Vol.2018-MPS-118 No.41
Vol.2018-BIO-54 No.41
2018/6/15

3次元畳み込みニューラルネットワークを用いた局所的相互作用に基づくタンパク質複合体予測構造評価

池田 光[1,a)]　石田 貴士[1,b)]

概要：
タンパク質間相互作用 (protein-protein interaction, PPI) の解明は生体現象の理解や創薬につながるが，この PPI を理解する上で，2 つのタンパク質がどのような複合体を形成するかは重要な情報となる．しかし，X 線結晶構造解析やクライオ電子顕微鏡などのような実験的な構造決定手法では莫大な費用と時間を要することが多く，計算機を用いた複合体構造の予測手法が開発されてきた．しかし，ドッキングによる予測は精度が必ずしも高いとは言えず，出力された予測構造を再評価するリランキングシステムが開発されてきた．既存のリランキングシステムの多くは，評価指標として 2 体間におけるエネルギーを用いており，3 次元での立体構造情報を考慮できていないという問題点を抱えている．一方で，近年，画像認識で用いられていた convolutional neural network を 3 次元に拡張した 3D convolutional neural network が生命科学領域でも成果を上げており，その有用性に注目が高まっている．本研究では，3D convolutional neural network 用いて多体で予測構造モデルを評価する手法を開発した．提案手法を用いて，ドッキングツールによって得られたタンパク質複合体予測構造について評価実験を行なったところ，データセットに含まれる 33 個のタンパク質複合体のうち 32 個のタンパク質複合体で既存手法より高い精度で評価を行うことに成功した．

Evaluation of protein complexes structure based on local protein-protein interaction using 3D convolution neural network

Hikaru Ikeda[1,a)]　Takashi Ishida[1,b)]

Abstract: Protein-protein interaction(PPI) leads to understanding of biological phenomena and drug discovery. In order to understand PPI, it is important what type of complex the two proteins form. However, experimental structure determination methods such as X-ray crystal structure analysis and cryo-electron microscopy often require enormous cost and time, and a prediction method of complex structures using computers is developed. However, prediction by docking is not necessarily high in accuracy, and a re-ranking system for reevaluating the output prediction structure has been developed. Many of existing re-ranking systems are used energy between two bodies as evaluation index, and it has the problem of not being able to consider three-dimensional structure information. In recent years, 3D convolutional neural network which extended convolutional neural network which was used in image recognition in three dimensions has also been successful in the life science field, and its usefulness attracts attention. In this research, we developed a method to evaluate predictive structural model with multiple bodies using 3D convolutional neural network. Using the proposed method, we performed an evaluation experiment on the protein complex predicted structure obtained by the docking tool.As a result, I succeeded in evaluating 32 protein complexes out of 33 protein complexes in the data set with higher accuracy than existing method.

[1] 東京工業大学 情報理工学院
Tokyo Institute of Technology, Meguro, Tokyo 101–0062, Japan
[a)] ikeda.h.am@m.titech.ac.co.jp
[b)] ishida@cs.titech.ac.jp

1. 序論

1.1 計算機によるタンパク質複合体構造予測

一般に，タンパク質は単体で生体内における役割を果た

すことは少なく，多くのタンパク質が他のタンパク質と相互作用することによってその機能を果たしている．この相互作用はタンパク質間相互作用と呼ばれ，生命現象において重要な役割を担っている．[1] このタンパク質間相互作用を理解する上で，2 つのタンパク質がどのような複合体を形成しているかは重要な情報となる．タンパク質複合体の構造を実験的に決定する手法としては，X 線結晶構造解析やクライオ電子顕微鏡がある．しかし，これらの手法では時間的・費用的コストがかかる．そこで，計算機によってタンパク質複合体構造を予測することができれば，構造決定の実験を行うべきタンパク質複合体を減らすことができ，コストを削減することが可能であり有用である．

