造予測のためには予測立体構造の評価手法（Model Quality Assessment Program，MQAP）が必要となる．MQAP の多くは単一，または複数の天然構造らしさを表す統計ポテンシャル関数や明示的に作られた特徴量を用いて機械学習を行うものである．統計ポテンシャル関数はこれまで数多く考案されてきた．既存の統計ポテンシャル関数の多くは原子対，残基対などでの相互作用を考慮したものであった．これらは二体間の相互作用であるため 3 次元構造を捉えることができていなかった．一方，3 次元畳み込みニューラルネットワーク（3D Convolutional Neural Network，3DCNN）は従来動作認識，物体認識に用いられてきたが，近年ではタンパク質立体構造データの解析に用いられ，成功を収めている．これらの点を踏まえ本研究では 3DCNN を用いて多体間での相互作用を考慮した MQAP を考案する．そのためにタンパク質の局所環境を定義し，その定義された環境の特徴量を 3DCNN の入力とし 2 値分類を行い，その学習済みモデルを用いてタンパク質全体の天然らしさのスコアを出力する．この手法を用いて天然構造とそれを模して作られた人工のモデルのプールであるデコイセットを評価した結果，既存手法と同等，もしくはそれ以上の数の天然構造を認識することに成功した．

Model quality assessment for protein tertiary structure prediction model using deep learning

Sato Rin[1,a)]　Ishida Takashi[1,b)]

Abstract: The three-dimensional structure of a protein is related to its function, and it is important information in life science application such as drug discovery. Determination of the three-dimensional structure is costly in terms of time and money, thus many methods for predicting three-dimensional structure using a computer have been developed but the accuracy is still insufficient. Thus, evaluation of predicted model quality is required and such methods are called model quality assessment program (MQAP). Most of MQAPs use machine learning and statistical potential functions expressing single or plural natural structure likelihoods and explicitly created feature quantities. Numerous statistical potential functions have been devised so far. Many of the existing statistical potential functions consider interactions in atom pairs, residue pairs, and the like. Since these are interactions between two bodies, it was not possible to capture whole three-dimensional structure information. 3D convolutional neural network (3DCNN) has been used for conventional motion recognition and object recognition, but in recent years it has been used successfully for analysis of protein three-dimensional structure data. Based on these points, in this research, we devise an MQAP that considers interaction among multiple bodies using 3DCNN. For that purpose, the local environment of the protein is defined, binary classification is performed with the feature amount of the defined environment as input of 3DCNN, and the score of the naturalness of the whole protein is output using the learned model. As a result of evaluating a decoy set which is a pool of native structures and artificial models, we succeeded in recognizing the same number of native structures as the existing method or more.

[1] 東京工業大学情報理工学院情報工学系
Department of Computer Science, School of Computing, Tokyo Institute of Technology, Tokyo, Japan
[a)] sato@cb.cs.titech.ac.jp
[b)] ishida@cs.titech.ac.jp

Vol.2018-MPS-118 No.40
Vol.2018-BIO-54 No.40
2018/6/15

1. 序論

ゲノム解読の技術の向上により多くの生物のゲノム解読が進んでいる．その一方でそのゲノムから翻訳されるアミノ酸配列の示すタンパク質がどのような立体構造をしているかの解明はそれに追いつけない状態にある．2017 年 12 月現在公的なタンパク質立体構造情報のデータベースである Protein Data Bank[1] に登録されているタンパク質立体構造の数は約 13 万件，それに対して公的なアミノ酸配列情報のデータベースである UniProt[2] に登録されているアミノ酸配列の数は約 1 億件である．タンパク質構造はタンパク質の機能に関わるため，タンパク質の機能の解明には不可欠であり，創薬等の生命科学を行う上で重要な情報となる．タンパク質構造を実験的に決定する方法は NMR や X 線結晶解析などいくつかあるが，どれも時間的，金銭的にコストがかかる．そこで計算機を用いて立体構造を予測する研究が以前より盛んに行われており，多くのモデリング手法が考案されてきた．

