Vol.2019-BIO-57 No.3
2019/3/8

グラフ畳み込みを用いた タンパク質予測立体構造の評価手法の開発

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．で立体構造予測では，複数のテンプレート構造や複数の予測手法の組み合わせが用いられることが多く，そこで生成される予測立体構造の評価手法 (Model Quality Assessment Program, MQAP) は有用な技術である．現在最も高精度な MQA 手法の 1 つである ProQ3D は，各残基ごとにクオリティ (局所スコア) し，その平均を取ることでタンパク質全体のクオリティ (大域スコア) を計算する．ここで局所スコアに機械学習アプローチを用いて大域スコアを予測することで精度の向上が期待されるが，タンパク質の残基の数は可変であるため，機械学習を導入することは困難である．この問題を解決するため本研究では，残基をノード，周辺残基との間にエッジとしたグラフ構造定義し，グラフ畳み込みとマルチタスク学習を用いて局所ラベルと大域ラベルを同時に学習する深層学習モデルを開発し，既存手法よりも高精度での予測に成功した．

Protein prediction structure model quality assessment using Graph Convolution

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 in life science application such as drug discovery. Determination of 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. However, the accuracy is still insufficient. Thus, evaluation of the quality of a predicted model is required and such software is called model quality assessment program (MQAP). ProQ3D, which is currently one of the best MQA method, assesses the model quality of each residue (local score) using deep learning method. As results, the quality of whole protein structure (global score) was simply calculated as the mean value of local scores. Thus, if we can use a machine learning method to integrate those local scores for getting a global score, it may improve the accuracy compared with that by the mean value. However, it is difficult to use machine learning method because number of residues in a protein is not fixed. To deal with this problem, we developed a novel single model assessment method using graph convolution. We defined a graph structure on a protein, whose node is a residue and as edge means close residues. By using multi-task learning based on local and global score, proposed method achieved better accuracy than previous methods.

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．2018 年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 15 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

[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

ノ酸配列の数は約 1 億 4 千万件である．

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，比較モデリング法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，天然構造との構造類似性が高い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．このような手法を総称して Model Quality Assessment Program (MQAP)[3] と呼ぶ．

MQAP は single model method[4], [5], [6], [7] と consensus method[8], [9] の 2 つに大きく分けられる．single model method は単一の予測立体構造を入力としてその構造の質を予測する．一方 consensus method は予測立体構造の集合を入力とし，構造の質が良いタンパク質はその他のタンパク質との構造が類似しているという前提に基づき構造の質を出力する．予測コンテストなどにおいては consensus method の方が高精度に予測できるが，質が悪いモデルが多くを占めている場合 single model method の方が精度良く予測することができることが知られている [10]．また精度が良い consensus method は single model method を入力特徴量とする手法が多いため，consensus method の改良のためにも single model method の開発は重要である．

近年最も精度がよい single model method の 1 つである ProQ3D[4] は各残基ごとの局所構造と天然構造の局所での構造類似性を S-score として定義し，これを Deep Neural Network を用いて学習する手法である．

この手法には大きく分けて 2 つの問題が挙げられる．1 つは全結合層により学習している点である．そのため，固定長のベクトルを入力とすることになり，ある一定のウィンドウサイズに区切って入力データを生成するが，これは配列上離れている情報を取り込むことができず，また局所空間を適切に捉えることができないことが考えられる．2 つ目の問題点は局所の良し悪しを表す局所ラベルのみで学習している点である．局所ラベルのみで学習しているため，GDT_TS の様なタンパク質全体での予測の良さを示す大域ラベルを学習することができない．また，大域ラベルを直接予測することができないため，ProQ3D では大域スコアを局所スコアの平均値として定義しているが，局所スコアを大域スコアに統合するのに機械学習を用いることで，単純な平均値よりも高精度な予測が期待される．しかし，タンパク質立体構造に含まれる残基の数は固定ではないため単純に機械学習を導入することは困難である．

近年，グラフ構造に畳み込み演算を定義する深層学習手法であるグラフ畳み込みが特に物性値予測等で数多く研究されている [11], [12], [13]．タンパク質立体構造に対してグラフ畳み込みを用いる手法も考案されつつあり，Fout らはタンパク質のインターフェース予測に各残基をノード，近傍 20 残基間をエッジとするグラフ構造をタンパク質立体構造に定義しグラフ畳み込みを用いた深層学習手法を考案し，従来手法である SVM を用いた手法よりも高精度の予測を達成している [14]．