計算機によるタンパク質複合体構造予測手法には，主にタンパク質の立体構造を使用する構造ベースの手法や，タンパク質のアミノ酸配列情報や共進化情報などに基づく手法が挙げられる．例えば，既知の相互作用情報を使用しない構造ベースの手法には ZDOCK[2]，MEGADOCK[3] を用いた手法などがあり，既知の相互作用情報を使用する構造ベースの手法には HOMCOS がある．

また，構造ベースの手法では，2 つのタンパク質のドッキングシミュレーションを行うが，この時ドッキングの仕方としてタンパク質を剛体と捉えてドッキングを行う rigid-body docking と flexible-body docking の 2 種類ある．実行時間の観点から，多くのタンパク質複合体を扱う場合は，一般的に rigid-body docking を主に用いる．

1.2 リランキングシステムと問題点

Rigid-body docking は高速かつある程度の性能を示すが，一方で予測制度が不十分であるという問題を抱えている．多くのドッキングツールでは高速化を測るため，厳密な結合エネルギーなどの計算を行っていない．そのため，ドッキングツールによって出力された予測複合体構造の中には，順位は良いが真の結合エネルギー値が悪く，構造として間違ったものが存在してしまうという問題点がある．

この問題を解決するものとして，リランキングシステムと言うものがある．これは，ドッキングツールによって得られたタンパク質複合体の予測構造に対して，新たな評価指標をもちいて再評価を行い予測構造の順位を改めるシステムである．

ZRANK[4] は Pierce らによって提案された現在最もよく用いられているリランキングシステムである．これらのリランキングシステムの多くは，近接している 2 個の原子について考えられたエネルギー関数を評価手法として用いている．これらの関数は 2 体間を捉えたものであり，重要な情報であるタンパク質複合体の立体構造の特徴を捉えることができない．そのため，天然に近い構造を正しく評価できていないという問題点がある．そこで，本研究では多体で考えることのできる 3D convolutional neural network(3D-CNN) をフレームワークとして用いることを提案する．

1.3 3D convolutional neural network

Neural network の中でも，畳み込み層をもつものを convolutional neural network (CNN) という．CNN はこれまで画像認識において広く成功を収めている．3D convolutional neural network (3D-CNN) とはこの C-NN において畳み込み層を 3 次元に拡張したものである．近年，3D-CNN は，物体認識 [5] や動作認識において成功を収めている．

1.4 研究目的とアプローチ

本研究では．現在のリランキングシステムでの評価が不十分であるという問題を解決するために，3D-CNN をもちいたタンパク質複合体予測構造の評価手法を提案する．

はじめに，ドッキングツールで生成されたタンパク質複合体予測構造の候補に対し，相互作用面上の局所環境を定義する．定義された局所環境を 3D-CNN の入力として学習し，モデルを作成する．次に，作成したモデルを用いて，局所環境の評価を行い，その評価値をもちいてタンパク質複合体予測構造の再評価することで，正解に近い構造をより精度よく特定することを目指す．

1.5 本論文の構成

本論文では，第 2 章において本研究で用いた 3D convolutional neural network ついて述べる．次に，第 3 章にて提案手法を述べ，第 4 章にて評価実験，第 5 章で結論と今後の課題について述べる．

2. 3D convolutional neural network

2.1 Convolutional neural network

Convolution neural network (CNN) は Lecun[7] らによって最初に提案された neural network である．CNN は層間のノードの計算をする点では NN と同様だが，その計算方法において畳み込み層とプーリング層の 2 つの層をもつという点で異なる．

畳み込み層

与えられた入力に対し，フィルタを用いて畳み込み演算を行い情報を抽出する．一般的には複数のフィルタを用いて行い，それぞれに出力値をもちこれらを重ねて活性化関数で繋ぐことでネットワークを構築する．フィルタをもちいることで，点ではなく領域で評価できるため，従来の 1 次元 neural network ではできなかった汎用的な画像の特徴の抽出が可能となった．

プーリング層

畳み込み層の出力に対し，次元の削減を行う．プーリングサイズ内の最大値を取り出す max pooling と平均値を取り出す average pooling の 2 種類のどちらかが

