Deep Learningを用いたタンパク質のコンタクト残基予測

福田宏幸
東京大学新領域創成科学研究科 メディカル情報生命専攻

１．背景

これまで得られているタンパク質のアミノ酸配列の数に比べて，実験的に得られている立体構造の数は極端に少なく，そのギャップを埋める為に，計算機によるタンパク質のアミノ酸配列情報のみを使ったタンパク質の立体構造予測は，構造生物学にとって重要な課題である。その中で，タンパク質のコンタクト残基ペア予測は，それが制約条件になり得る為，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（Multiple Sequence Alignment：MSA）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，こちらも精度向上の為に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。

２．方法

２－１．データセット：１）Non-redundantなアミノ酸配列をPISCES cull pdb serverより取得。 ２） タンパク質の構造が座標として記録されたPDBファイルを取得し，コンタクト残基を特定。 （ Cβ 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Cα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）に分割して使用した。 ４）多重アライメントは，HHBlitsを使用し，E-value 0.001の条件で作成した。アライメント構築の為のデータベースは，UniProt20_2016 libraryを使用した。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した。

２－２．モデル：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，多重アライメントから計算された特徴量（①アライメント中のGAP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Multi Layer Perceptron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。これをサイズL×Lピクセル，441チャンネルの画像と考え，続くフィルターサイズ1×1のCNN

に入力し,チャンネルの次元を 441 から 128 に次元圧縮する。(B)では,①先程の CNN の出力 ②多重アライメント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，ランダムに連続する 250 残基を選んで使用した。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。**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。

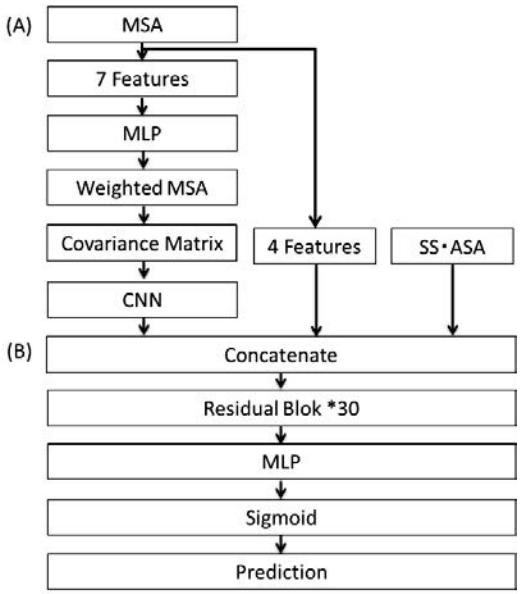

図１：今回我々が使用した Deep Learning のネットワーク。

３．結果

表 1 に，CASP11 dataset での，実験結果を示す。多重アライメントからコンタクト残基を予測する代表的な手法である，PSICOV，CCMpred や，近年別のチームで開発された Deep Learning を用いた手法との比較を行った。結果，我々の手法がいずれの領域のコンタクト予測においても，大幅な精度の向上を実現した。

表１ 実験結果

２つのコンタクト残基ペア間の距離（間にある残基数）によってShort（６残基から１３残基），Medium（１４残基から２３残基），Long（２４残基以上）に分けて集計している。予測精度は，残基長をＬとして，上位L/10，L/5，L/2，Ｌ個の残基ペアについてコンタクトを予測した時の正解率を精度としている。

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
DeepCov	0.69	0.58	0.4	0.25	0.67	0.6	0.43	0.29	0.7	0.66	0.53	0.4
OurMethod	**0.86**	**0.74**	**0.49**	**0.2**	**0.85**	**0.76**	**0.56**	**0.36**	**0.83**	**0.78**	**0.68**	**0.53**

また，Reedbush Ｌの４枚のGPUを並列に使用することで，従来の約４倍の計算速度を実現することが出来た。

４．結論

・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，予測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た。

謝辞

本研究は，東京大学情報基盤センターの若手・女性利用に採択され，同センターのReedbush Ｌを使用して行われたものです。

参考文献

Jones DT, Buchan DW, Cozzetto D, Pontil M. PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. Bioinformatics. 2012 Jan 15;28(2):184-90. doi: 10.1093/bioinformatics/btr638. Epub 2011 Nov 17.

Seemayer S, Gruber M, Söding J. CCMpred -- fast and precise prediction of protein residue-residue contacts from correlated mutations. Bioinformatics. 2014 Nov 1;30(21):3128-30. doi: 10.1093/bioinformatics/btu500. Epub 2014 Jul 26.

G. Wang and R. L. Dunbrack, Jr. PISCES: a protein sequence culling server. Bioinformatics, 19:1589-1591, 2003.

Remmert M, Biegert A, Hauser A, SoÈding J. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nat Methods. 2011 Dec

Monastyrskyy B, D'Andrea D, Fidelis K, Tramontano A, Kryshtafovych A. New encouraging developments in contact prediction: Assessment of the CASP11 results. Proteins. 2016

Sep;84 Suppl 1:131-44. doi: 10.1002/prot.24943. Epub 2015 Nov 17.

Magnan CN, Baldi P. SSpro/ACCpro: almost perfect prediction of protein secondary structure and relative solvent accessibility using profiles, machine learning and structural similarity. Bioinformatics. 2014 Sep 15;30(18):2592-7. doi: 10.1093/bioinformatics/btu352. Epub 2014 May 24.