モデリングの手法が様々存在し，また比較モデリングに関してはテンプレートに用いるタンパク質が異なると結果も異なるため，多様な予測立体構造を得ることができるが．その一方でそれらの予測立体構造のうち一番天然構造らしい構造を選ぶ必要があり，多くの手法が開発されてきた．ここで天然構造らしいとは天然構造との構造類似性が高いことを表す．これを Model Quality Assessment Program (MQAP)[3] と呼ぶ．多くの MQAP は単一，または複数の天然構造らしさを表す統計ポテンシャル関数 [4], [5] によって構成されており，近年では明示的に作られた特徴量に基づく機械学習による予測モデルなども提案されている [6], [7]．統計ポテンシャル関数は Protein Data Bank に既知の天然構造から構造的な特徴の分布に基づき，統計的につくられたポテンシャル関数のことであり数多く考案されてきた．これらの多くの統計ポテンシャル関数は原子対，残基対など 2 体間の相互作用を捉えたものが主であった [8]．しかし，タンパク質は 3 次元構造であるために立体構造の特徴を捉えられていなかった．近年このことを踏まえ多体間のポテンシャル関数が考案されてきた [8], [9], [10]．しかしこれらの多体間のポテンシャル関数は既存の 2 体間のものより良い精度を出すには至っていない．これは多体にすると問題がより複雑になりパラメーターも増えたことによるものだと思われる．よって多体間の相互作用を捉えるにはこれまでの統計的ポテンシャル関数を作る手法とは違う新たな手法が必要である．

畳み込み層をもつニューラルネットワーク である畳み込みニューラルネットワークはこれまで多くの分野で成功を収めた．これを 3 次元に拡張したものが 3 次元畳み込みニューラルネットワーク（3 Dimentional Convolutional Neural Network，3DCNN）である．3DCNN は従来動作認識 [11] や物体認識 [12] に用いられてきたが，近年ではタンパク質立体構造の解析に用いられ始めている [13], [14]．その中で明示的な特徴量を用いて機械学習をした既存手法より良い精度を収め，タンパク質立体構造の解析において 3DCNN の有効性を示唆した．このことを踏まえ，3DCNN を用いれば MQAP の分野においても有効であると考え，またタンパク質立体構造を多体で捉えることを可能になることから本研究では 3DCNN を用いることとした．

2. 提案手法

2.1 提案手法の流れ

本研究では多体間での相互作用を扱うために 3DCNN を用いる手法を提案する．3DCNN は 4 階のテンソルを入力として畳み込み計算をするが，この入力は bounding box と呼び，それぞれの格子を voxel と呼ぶ．また画像の各ピクセルごとの RGB に相当するものを voxel では channel と呼ぶ．タンパク質全体の天然構造らしさを評価するため本来タンパク質全体を 1 つの bounding box とし，それを評価することが考えられるが，タンパク質ごとに大きさが異なり，また 3 次元構造は画像のような 2 次元のものよりパラメーターが多く，大きな bounding box を扱うのは困難である．よって本研究では分割統治の考えで，局所構造に注目しその良し悪しを判定し，最後にそれらの結果をあわせてタンパク質全体の評価を行う．そのために，局所環境を定義することでタンパク質を分割する．以下のセクションで詳細を記述する．

2.2 局所環境の定義

タンパク質のある残基についての局所環境を定義する必要がある．

ある残基の周辺環境を bounding_box の立方体にするために，その立方体を作る基底が必要となる．タンパク質立体構造は PDB フォーマット形式で表され，各原子の xyz 座標が記述されている．この軸を単純に基底として用いると，回転した同じタンパク質の局所環境が入力となったときに同じ局所環境として認識することができない．そこで

$$\boldsymbol{a} = (C\alpha, C)^{\mathrm{T}}, \boldsymbol{b} = (C\alpha, N)^{\mathrm{T}}, \boldsymbol{c} = \boldsymbol{a} \times \boldsymbol{b}$$