本研究では，グラフ畳み込みを用いることで，局所ラベルと大域ラベルを同時にマルチタスクで学習することにを可能とし，それにより高精度な single model での MQAP を開発した．構造予測のコンペティションである CASP[15](Critical Assessment of protein Structure Prediction)11, 12 のテストセットを用いた実験では既存手法よりも高精度での予測に成功した．

2. 提案手法

2.1 学習に用いるデータセット

CASP7-10 で用いたれた 438 個の天然構造と，それらに対して参加グループによってモデリングされた平均 274.3 個のデコイ構造からなるデコイセットを学習に用いる．天然構造単位でデータセットを学習データと検証データを 8:2 で分割し，学習 データのデコイ構造を 25%にランダムサンプリングした．また Scwrl4[16] を用いて側鎖を最適化した．

2.2 定義するグラフとグラフ上の畳み込み演算

各残基をノードとし，各残基の CA 間距離が 8Å 以内のノード間にエッジがあると定義した．このグラフ構造に対して先行研究 [14] で用いられたグラフ畳み込み演算の定義のうち以下の 2 種類のグラフ畳み込み演算を用いた．

- NodeAverage
  注目ノードと周辺ノードの重みを分けて足し合わせるモデルである．

$$z_i = \sigma\left(W^{\mathrm{c}} x_i + \frac{1}{|\mathcal{N}_i|}\sum_{j\in\mathcal{N}_i} W^{\mathrm{N}} x_j + b\right)$$

- NodeEdgeAverage
  NodeAverage にエッジの特徴も加えて畳み込みをするモデルである．

$$z_i = \sigma\left(W^{\mathrm{c}} x_i + \frac{1}{|\mathcal{N}_i|}\sum_{j\in\mathcal{N}_i} W^{\mathrm{N}} x_j + \frac{1}{|\mathcal{N}_i|}\sum_{j\in\mathcal{N}_i} W^{\mathrm{E}} A_{ij} + b\right)$$

2.3 入力特徴量

2.3.1 ノードの特徴量

以下の 3 種類の特徴量をあわせて用いる．

- **Base feature** 先行研究 [14] に用いられている Psi-balst[17] から得られる Position-Specific Scoring Matrix(PSSM)，Relative accesible Surface Area (observed RSA)[18]，Half Sphere Exposure[19]，に Secondary Structure (Observed SS)[20] を加えた 26 次元の特徴量
- **Profile based feature**
  ProQ3D[4] や SVMQA[7] に用いられている Psi-blast から得られる PSSM を用いて予測される RSA(predicted RSA)[21] や SS (predicted SS)[21]，また observed SS と predicted SS が一致しているかどうかを表す特徴量を加えた 5 次元の特徴量
- **Rosetta energy feature**
  ProQ3D でも用いられている，生体分子のデザインに用いる統計ポテンシャルと物理的ポテンシャルで構成されるエネルギー関数である Rosetta[22] を用いて計算される各残基毎の様々なエネルギーを表す 20 次元の特徴量

2.3.2 エッジの特徴量

先行研究 [14] を参考に残基間の距離，それぞれの残基のアミド面の法線ベクトルがなす余弦，アミド結合があるかどうかの計 3 次元とした．

2.4 ラベル

本手法ではマルチタスクでの学習を行うため，局所・大域の 2 つのラベルが与えられている．

2.4.1 局所ラベル

局所ラベルとして本研究では以下のラベルを定義する．天然構造とデコイ構造の各残基毎の局所での構造類似性に関して，構造を重ね合わせたとき以下の式により局所ラベルを定義する．

$$\text{local label} = \begin{cases} 1 & \text{if } \frac{1}{4}\left(\sum_{i=-2}^{1} p_i\right) > 0.5 \\ 0 & \text{(otherwise)} \end{cases}$$

(p_i denotes rate of residues under distance cutoff $\leq 2^i \mathring{A}$)