Vol.2018-MPS-118 No.41
Vol.2018-BIO-54 No.41
2018/6/15

主に用いられる．

また，ある層の中のノードのうち幾つかを無効にして学習を行い，次の層では別のノードを無効にして学習を繰り返すドロップアウト法を用いることがある．これにより，ノード間の依存度を小さくし，汎化性能を上げることで過学習を抑える事ができる．

2.2 3D convolutional neural network

3DCNN とは画像認識において広く使われている CNN において畳み込み層を 3 次元に拡張したものである．

Maturana らは物体を voxel と呼ばれる 3 次元の格子状にすることで，3 次元のまま物体の特徴を抽出し，識別を行う物体認識 [5] を行った．Ji らは画像の次元に時間軸を加え，動作認識を行った．これらはいずれも入力から手動によって作成された複雑な特徴に基づいて分類器を構築する従来の手法より成果を上げている．物体認識においては，Maturana らが行ったように物体を voxel に分割し，各 voxel に画像の各ピクセルにおける RGB に相当する channel と呼ばれる特徴量を作成し，入力にもちいるという方法が広く用いられている．

また，J. Jiménez らは 3DCNN を用いてタンパク質の binding site の予測において，従来の幾何学的・化学的・進化的特徴を用いた手法に比べより良い精度を出すことに成功している．さらに，Torng らは 20 種類のアミノ酸の周辺の局所環境を解析しタンパク質構造内の環境と最も適合するアミノ酸の予測を行い，従来の手作業によって得た特徴量を用いたモデルよりも予測精度を 2 倍に向上している．この 2 つの手法も先に述べた voxel を用いている．

以上からもわかるように，タンパク質構造予測などの生命科学領域においても 3D-CNN が有用な技術であると見られている．

3. 提案手法

既存手法は，現状精度が高いとは言えないという問題点がある．この理由として，近接している 2 つの原子について考えられたエネルギーを評価手法として用いていることが挙げられる．多くのポテンシャル関数が 2 体間を捉えており，重要な情報であるタンパク質複合体の立体構造の特徴を捉えることができない．そこで，多体で考えることのできる 3D convolutional neural network(3D-CNN) をフレームワークとして用いたタンパク質複合体予測構造の評価手法を提案する．

3D-CNN の入力には，入力サイズを一様にするため，相互作用面上にある残基を中心とした局所環境を新たに定義して用いた．これらの局所環境を用いて学習し，モデルを作成したのち，作成したモデルを用いて，タンパク質複合体予測構造に含まれる複数の局所環境を評価し，その評価値を総合してタンパク質複合体予測構造のスコアとした．

3.1 予測構造のラベル付け

各予測構造に対して，正解構造との LRMSD(Ligand Root Mean Square deviation) を計算し，Critical Assessment of PRediction of Interactions(CAPRI) におけるタンパク質複合体の acceptable の基準である 10Å 以内 [10] を正例とし，それ以外を負例とした．

3.2 相互作用面の局所環境の定義

本研究では 3DCNN の入力に用いる局所環境を bounding box と呼び，以下のように定義した．ある残基に対して，その残基のもつ C_α 原子 (官能基と隣接した 1 番目の炭素原子) と N 原子の結合 ($\overrightarrow{C_\alpha N}$) および C_α 原子とカルボキシ基の C 原子の結合 ($\overrightarrow{C_\alpha C}$) と 2 つのベクトルの外積 ($\overrightarrow{C_\alpha N} \times \overrightarrow{C_\alpha C}$) によって得られるベクトルの 3 つのベクトルから得られる正規直交基底を用いて局所環境の基底を定める．[9] この基底を用いて，タンパク質複合体中の残基の C_α 原子を中心として 12Å × 12Å × 12Å の立方体を bounding box とする．

また，本研究では相互作用面に注目してタンパク質複合体予測構造の評価を行うため，相互作用面上に存在する bounding box を用いる．そこで，bounding box 内にタンパク質複合体を構成する 2 つのタンパク質両方の原子が含まれているものを相互作用面上の bounding box とした．