として正規直交基底を得る [13]．これは残基によって一意に決定される基底であるため上記の問題が解決できる．この得られた正規直交基底を局所環境の基底として定める．図 1 のように 1 残基の $C\alpha$ 原子を中心に，この基底を用いて

$$v = (x, y, z)^t \text{ただし} -d \leq x, y, z < d$$

で表される $2d \times 2d \times 2d$ の立方体内にある原子，残基の情報を本研究の局所環境として定義する．今回は d=6[Å] とした．

Vol.2018-MPS-118 No.40
Vol.2018-BIO-54 No.40
2018/6/15

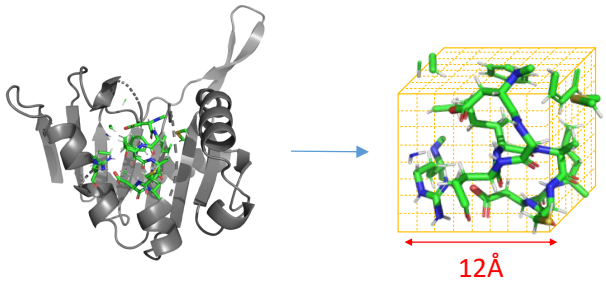

図 1　局所構造を抽出する

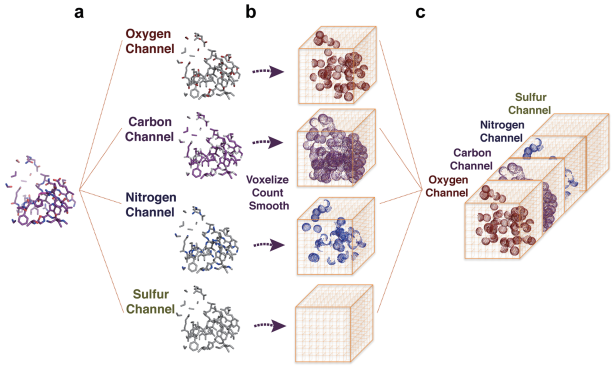

図 2　特徴量生成 [13]

2.3　特徴量の生成

Wen Torng らは図 2 のように channel を（炭素，酸素，窒素，硫黄）の 4 つとして，それぞれの voxel について各原子が存在する場合，対応する channel の値を 1 とし残りの 3 つの channel は 0 の値を格納する．その後分散を各原子のファンデルワールス半径の平均としてガウシアンフィルタをかける．[13]

本研究ではこれに加えて，各原子の属する残基の情報を表 1 の物性値を正規化して，対応する channel に格納する．

さらに，Stride[18] によって出力された 2 次構造情報の 7 class を同様に channel に追加する．また，DEPTH[19] を用いて各原子のタンパク質表面からの距離を計算し channel に追加した．計 15 channel となり，入力サイズは (12, 12, 12, 15) となった．

2.4　提案手法のスコア

本研究では 1 つのモデルに対して残基の数だけ出力を得るため，これらをタンパク質全体のスコアに変換する必要がある．各残基からのスコアを p_i とおく．複数試したが，本研究ではスコアは

$$SCORE = \frac{1}{N}\sum_{i=0}^{N} p_i$$

として変換した．

2.5　学習に用いたデータセット

あるアミノ酸配列について既知の天然構造と，その天然構造に似せモデリング等によって作ったニセの構造（デコイという）のプールをデコイセットと呼ぶ．本研究では Critical Assessment of protein Structure Prediction (CASP) で行われた構造予測のコンペティションで提出された予測立体構造をデコイとすることで，CASP8 から CASP10 まで [20][21][22] のドメイン毎に分けられた天然構造とその予測立体構造を計 436 個のデコイセットを取得し，これをデータセットとした．それぞれ平均して 434.7 個のデコイ構造が存在する．

表 1　アミノ酸の物性

アミノ酸	分子量 [15]	等電点 [16]	疎水性スコア [17]
A	71.0779	6.01	1.8
C	103.1429	5.05	2.5
D	115.0874	2.85	-3.5
E	129.1140	3.15	-3.5
F	147.1739	5.49	2.8
G	57.0513	6.06	-0.4
H	137.1393	7.60	-3.2
I	113.1576	6.05	4.5
K	128.1723	9.60	-3.9
L	113.1576	6.01	3.8
M	131.1961	5.74	1.9
N	114.1026	5.41	-3.5
P	97.1152	6.30	-1.6
Q	128.1292	5.65	-3.5
R	156.1857	10.76	-4.5
S	87.0773	5.68	-0.8
T	101.1039	5.60	-0.7
V	99.1311	6.00	4.2
W	186.2099	5.89	-0.9
Y	163.1733	5.64	-1.3

