| L1111-08 | ニューラルネットワークを用いたタンパク質ドメインリンカー予測器の構築 | | | | | |
|---|---|---|---|---|---|---|
| | 氏名 | 松沢佑紀 | 主査 | 黒田 | 副査 | 長澤・中澤・太田・篠原 |

[ 背景・目的 ] タンパク質にはマルチドメインタンパク質と呼ばれる複数の構造的に独立した構造ドメインからなるタンパク質が存在する。マルチドメインタンパク質は巨大なタンパク質であり、一般的に発現や結晶化が困難であることが知られているため、タンパク質を個々の構造ドメインに分割することで、より容易に解析を行うことが有効とされている。そのため、ドメイン境界であるリンカー部位をアミノ酸配列から特定する方法が研究されてきた。現在、タンパク質のドメインリンカーの同定ツールは複数のものが知られており、当研究室においても、ドメインリンカー予測器を開発してきた。しかし、これらの予測機の予測精度はあまり高くなく、マルチドメインタンパク質に限定しても精度は 30%を切っている。そこで本研究ではニューラルネットワーク（以下、NN）の一種であるリカレントニューラルネットワーク(RNN)を用いてドメインリンカー予測器の予測精度の向上を目的とする。

[ 手法 ] SCOPe、CATH の 2 つのドメインデータベースを利用して、ドメインのデータを取得した。これらのタンパク質ドメインのデータについて、代表化と目視による確認を行い、マルチドメインタンパク質を 3529 配列とシングルドメインタンパク質を 4131 配列のデータセットを構築した。次に、これらのデータセットの各タンパク質の残基について特徴量を作成した。特徴量にはアミノ酸残基、PSSM(Position Specific Score Matrix)、MSA(Multiple Sequence Alignment)から作成した 3 種、44 次元の特徴を用いた。予測器の学習では学習データセット中のタンパク質について、各残基がドメイン、リンカーのどちらに属するかを学習させ、リンカー残基の誤検出率を最小化した。最後に、構築した予測器を用いてテストを行った。評価方法はテスト用に用意したタンパク質のすべての残基に対して、予測器によりリンカーかドメインか予測し、最も確率の高い 1 残基が閾値である 0.5 をこえていればこれをリンカーと判定した。そして、この残基が実際のリンカー部位の±5 残基以内に含まれていれば予測を成功とした。正答率を示す *Accuracy*、予測器の精度を示す *Precision*、感度を示す *Sensitivity*、それらを総合して評価するための *F1 score* を算出した。テストはマルチドメインタンパク質 300 配列シングルドメインタンパク質 300 配列を用いて評価した。同様の方法で既存のドメインリンカー予測器でも予測を行い性能の比較を行った。

[ 結果および考察 ] 構築した予測器を用いてテストを行ったところ、*Precision* が 0.46、*Sensitivity* が 0.21 という値が得られた。これは *Precision* については他の予測器と比較し最も高い値であり、誤検出が他の予測器よりも少ないといえる。シングルドメインタンパク質についても、0.81 の高い *Accuracy* を得ることができた。この予測精度の向上の要因としては、シングルドメインタンパク質を加えることによるデータセットの偏りの軽減と、RNN,MSA により、非局所的な特徴を捉えることができたことが考えられる。

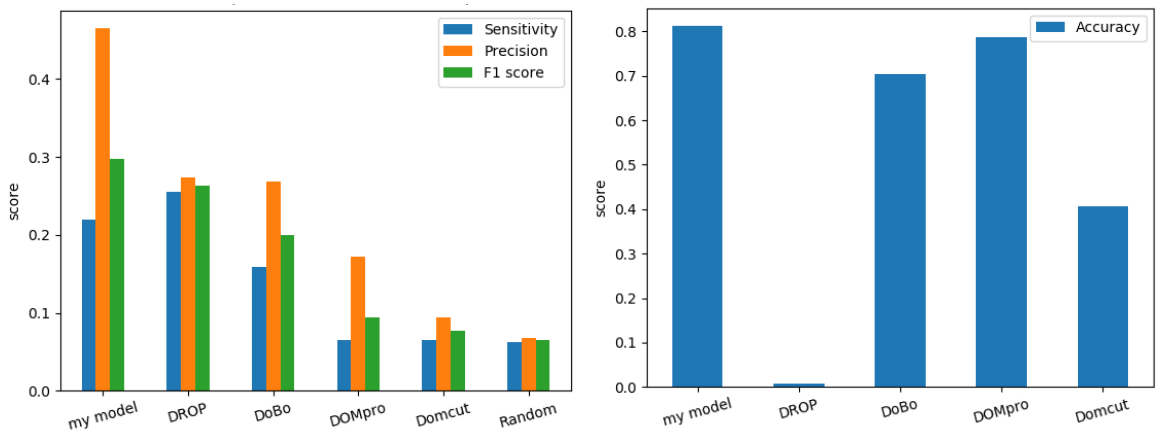


**図 1. 構築した予測器の性能比較 左:マルチドメインタンパク質、右：シングルドメインタンパク質**