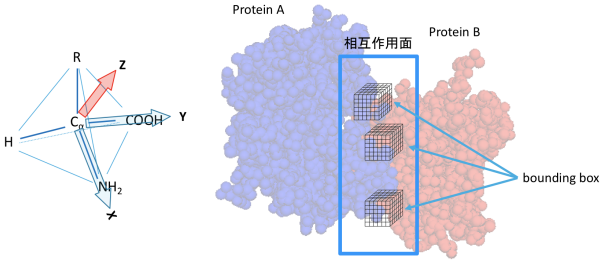


図 1 局所環境の定義

3.3 特徴量

作成した 12Å × 12Å × 12Å の bounding box を更に 2Å × 2Å × 2Å の voxel に分割した．各 voxel において，画像の RGB に相当する channel をその立体格子に含まれる原子の種類 (炭素，窒素，酸素，硫黄の 4 種) とその原子の属するアミノ酸種 (20 種) の計 24channel とした．各 voxel では原子が存在する場合，その原子種とその原子が属するアミノ酸種に対応する channel に 1 を格納し，他の全ての channel に 0 を格納する．ここで，表 1 より，特徴量に含まれる原子の直径はすべて約 2Å 以内のため，各 voxel が単一の原子を収容することが保証されている．

3.4 局所環境のラベル付け

データセット内のタンパク質複合体予測構造それぞれに対し，相互作用面上の bounding box は複数得られる．正

表 1　原子の直径

原子	直径 (Å)
炭素 (C)	1.54
窒素 (N)	1.5
酸素 (O)	1.46
硫黄 (S)	2.04

例のタンパク質複合体の相互作用面は局所的にどこをみても正しい構造を取り，負例のタンパク質複合体の相互作用面は局所的にどこを見ても正しくない構造を取っていると仮定し，各 bounding box のラベルをその bounding box が属するタンパク質複合体のラベルとした．

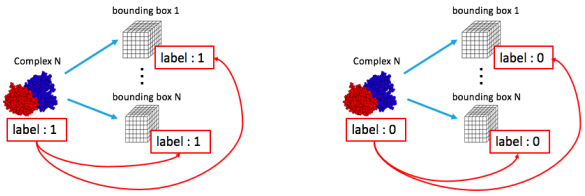


図 2　局所環境のラベル付

3.5　ネットワークの構築

本研究では，keras 2.1.2(backend は Tensorflow 1.4.1) を用いて，3DCNN の実装を行った．

入力は，12Å × 12Å × 12Å を 2Å × 2Å × 2Å に分割し，24channel をもつ 6 × 6 × 6 × 24 の bounding box で，出力は入力の bounding box に対する予測値で 0 ～ 1 の値である．

各層の詳細を以下に示す．

Conv3D 層

入力に対して畳み込み演算を行う．この時，フィルタの数は 16，フィルタサイズは 3 × 3 × 3，活性化関数に Relu 関数を用いた．

Dropout 層

過学習を抑える．Drop 率は 0.2 とした．

Maxpooling3D 層

次元の削減を行う．プーリングサイズは 2 × 2 × 2 とし，プーリング手法は Max pooling を用いた．

Dense 層

全結合を行う．最後の Dense 層以外は活性化関数に Relu 関数を用い，最後の Dense 層では sigmoid 関数を用いることで，出力値をそのまま入力の予測値とした．