本研究では天然構造とデコイ構造の類似度からラベル付けを行った．2 つの立体構造の類似度は主に RMSD[23], GDT_TS[24], TMscore[25] が用いられる．本研究では TM-score というソフトウェアを用いて上記を計算し，天然構造との GDT_TS が 0.9 より大きい立体構造から得られる voxel は全て正例，0.35 より小さい立体構造から得られる voxel は全て負例としてラベル付をした．また負例は 1/10 にダウンサンプリングした．結果サンプル数は 193 万，うち正例は 27%，負例は 63%となった．

2.6　学習方法

ニューラルネットワークの学習のためのライブラリは keras[26] を用い，層構造は補足図 3 のものを用いた．これは [13] の層構造をもとに，本研究の入力サイズを加味して改良したものである．Drop out 層は p=0.3，活性化関数は出力層のみ sigmoid，それ以外は relu，最適化アルゴリズム は adam[27]，loss は binary_crossentropy を用いた．重みパラメーターの総数は約 220 万となった．

Vol.2018-MPS-118 No.40
Vol.2018-BIO-54 No.40
2018/6/15

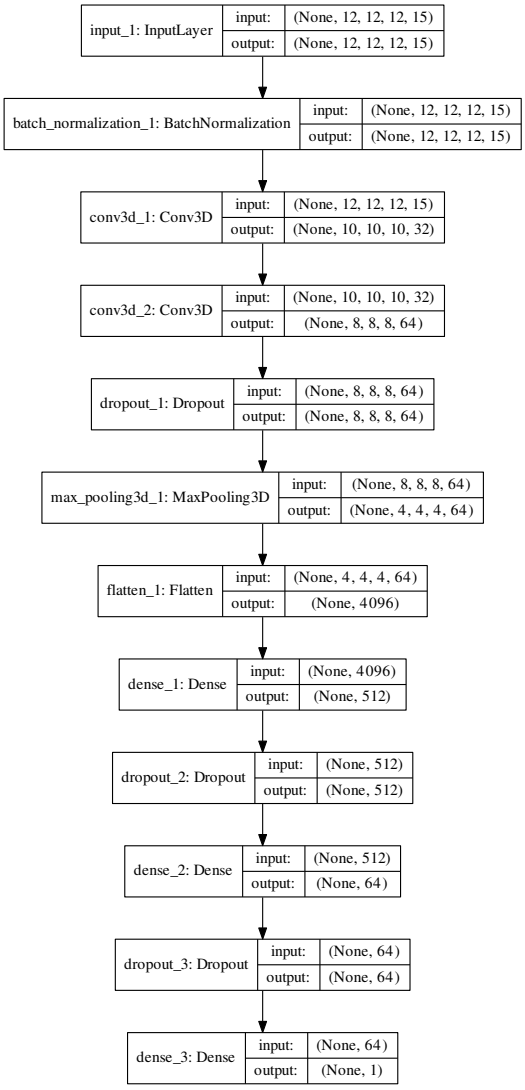

図 3　層構造

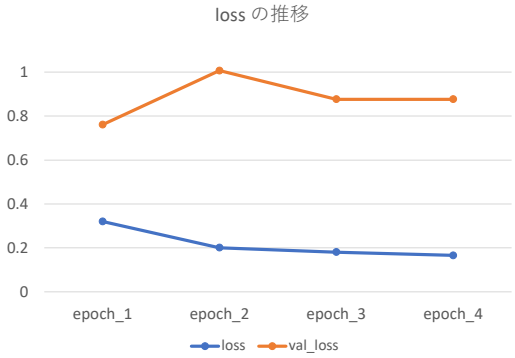

図 4　損失の推移の推移

3. 性能評価

この章では前章で学習したモデルを用いてデコイセットを評価し，予測立体構造に対する天然構造らしさのスコアを出力し，その出力の性能を既存手法と比較する．

3.1 4 章の学習結果

性能評価の前に，まず前章の学習結果について記述する．損失の推移は図 4 のような結果となった．ただしこれはタンパク質全体の評価ではなく，局所環境の評価についてである．train_loss と validation_loss が大きく乖離しており，また学習が 1epoch 目で終えていることがわかる．これは問題が簡単であるためにテストデータを全データの学習する前にすでにある程度学習しきっているために起きていることが考えられる．今回の学習では既知の天然構造が約 13 万あるのに対し，たかだか 400 程度の天然構造とそのデコイ構造しか用いていないため，validation_loss を小さくするためには更に多くのタンパク質を用いる必要があると思われる．

本研究ではこれらのモデルのうち validation loss が一番小さい epoch_1 のモデルを用いて性能評価を行う．

