EX18320 (東京大学情報基盤センター推薦課題)

*福田 宏幸（東京大学大学院 新領域創成科学研究科）*

# Deep Learningを用いたタンパク質のコンタクト残基予測

## Abstract

アミノ酸配列情報のみを使ったタンパク質のコンタクト残基ペア予測は、タンパク質の立体構造予測にとって重要なステップと考えられており、精力的な研究がなされている。近年Pottsモデルの導入などによりコンタクト残基ペア予測は大幅な改善がみられているが、立体構造予測にとっては未だに十分な精度は得られているとはいえず、改良の余地がある。また、既存のコンタクト残基ペア予測手法のほとんどは、類縁配列の多重アライメントから進化過程での残基間の変異の相関を読み取り、予測に利用しているが、多重アライメントが正しいという保証はなく、こちらも多くの研究がなされている。そこで本研究では、深層学習を用いて、多重アライメント中の各配列の重み付けとコンタクト予測を1つのネットワークで同時に学習することで、コンタクト予測に適した多重アライメントの重み付けを学習し、トータルでの精度向上を目指す。深層学習には、Residual Networkを用い層を深く重ねることで精度の向上を実現している。

## Methods

**Dataset:** １）Non-redundantなアミノ酸配列をPISCES cull pdb serverより取得。２）PDBファイルを取得し、コンタクト残基を特定。（ $C_\beta$ 間の距離が8 Å以内の残基をコンタクト残基と定義。Glycineの場合は$C_\alpha$座標を用いた。）３）700残基以上と25残基以上のタンパク質を除く。残った14680個のタンパク質を、11744個（Training）と2936個（Validation）に分割して使用。 ４）多重アライメントは、HHBlitsを使用して計算（E-value was set to 0.001 on the UniProt20_2016 library.）。予測2次構造と露出溶媒面積はScratch-1Dを用いて計算した。Testには、CASP11 (Critical Assessment of Techniques for Protein Structure Prediction) で使用された105種類のドメインを使用。※今回の実験は、CASP13への参加が目的になっており、Test setとのRedundancyを除いていない。

**Model:** 我々の使用したネットワーク構造を図1に示す。ネットワークは、多重アライメントの配列間の重み付けをする部分(A)と、重み付けされた多重アライメントと予測2次構造等から、コンタクト確率を予測する部分(B)から構成される。(A) では、MSAから計算された特徴量（①GAPの割合②クエリ配列との一致率③多重アライメント全体のコンセンサス配列との一致率④配列本数と①～③の平均）をMLPに入力し、それぞれの配列に対して重みを出力する。得られた重み付き多重アライメントから、既存手法と同様に大きさL×Lの441個の共分散行列を計算する。これをCNNの入力とする。(B)では、①CNNの出力 ②MSAから計算されるカラムごとのEntropy、PSSM、カラム間のMutual Information ③予測された2次構造と露出溶媒面積を、60層のResidual Networkに入力し、コンタクト確率を得る。Training時には、計算量を減らすため250残基を超える配列については、ランダムに250残基をクリッピングして使用した。Trainingには、ADAM optimizerを用い、学習率を0.0005とした。過学習を防ぐため、DropuutとL2正則化を用いている。**<u>計算には、東京大学情報基盤センターのReedbush Lを使用。搭載されている4枚のNVIDIA Tesla P100を用い、それぞれのGPUで並列に勾配を計算。CPUが計算された勾配を平均しパラメータを更新している。</u>**

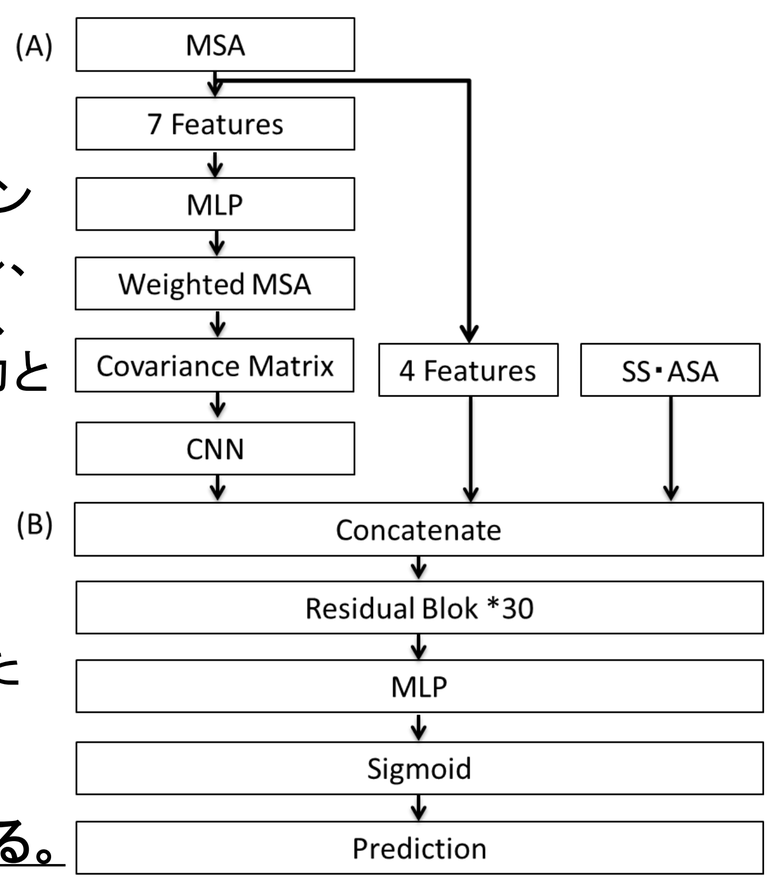


図1

## Results

表1に、CASP11 datasetでの、実験結果を示す。既存手法と比較して、大幅な精度の向上を実現した。

| Method | Short | | | | Medium | | | | Long | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | L/10 | L/5 | L/2 | L | L/10 | L/5 | L/2 | L | L/10 | L/5 | L/2 | L |
| PSICOV | 0.32 | 0.24 | 0.16 | 0.12 | 0.35 | 0.27 | 0.19 | 0.13 | 0.4 | 0.35 | 0.26 | 0.2 |
| CCMpred | 0.36 | 0.28 | 0.18 | 0.13 | 0.41 | 0.32 | 0.22 | 0.15 | 0.45 | 0.41 | 0.32 | 0.24 |
| MetaPSICOV | 0.67 | 0.56 | 0.38 | 0.24 | 0.69 | 0.59 | 0.43 | 0.29 | 0.68 | 0.63 | 0.53 | 0.41 |
| DeepCONV | 0.69 | 0.58 | 0.40 | 0.25 | 0.67 | 0.60 | 0.43 | 0.29 | 0.70 | 0.66 | 0.53 | 0.40 |
| **Our Method** | **0.88** | **0.78** | **0.52** | **0.30** | **0.88** | **0.79** | **0.59** | **0.38** | **0.85** | **0.82** | **0.73** | **0.58** |

また、1epochあたりの計算時間はGeForce NIVDIA TITAN X 1枚での計算 約8時間に比べ、約1時間半に短縮された。

## Conclusions

・コンタクト残基の予測に深層学習を用い、多重アライメントの重み付けを含めてトータルで最適化することで、精度の向上を実現した。

・ 複数のGPUを用いて並列計算することで計算時間が線形的に短縮され、より深いネットワークを構築し、精度の向上に寄与できる。

This research is partially supported by Initiative on Promotion of Supercomputing for Young or Women Researchers, Information Technology Center, The University of Tokyo.