この局所ラベルを用いて二値分類として局所の構造を学習する．

2.4.2 大域ラベル

大域ラベルとして天然構造とデコイ構造の構造全体の GDT_TS を用いて回帰問題として学習する．予測問題としての最終的な正解ラベルはこの大域ラベルである．

2.5 学習モデル

提案手法では局所ラベルと大域ラベルをマルチタスクで学習するために図 1 のようなモデルを用いる．損失関数は局所ラベルとの差を表す local loss と大域ラベルとの差を表す global loss による以下の式で定める単純な和とする．

$$Loss = local\,loss + global\,loss$$

各ノード毎の特徴量を 1 つにまとめ，グラフ全体の出力を得ることを readout と呼ぶが本研究では平均値のベクトルを得る操作を readout function とした．層構造の詳細は図 2 に示す．ただし Graph Convolution は G.Conv と表記する．

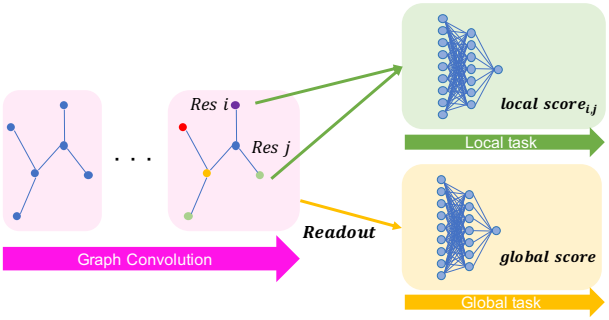

図 1 マルチタスクでの学習の概要

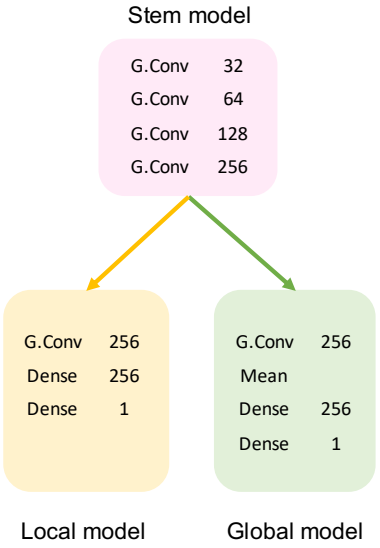

図 2 モデルの詳細

3. 実験結果

3.1 検証データでの性能比較

様々なパラメーターでの精度の比較を行った結果が以下の表である．ここで CC validation は検証データの大域ラベルと大域スコアにおいての，CC validation(local) は局所スコアの平均値と大域ラベルにおいてのピアソンの相関係数を表す．

3.1.1 マルチタスクによる精度比較

マルチタスクにより精度が向上するかを検証した結果が表 1 である．局所ラベルをマルチタスクに学習するほうが

表 1　検証データセットに対するマルチタスク学習による比較

local loss	global loss	CC validation	CC validation(local)
✗	✓	0.887	-
✓	✗	-	0.852
✓	✓	**0.906**	0.842

表 2　使用したノード特徴量による精度への影響

Base Feature	Profile based	Rosetta Energy	CC validation
✓	✗	✗	0.906
✓	✓	✗	0.910
✓	✗	✓	0.909
✓	✓	✓	**0.916**

表 3　グラフ畳み込みモデルによる精度比較

グラフ畳み込みモデル	Val Pearson
Node Average	**0.916**
Node Edge Average	0.912

表 4　テストセットの詳細

Dataset	タンパク質数	平均デコイ数
CASP11 stage2	88	150
CASP12 stage2	51	150

精度が向上することが示された．また既存手法のように局所スコアの単純な平均値を大域スコアとするよりも精度が良い結果となった．

3.1.2　用いるノード特徴量による比較

ノード特徴量に Profile based feature や Rosetta energy feature のようなハイレベルな特徴量を加え精度が向上するかの検証結果が表 2 である．それぞれハイレベルな特徴量を加えることで精度が向上する結果となった．

3.1.3　グラフ畳み込みモデルによる比較

学習に用いるグラフ畳み込み演算の定義によって精度が変わるかを検証した．エッジの特徴量も畳み込みをする Node Edge Average のほうが精度が下がる結果となった．エッジの特徴量は更新されないモデルを用いているため精度が下がったことが考えられる．

3.2　既存手法との比較

3.2.1　テストセット

CASP11, 12 stage 2 のデコイセットを用いて，提案手法の精度を既存手法と比較する．予測立体構造とその GDT_TS は http://predictioncenter.org/から取得した．デコイセットの詳細は表 4 に示す．