まず，このモデルに関して bounding_box の単位で学習できているかを確認する．そのための評価指標として ROC 曲線下の面積である AUC（Area Under the Receiver Operating Characteristic curve）を計算する．学習したモデルを用いたテストデータによる ROC 曲線は図 5 のようになった．この結果から汎化性能をある程度もって学習できていることが示された．

3.2 テストデータセット

次に実際に MQAP に本手法を適用して性能を評価していく．MQAP の性能評価はデコイセットに対して手法を適用してスコアを出力し，スコア最上位のモデルが天然構造であるか，また天然構造に近い構造をどれだけスコア上位で検出できるかなどが指標とされる．本研究では性

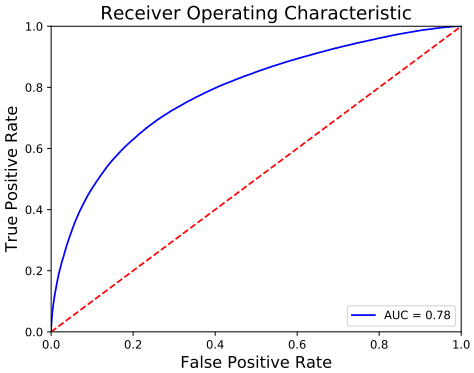

図 5　epoch_1 のモデルをもちいてテストデータを予測した際の ROC 曲線

能比較に頻繁に用いられる 'R' Us（4state_reduced，Fisa，Fisa_casp3，Lmds，Lattice_ssfit，hg_structal，ig_structal，ig_structal_hires）デコイセット，ROSETTA デコイセット，I-TASSER デコイセットを用いた．一般的には MOULDER デコイセットも用いられることが多いが，いくつかのタンパク質で TMscore[25] を実行できなかったため，今回は用いなかった．詳細は表 2 に記述する．

表 2　デコイセットの詳細

デコイセット	タンパク質の数	平均のデコイの数
4state_reduced	7	665
Fisa	4	1432
Fisa_casp3	5	1432
Lmds	11	439
Lattice_ssfit	8	2000
hg_structal	29	29
ig_structal	61	61
ig_structal_hires	20	19
ROSETTA	58	100
I-TASSER	56	438.2

3.3　性能評価の指標

手法の MQAP の性能としてどれだけ天然構造を識別できるか，また天然構造に近い構造を識別できるかなどを評価するために，上記のデコイセットを用いて以下の指標で性能評価を行う．

- 手法を用いて最もよいスコアの立体構造が天然構造であるデコイセットの数
- 各立体構造にスコアを付した時の天然構造の Zscore
- 天然構造を省いて最もよいスコアの立体構造の天然構造に対する TMscore
- 天然構造を省いた時のスコアと TMscore の相関係数

3.4　結果

上記のデコイセットを用いて既存手法のポテンシャル関数である DFIRE, DOPE, RW, RWplus, dDFIRE, OPUS-PSP，GOAP と比較した．各表は最も良い値は太字で示されている．