3.6　予測構造のスコアの定義

タンパク質複合体予測構造の相互作用面からは複数の bounding box が得られるため，これらの bounding box の予測値を総合的に判断する必要がある．また，タンパク質複合体によって相互作用面の大きさは異なり，得られる bounding box の個数は一様でないので，作成された bounding box の個数も考慮に入れる必要がある．その補正も含めたスコアとして，本研究では，bounding box の予測値の平均を bounding box 数で割ることで新たなスコアとした．

$$\overline{pred} = \frac{\sum_{k=1}^{m} pred_k}{m} \tag{1}$$

$$Score_N = \frac{\overline{pred}}{m} \tag{2}$$

このスコアの高い順にソートすることで，予測構造の新しい順位を出力する．

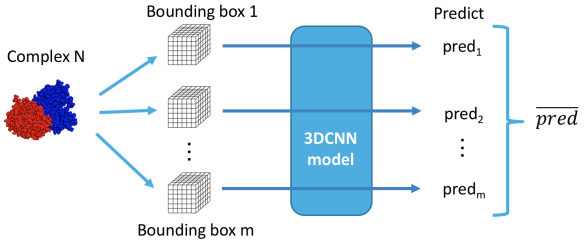


図 3　スコアの定義

4.　評価実験

4.1　データセット

本研究では，タンパク質複合体のデータとして，Vreven らによるタンパク質ドッキングのテストケースである benchmark5.0[11] に含まれる 230 個のタンパク質複合体のうち，残基数が 300 以下かつ第 3 章 1 節におけるラベル付けにおいて正例数が 10 個より多く存在する unbound 構造の 33 個のタンパク質複合体を用いた．このうち，PDBID が CP57 のものについては，条件を満たすが 3P57 から作成されたものであり類似性が高いため今回のデータセットから除いた．

タンパク質タンパク質ドッキングツールは MEGADOCK を用い，回転角を 6°刻み，回転角ごとに　2 つの予測構造を出力することで，各タンパク質複合体に対して，108000 個の予測複合体を作成した．この予測複合体に対して，第 3 章 1 節で示したラベル付を行い，正例の予測構造と top3000 の予測構造をデータセットとした．

データセット中の正例数の平均は 334.3 であり，第 3 章 2 節における作成された bounding box 数の平均は 48.9 だった．

4.2　評価指標

本研究では，精度の評価指標として，AUC(Area under Receiver Operating Characteristic curve) と Success Rate, Success Num を用いた．

AUC(Area under Receiver Operating Characteristic curve)

ROC 曲線の下の面積を表し，2 クラス分類問題におけ

Vol.2018-MPS-118 No.41
Vol.2018-BIO-54 No.41
2018/6/15

る分類器の性能の良さを表し，値が大きいほどよい分類器と評価できる．

Success Rate, Success Num

リランキングシステムは予測順位の高い位置にある正解の構造の個数がリランキング前より後のほうが多いことが望まれる．そこで，予測順位の上位にどれほど正例を集めることができたか表す評価指標が必要となる．そこで，作成した予測構造 108000 個中の正例数にしめる topN 中の正例数の割合を SuccessRate(N)，topN 中の正例数を SuccessNum(N) と定義した．(本研究では，$N = 1, 10, 100, 1000, 3000$ を用いた．)

$$SuccessRate(N) = \frac{\text{top N 中の正例数}}{\text{top 108000 中の正例数}} \times 100 \quad (3)$$

$$SuccessNum(N) = \text{topN 中の正例数} \quad (4)$$

4.3 実験

4.3.1 計算機環境

tsubame3.0 の f ノードを用いて実験を行った．

表 2 f ノードの詳細

CPU	Intel Xeon E5-2680 v4 2.4GHz ×2CPU
コア数	28 コア
メモリ	240GB
GPU	NVIDIA TESLA P100 for NVlink-Optimized Servers Γ4

4.3.2 学習方法と詳細

データセット内の各タンパク質複合体の予測構造において，正例と負例の数が 1 : 1 になるように負例のダウンサンプリングをしたのち，各タンパク質複合体について第 3 章 2,3,4 節で示した方法によって作成された bounding box を入力として学習を行った．その際，ダウンサンプリングを行ったデータセットに対し，タンパク質複合体を 1 つの unit とした 5 fold cross varidation を行った．だだし，test データではダウンサンプリングを行わず，top 3000 の予測構造と top 3000 に含まれない正例と判断された予測構造全てを用いた．