3.2.2　比較手法

既存の single model 手法に対して精度を比較する．比較する既存手法として CASP 等で良い成績を残している最新の手法である ProQ3D[4], ProQ3[23], DeepQA[5], VoroMQA[6] を用いる．既存手法はデフォルトパラメータを用いて実行した．

表 5　CASP11 stage2

method	CC	ρ	$Loss$	$p\text{-}value(CC)$	$p\text{-}value(\rho)$
Proposed	**0.572**	**0.529**	**3.830**	-	-
ProQ3D	0.497	0.465	6.278	**3.33E-03**	**8.08E-03**
ProQ3	0.457	0.430	5.460	**8.53E-05**	**3.41E-04**
VoroMQA	0.432	0.415	6.433	**2.62E-06**	**4.17E-05**
DeepQA	0.411	0.396	7.581	**7.19E-07**	**1.21E-05**

表 6　CASP12 stage2

method	CC	ρ	$Loss$	$p\text{-}value(CC)$	$p\text{-}value(\rho)$
Proposed	**0.702**	0.635	5.976	-	-
ProQ3D	0.690	**0.641**	7.503	5.00E-01	7.42E-01
ProQ3	0.636	0.584	**5.435**	**4.55E-04**	**1.31E-02**
VoroMQA	0.593	0.540	7.764	**4.06E-06**	**3.39E-05**
DeepQA	0.572	0.537	7.634	**1.23E-07**	**6.38E-05**

3.2.3　評価指標

それぞれの手法から得られたスコアと GDT_TS から計算されるピアソン (CC), スピアマン (ρ) の相関係数を用いる．また手法により選択されたモデルと最も GDT_TS の大きいモデルの差を $Loss$ とする．

3.2.4　比較結果

以下の表 5, 6 に比較結果を示す．CC, ρ, $Loss$ について最も良いものを太字で示し，提案手法との対応のある t 検定の結果を $p\text{-}value(CC)$, $p\text{-}value(\rho)$ と表し，有意水準 5%で有意なものを太字で示す．CASP11 stage2 において提案手法は既存手法よりも有意に精度が良い結果を示した．CASP12 stage2 においても ProQ3D 以外の手法で有意に精度が良く，また ProQ3D よりも精度が良い，または同等の精度を示した．

4.　結論

4.1　本研究の結論

本研究では既存手法の問題点を解決するためにグラフ畳み込みとマルチタスク学習を組み合わせた新たな Single model method のタンパク質予測立体構造評価手法を開発した．マルチタスク学習が精度を向上させることを確認し，テストセットでの性能比較において CASP11 stage 2 では最も精度が良いとされる既存手法の 1 つである ProQ3D よりも有意に精度よく予測することに成功した (p-value=0.003)．

4.2　今後の課題

今回エッジの特徴量を畳み込み演算に加えることで精度が低下したが，これは Node Edge Average ではエッジの特徴量が更新されないことや，エッジの特徴量が適切でないことが理由として考えられる．エッジの特徴量を更新する畳み込み演算を定義し，またエッジの特徴量に残基対のエネルギーや Profile から予測されるコンタクト予測等

の特徴量を加えることで精度の向上が期待できる．また near-native なモデルを選択するユースケースにおいて順位相関が特に重要であるが，大域ラベルの回帰問題ではなくランク学習にすることで順位相関の指標が向上することが期待できる．

参考文献

[1] S. K. Burley *et al.*: RCSB Protein Data Bank: biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy, *Nucleic Acids Res.*, Vol. 47, No. October 2018, pp. 464–474 (online), DOI: 10.1093/nar/gky1004 (2018).

[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] K. Uziela *et al.*: ProQ3D: Improved model quality assessments using deep learning, *Bioinformatics*, Vol. 33, No. 10, pp. 1578–1580 (online), DOI: 10.1093/bioinformatics/btw819 (2017).

[5] R. Cao *et al.*: DeepQA: Improving the estimation of single protein model quality with deep belief networks, *BMC Bioinformatics*, Vol. 17, No. 1, pp. 1–9 (online), DOI: 10.1186/s12859-016-1405-y (2016).