表 3　天然構造認識

Decoy sets	RWplus	dDFIRE	OPUS-PSP	GOAP	提案手法	#targets
4state_reduced	6	**7**	**7**	**7**	6	7
fisa	**3**	**3**	**3**	**3**	**3**	4
fisa_casp3	4	4	**5**	**5**	4	5
lmds	7	6	**8**	7	6	10
lattice_ssfit	**8**	**8**	**8**	**8**	**8**	8
hg_structal	12	16	18	22	**23**	29
ig_structal	0	26	20	**47**	41	61
ig_structal_hires	0	16	14	**18**	**18**	20
ROSETTA	20	12	39	45	**51**	58
I-TASSER	**56**	48	55	45	48	56
No.	116	146	177	207	**208**	258

表 4　天然構造の Zscore の平均

Decoy sets	RWplus	dDFIRE	OPUS-PSP	GOAP	提案手法	#targets
4state_reduced	-3.51	-4.15	-4.49	-4.38	**5.50**	7
fisa	**-4.79**	-3.80	-4.24	-3.97	4.70	4
fisa_casp3	-5.17	-4.83	-6.33	-5.27	**6.51**	5
lmds	-1.03	-2.44	**-5.63**	-4.07	4.69	10
lattice_ssfit	-8.85	-10.12	-6.75	-8.38	**13.83**	8
hg_structal	-1.74	-1.33	1.87	**-2.73**	2.38	29
ig_structal	1.11	-1.02	0.69	-1.62	**2.11**	61
ig_structal_hires	0.32	-2.05	-0.77	**-2.35**	2.30	20
ROSETTA	-1.47	-0.83	-3.00	-3.70	**4.50**	58
I-TASSER	-5.77	-5.03	**-7.43**	-5.36	6.22	56
No.	-2.08	-2.50	-2.71	-3.57	**4.27**	258

表 5　最良モデルの TMscore の平均

Decoy sets	Rwplus	dDFIRE	OPUS-PSP	GOAP	提案手法	#targets
4state_reduced	0.667	0.732	0.755	**0.818**	0.817	7
Fisa	0.434	0.454	0.405	**0.475**	0.392	4
Fisa_casp3	0.277	0.309	0.270	0.300	**0.347**	5
Lmds	0.346	0.364	0.339	0.339	**0.365**	10
Lattice_ssfit	0.251	**0.266**	0.248	0.248	0.249	8
hg_structal	**0.891**	**0.891**	**0.891**	0.889	0.885	29
ig_structal	0.948	0.948	**0.953**	0.946	0.943	61
ig_structal_hires	**0.950**	0.946	0.946	0.944	0.939	20
ROSETTA	0.505	0.480	0.506	**0.511**	0.458	58
I-TASSER	0.577	**0.578**	0.547	0.567	0.573	56
All	0.688	0.686	0.684	**0.691**	0.679	258

3.4.1　天然構造認識

表 3 は天然構造がスコア最上位である回数を示している．提案手法は既存手法に遜色ない性能を示した．特に ROSETTA デコイセットに関しては他既存手法より多くの天然構造を検出できた．また表 4 では天然構造のスコアを Zscore 化した平均を表している．この値が大きいほどよりよく天然構造を認識できていることを表す．提案手法は多くのデコイセットで既存手法よりも大きな値となった．

3.4.2　最良モデルの TMscore の平均

実際の問題では天然構造は存在しないため，予測立体構造のプールから天然構造に最も近い構造を選択できることが求められる．ここでは天然構造をデコイセットから除いた際の，スコア最上位の構造がどれだけ天然構造に近いかを表 5 で比較する．2 つの立体構造がどれだけ類似しているかを示す指標として TMscore を用いる．提案手法は既存手法と並ぶ性能となった．