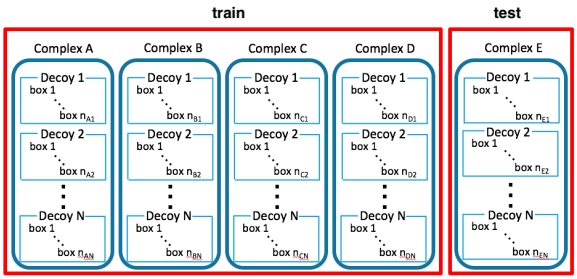


図 4 5 fold cross validation

Neural network を最適化するアルゴリズムには，学習率を 0.001 とした Adam を用いた．epoch 数は 30 を指定し，earlystopping において validation loss を監視対象の値とし，バッチサイズは，タンパク質複合体の予測構造体単位で 250 とした．

4.3.3 学習の経過

1 epoch に約 3 時間所要した．データセットを 5 分割して行った実験の平均 epoch 数は 14 であった．そのため，すべての学習を終えるのに約 3 日要した．学習の経過の様子を以下に示す．

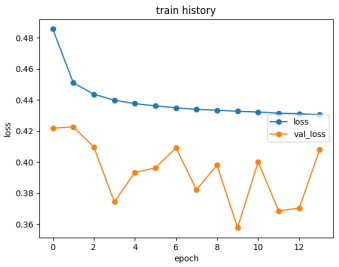


図 5 loss の推移

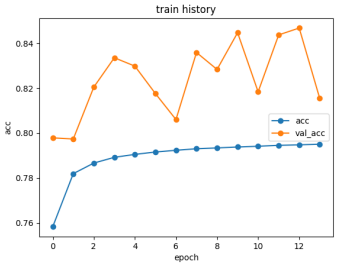


図 6 accuracy の推移

4.4 結果

データセットのタンパク質複合体の AUC の値は，平均 0.75 となった．AUC の最大値は PDBID が 1AY7(RNase Sa / Barstar) と 1KTZ(TGF-β / TGF-β receptor) の 0.98，最小値は 2J0T(MMP1 Intersitial collagenase / Metalloproteinase inhibitor 1) の 0.32 となった．

SuccessRate と SuccessNum の平均は以下の様になった．

N	1	10	100	1000	3000
SuccessRate(%)	0.53	4.74	16.7	61.0	97.2
SuccessNum(個)	0.64	5.45	35.8	193.7	354.0

これらについて，次に既存研究である ZRANK を行った場合との比較を行う．

4.5 比較実験

第 4 章 2 節の評価指標を用いて，既存研究の ZRANK との比較を行った．

4.5.1 精度

テストデータに含まれるタンパク質複合体予測構造に対して，ZRANK により再評価を行った．ZRANK によって付与されたスコアは，小さいほうがより天然構造に近くなるため，−1 をかけることで，天然構造に近いほどスコアが高くなるよう調整した．ZRANK を用いた場合での AUC を計算すると，平均 0.46，分散 0.03 であった．提案手法によってリランキングを行った結果，AUC は平均 0.75，分散 0.02 となったため，既存手法より精度がよくなったと言える．

また，以下の表の様に，平均の SuccessRate，SuccessNum も全てで既存手法より提案手法のほうが高くなり，提案手法の有効性が認められた．

表 3　ZRANK との比較

		SuccessRate(%)				
	AUC	1	10	100	1000	3000
ZRANK	0.46	0.01	0.22	2.66	23.5	87.2
3DCNN	0.75	0.53	4.74	16.7	61.0	97.2

		SuccessNum(個)				
	AUC	1	10	100	1000	3000
ZRANK	0.46	0.09	1.00	10.4	79.6	267.5
3DCNN	0.75	0.64	5.45	35.8	193.7	354.0