[6] K. Olechnovič, Č. Venclovas: VoroMQA: Assessment of protein structure quality using interatomic contact areas, *Proteins Struct. Funct. Bioinforma.*, Vol. 85, No. 6, pp. 1131–1145 (online), DOI: 10.1002/prot.25278 (2017).

[7] B. Manavalan, J. Lee: SVMQA: support-vector-machine-based protein single-model quality assessment, *Bioinformatics*, Vol. 33, No. 16, pp. 2496–2503 (online), DOI: 10.1093/bioinformatics/btx222 (2017).

[8] J. Lundström *et al.*: Pcons: a neural-network-based consensus predictor that improves fold recognition., *Protein Sci.*, Vol. 10, No. 11, pp. 2354–62 (online), DOI: 10.1101/ps.08501.are (2001).

[9] M. J. Skwark, A. Elofsson: PconsD: Ultra rapid, accurate model quality assessment for protein structure prediction, *Bioinformatics*, Vol. 29, No. 14, pp. 1817–1818 (online), DOI: 10.1093/bioinformatics/btt272 (2013).

[10] A. Kryshtafovych *et al.*: Methods of model accuracy estimation can help selecting the best models from decoy sets: Assessment of model accuracy estimations in CASP11, *Proteins*, Vol. 84, No. May 2014, pp. 349–369 (online), DOI: 10.1002/prot.24919 (2016).

[11] S. Kearnes *et al.*: Molecular graph convolutions: moving beyond fingerprints, *J. Comput. Aided. Mol. Des.*, Vol. 30, No. 8, pp. 595–608 (online), DOI: 10.1007/s10822-016-9938-8 (2016).

[12] J. Gilmer *et al.*: Neural Message Passing for Quantum Chemistry, *J. Med. Chem.*, Vol. 61, No. 5, pp. 1951–1968 (online), DOI: 10.1021/acs.jmedchem.7b01484 (2017).

[13] K. T. Schütt *et al.*: SchNet - A deep learning architecture for molecules and materials, *J. Chem. Phys.*, Vol. 148, No. 24 (online), DOI: 10.1063/1.5019779 (2018).

[14] A. Fout *et al.*: Protein Interface Prediction using Graph Convolutional Networks, *Adv. Neural Inf. Process. Syst. 30* (I. Guyon *et al.*, eds.), Curran Associates, Inc., pp. 6530–6539 (online), available from ⟨http://papers.nips.cc/paper/7231-protein-interface-prediction-using-graph-convolutional-networks.pdf⟩ (2017).

[15] J. Moult *et al.*: Critical assessment of methods of protein structure prediction (CASP)—Round XII, *Proteins Struct. Funct. Bioinforma.*, Vol. 86, No. S1, pp. 7–15（オンライン），DOI: 10.1002/prot.25415 (2018).

[16] G. G. Krivov, M. V. Shapovalov, R. L. Dunbrack: Improved prediction of protein side-chain conformations with SCWRL4, *Proteins Struct. Funct. Bioinforma.*, Vol. 77, No. 4, pp. 778–795 (online), DOI: 10.1002/prot.22488 (2009).

[17] D. J. Lipman *et al.*: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Res.*, Vol. 25, No. 17, pp. 3389–3402 (online), DOI: 10.1093/nar/25.17.3389 (1997).

[18] S. Mitternacht: FreeSASA: An open source C library for solvent accessible surface area calculations., *F1000Research*, Vol. 5, p. 189 (online), DOI: 10.12688/f1000research.7931.1 (2016).

[19] T. Hamelryck: An amino acid has two sides: A new 2D measure provides a different view of solvent exposure, *Proteins Struct. Funct. Bioinforma.*, Vol. 59, No. 1, pp. 38–48 (online), DOI: 10.1002/prot.20379 (2005).

[20] 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).

[21] C. N. Magnan, 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. 2592–7 (online), DOI: 10.1093/bioinformatics/btu352 (2014).

[22] R. F. Alford *et al.*: The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design, *J. Chem. Theory Comput.*, Vol. 13, No. 6, pp. 3031–3048 (online), DOI: 10.1021/acs.jctc.7b00125 (2017).

[23] K. Uziela *et al.*: ProQ3: Improved model quality assessments using Rosetta energy terms, *Sci. Rep.*, Vol. 6, No. June, pp. 1–10 (online), DOI: 10.1038/srep33509 (2016).