表 6　各手法のスコアと TMscore の相関係数

Decoy sets	DFIRE	Rwplus	dDFIRE	OPUS-PSP	GOAP	提案手法	#targets
4state_reduced	-0.635	-0.606	-0.693	-0.589	**-0.694**	0.689	7
Fisa	-0.446	**-0.462**	-0.461	-0.282	-0.347	0.428	4
Fisa_casp3	-0.243	**-0.240**	-0.149	-0.095	-0.221	0.145	5
Lmds	-0.118	-0.147	**-0.248**	-0.091	-0.146	0.166	10
Lattice_ssfit	-0.094	**-0.097**	-0.070	-0.051	-0.058	0.027	8
hg_structal	-0.817	-0.806	-0.796	-0.752	**-0.825**	0.722	29
ig_structal	-0.785	-0.782	-0.766	-0.779	**-0.865**	0.714	61
ig_structal_hires	-0.876	**-0.879**	-0.844	-0.832	-0.885	0.744	20
ROSETTA	-0.441	-0.444	-0.393	-0.343	**-0.476**	0.263	58
I-TASSER	**-0.519**	-0.488	-0.525	-0.284	-0.477	0.381	56
All	-0.593	-0.586	-0.579	-0.499	**-0.612**	0.485	258

3.4.3　スコアと TMscore の相関係数

各手法のスコアと TMscore について，ピアソンの相関係数を用いて表 6 で比較する．相関係数が大きいほど，天然構造に近い構造を得ることができる．この指標では既存手法に劣る結果となった．

4. 結論

4.1　本研究の結論

本研究では多体間の相互作用を捉えた予測立体構造の評価手法を開発するために，3 次元畳み込みニューラルネットワークを用いてタンパク質の局所環境を評価し，タンパク質全体のスコアを局所環境の評価の平均として出力する予測立体構造の評価手法を開発した．デコイセットを用いて性能評価をした結果，天然構造認識については既存手法よりも良い性能であり，またスコア最上位の TMscore の平均についても既存手法に並ぶ性能を示した．このことは，3 次元畳み込みニューラルネットワークを用いてタンパク質の局所構造を評価することの妥当性を示唆していると考えられる．

4.2　今後の課題

本研究の学習においてはデコイセットを用いて学習しているが，デコイセットは現在取得可能なものを全て集めても 1000 に満たない．これは既知のタンパク質立体構造約 13 万には遠く及ばない数字であり，このままのフレームワークでは層構造や特徴量を工夫しても精度の改善は多少しか見込めないため，他の学習方法を模索する必要がある．

参考文献

[1] P. W. Rose *et al.*: The RCSB protein data bank: Integrative view of protein, gene and 3D structural information, *Nucleic Acids Res.*, Vol. 45, No. D1, pp. D271–D281 (online), DOI: 10.1093/nar/gkw1000 (2017).

[2] A. Bateman *et al.*: UniProt: The universal protein knowledgebase, *Nucleic Acids Res.*, Vol. 45, No. D1, pp. D158–D169 (online), DOI: 10.1093/nar/gkw1099 (2017).

[3] D. Kihara, H. Chen, Y. D. Yang: Quality assessment of protein structure models., *Curr. Protein Pept. Sci.*, Vol. 10, No. 3, pp. 216–228 (online), DOI: 10.2174/138920309788452173 (2009).

[4] H. Zhou, J. Skolnick: GOAP: A generalized orientation-dependent, all-atom statistical potential for protein structure prediction, *Biophys. J.*, Vol. 101, No. 8, pp. 2043–2052 (online), DOI: 10.1016/j.bpj.2011.09.012 (2011).

[5] A. Sali, A. Sali: Statistical potential for assessment and prediction of protein structures, *Protein Sci.*, pp. 2507–2524 (online), DOI: 10.1110/ps.062416606.Instead (2006).

[6] B. Manavalan, J. Lee: SVMQA: supportvector-machine-based protein single-model quality assessment, *Bioinformatics*, Vol. 33, No. April 2017, pp. 2496–2503 (online), DOI: 10.1093/bioinformatics/btx222 (2017).

[7] A. Ray, E. Lindahl, B. Orn Wallner: Improved model quality assessment using ProQ2, *BMC Bioinformatics*, Vol. 13, pp. 1–12 (online), DOI: 10.1186/1471-2105-13-224 (2012).

[8] M. Masso: All-Atom Four-Body Knowledge-Based Statistical Potentials to Distinguish Native Protein Structures from Nonnative Folds, Vol. 2017 (2017).

[9] Y. Feng, A. Kloczkowski, R. L. Jernigan: Four-body contact potentials derived from two protein datasets to discriminate native structures from decoys, *Proteins Struct. Funct. Genet.*, Vol. 68, No. 1, pp. 57–66 (online), DOI: 10.1002/prot.21362 (2007).