4.5.2 実行時間

ZRANK は 54000 個の予測構造に対し，約 3 時間の時間を所要する．[4] つまり，1 つのタンパク質複合体予測構造に対して，0.2 秒かかる．一方，提案手法においては，モデルの学習には 4.3.2 にもあるように 3 日かかってしまう．学習済みのモデルを使用する場合においても，1 つのタンパク質複合体予測構造に対して bounding box を作成することに 2.4 秒，モデルによって再評価を行うことに 0.17 秒，合計で 2.57 秒かかる．従って，提案手法は 1 つのタンパク質複合体予測構造に対して，既存手法より 2.37 秒多く必要となり，計算時間の面で既存手法より劣る結果となった．

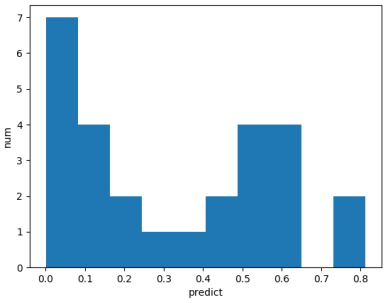
4.6 考察

4.6.1 予測精度

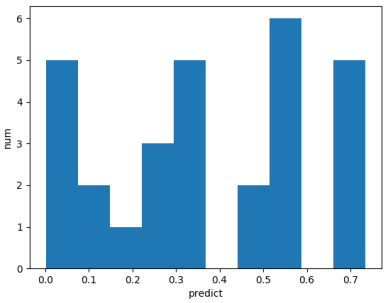
予測構造の評価において，3DCNN を用いた手法は既存手法の ZRANK に比べ，データセット内の 33 タンパク質複合体中 32 のタンパク質複合体で，その予測精度が向上した．この結果により，より天然に近い構造を精度良く評価するためには，既存研究で用いられているような 2 体間のポテンシャル関数より，多体で捉えることのできる 3DCNN の方が有効であることがわかる．

4.6.2 スコアの算出方法

スコアの定義方法について，現在の方法であると，予測値がとても悪い boudning box は数個あるが後はほどほどの予測値である場合に正しくスコアをつけることが出来ない．PDBID:1AY7 の予測複合体の内，MEGADOCK での予測順位が 960 位だった負例の予測構造と 95716 位の正例の予測構造を例に上げる．これらは，提案手法により 10 位，3 位にリランキングされている．この予測構造がもつ bounding box の予測値をヒストグラムにしたものが以下である．



MEGADOCK で 960 位の予測構造



MEGADOCK で 95716 位の予測構造

図 7　予測構造のヒストグラム

これを見ると，95716 位の予測構造は予測値がよい bounding box の方が多く，正しく再評価出来ていると言える．一方で，960 位の予測構造は予測値が最も悪い bounding box が最もよい bounding box の個数より多くある．本来であればスコアは低いはずであるが，全ての bounding box の予測値の平均を取ってしまっているために予測値の低い bounding box の影響を強く反映できず，最終的な評価値として上位になってしまう．そこで，極端な値に対しては閾値を設け，それ以上であれば重み付けをしてその影響を大きくすることでより正確なスコアが計算できると考える．

4.6.3 学習

学習時において，validation accuracy が accuracy より高い値になってしまったため，train データに対して test データが予測し易いデータになっている可能性がある．つまり，train データと test データに似たようなタンパク質複合体がある場合，簡単な問題となってしまう．そこで EMBOSS Needle[13] を使ってデータセット内のタンパク質の配列の類似性を調べが，相同性は見られなかった．従って，データ量の不足，earlystop により学習が必要以上に速

く切り上げられてしまった可能性があるが，これらについては本研究で調査を行うことができなかった．

5. 結論と課題

5.1 結論

本研究では．現在のリランキングシステムでの評価が不十分であるという問題を解決するために，リランキングシステムで用いられるタンパク質複合体予測構造を評価するスコア関数を改良するという目的で研究を行なった．既存研究はスコア関数として 2 つの原子間におけるエネルギーを用いているため，近接している 2 体間しか考えられていないという問題があった．本研究では，周囲の情報を取り込み多体での評価を行うために，画像認識でよく用いられる CNN を 3 次元に拡張した 3DCNN を用いた手法を提案し，精度よくリランキングを行うことが可能であることがと示された．しかし，計算時間の面で既存手法に劣る結果となった．

5.2 今後の課題

本研究ではラベルをつける際に，正解構造との LRMSD を用いたが，タンパク質複合体の予測構造の評価において，相互作用面以外の構造は冗長になりうる．そこで，相互作用面だけに着目した iRMSD[12] を用いることでより正確なラベル付けが可能となる．

また，学習時間と計算機のメモリの関係上，残基数を 300 以下のタンパク質複合体と条件を設けた．しかし，マルチノード・マルチプロセスによる分散学習により，これら 2 つの問題を解決出来るのではないかと考えている．これにより，今回使用した benchmark5.0 の全てのタンパク質複合体において提案手法の精度評価を行うことで，さらなる精度の向上を見込める可能性がある．

さらに，本研究では unbound 構造を用いてドッキングツールによって出力された構造から正例を抽出し，学習をおこなっているため，正例を十分に得られていないという問題がある．これについては，MEGADOCK でドッキングシミュレーションを行う際にブロック機能を用いて，相互作用面から極端に離れた位置でのドッキングをブロックし，より正解に近い構造を取り出すことや，学習の際に正例の bounding box を回転させ，新たな正例を作成することで正例数を増やすことができ，より精度の良い結果が得られると考える．

参考文献

[1] Stelzl, U., Worm, U., Lalowski, M., et al. : A Human Protein-Protein Interaction Network: A Resource for Annotating the Proteome, *Cell*, Vol.122, No.6, pp.957-968(2005).
[2] Chen, R., Li, L., and Weng, Z. : ZDOCK: An initial-stage protein-docking algorithm, *Proteins*, Vol.52, No.1, pp.80-87(2003).
[3] Ohue, M., Shimoda, T., Suzuki, S., et al. : MEGADOCK 4.0: An ultra-high-performance protein-protein docking software for heterogeneous supercomputers, *Bioinformatics*, Vol.30, No.22, pp.3281-3283(2014).
[4] Pierce, B. and Weng, Z. : ZRANK: Reranking protein docking predictions with an optimized energy function, *Proteins*, Vol.67, No.4, pp.1078-1086(2007).
[5] Maturana, D. and Scherer, S. : VoxNet: A 3D Convolutional Neural Network for Real-Time Object Recognition, *Iros*, pp.922-928(2015).
[6] Pierce, B. and Weng, Z. : A combination of rescoring and refinement significantly improves protein docking performance, *Proteins*, Vol.72, No.1, pp.270-279(2008).
[7] LeCun, Y. and Bengio, Y. : Convolutional networks for images, speech, and time series, *The Handbook of Brain Theory and Neural Networks*, Vol.3361, pp.255-258(1995).
[8] Jiménez, J., Doerr, S., Martínez-Rosell, G., et al. : DeepSite: Protein-binding site predictor using 3D-convolutional neural networks, *Bioinformatics*, Vol.33, No.19, pp.3036-3042(2017).
[9] Torng, W. and Altman, R. B. : 3D deep convolutional neural networks for amino acid environment similarity analysis, *BMC Bioinformatics*, Vol.18, No.1, (2017).
[10] Janin, J. : Assessing predictions of protein-protein interaction: The CAPRI experiment, *Proteins*, Vol.14, No.2, pp.278-283(2005).
[11] Vreven, T., Moal, I. H., Vangone, A., et al. : Updates to the Integrated Protein-Protein Interaction Benchmarks: Docking Benchmark Version 5 and Affinity Benchmark Version 2, *Journal of Molecular Biology*, Vol.427, No.19, pp.3031-3041(2015).
[12] Armougom, F., Moretti, S., Keduas, V., et al. : The iRMSD: A local measure of sequence alignment accuracy using structural information, *Bioinformatics*, Vol. 22, (2006).
[13] Needleman, S. B. and Wunsch, C. D. : A general method applicable to the search for similarities in the amino acid sequence of two proteins, *Journal of Molecular Biology*, Vol.48, No.3, pp.443-453(1970).