[10] B. Krishnamoorthy: Development of a Four-Body Statistical Pseudo-Potential to Discriminate Native from Non-Native Protein Conformations (2001).

[11] J. Carreira, A. Zisserman: Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset, (online), DOI: 10.1109/CVPR.2017.502 (2017).

[12] D. Maturana, S. Scherer: VoxNet : A 3D Convolutional Neural Network for Real-Time Object Recognition.

[13] W. Torng, R. B. Altman: 3D deep convolutional neural networks for amino acid environment similarity analysis, pp. 1–23 (online), DOI: 10.1186/s12859-017-1702-0 (2017).

[14] J. Jiménez *et al.*: DeepSite: Protein-binding site predictor using 3D-convolutional neural networks, *Bioinformatics*, Vol. 33, No. 19, pp. 3036–3042 (online), DOI: 10.1093/bioinformatics/btx350 (2017).

[15] S. D. Maleknia, R. Johnson: Mass Spectrometry of Amino Acids and Proteins, *Amin. Acids, Pept. Proteins Org. Chem.*, Vol. 5, pp. 1–50 (online), DOI: 10.1002/9783527631841.ch1 (2012).

[16] 菅. 二三男：マクマリー・生物有機化学 生化学編，丸善，第 2 edition (2007).

[17] J. Kyte, R. F. Doolittle: A simple method for displaying the hydropathic character of a protein, *J. Mol. Biol.*, Vol. 157, No. 1, pp. 105–132 (online), DOI: 10.1016/0022-2836(82)90515-0 (1982).

[18] M. Heinig, D. Frishman: STRIDE: A web server for secondary structure assignment from known atomic coordinates of proteins, *Nucleic Acids Res.*, Vol. 32, No. WEB SERVER ISS., pp. 500–502 (online), DOI: 10.1093/nar/gkh429 (2004).

[19] K. P. Tan *et al.*: Depth: a web server to compute depth, cavity sizes, detect potential small-molecule ligand-binding cavities and predict the pKa of ionizable residues in proteins., *Nucleic Acids Res.*, Vol. 41, No. Web Server issue, pp. 314–321 (online), DOI: 10.1093/nar/gkt503 (2013).

[20] J. Moult *et al.*: Critical assessment of methods of protein structure prediction-Round VIII, *Proteins Struct. Funct. Bioinforma.*, Vol. 77, No. SUPPL. 9, pp. 1–4 (online), DOI: 10.1002/prot.22406 (2009).

[21] J. Moult *et al.*: Critical assessment of methods of protein

structure prediction (CASP)-round IX, *Proteins Struct. Funct. Bioinforma.*, Vol. 79, No. SUPPL. 10, pp. 1–5 (online), DOI: 10.1002/prot.23200 (2011).

[22] J. Moult *et al.*: Critical assessment of methods of protein structure prediction, *Proteins*, Vol. 82, No. 0 2, pp. 1–6 (online), DOI: 10.1002/prot.24452.Critical (2015).

[23] W. Kabsch: A solution for the best rotation to relate two sets of vectors, *Acta Crystallographica Section A*, Vol. 32, pp. 922–923 (online), DOI: 10.1107/S0567739476001873 (1976).

[24] A. Zemla: LGA: A method for finding 3D similarities in protein structures, *Nucleic Acids Res.*, Vol. 31, No. 13, pp. 3370–3374 (online), DOI: 10.1093/nar/gkg571 (2003).

[25] J. Xu, Y. Zhang: How significant is a protein structure similarity with TM-score = 0.5?, *Bioinformatics*, Vol. 26, No. 7, pp. 889–895 (online), DOI: 10.1093/bioinformatics/btq066 (2010).

[26] F. Chollet: keras, https://github.com/fchollet/keras (2015).

[27] D. P. Kingma, J. Ba: Adam: A Method for Stochastic Optimization, pp. 1–15 (online), DOI: http://doi.acm.org.ezproxy.lib.ucf.edu/10.1145/1830483.1830503 (2014).