平成 ３１年 ３月 ２０日

平成 30 年度研究助成実績報告書

一般財団法人日本産業科学研究所

宮地 尚 理事長 殿

下記の通り一般財団法人日本産業科学研究所研究助成金の研究実績を報告致します。

申請者氏名：清水　謙多郎

所属：国立大学法人東京大学大学院農学生命科学研究科 応用生命工学専攻

研究題目 深層学習を用いたタンパク質相互作用予測システムの開発

Development of a protein interaction prediction system by using deep learning algorithm

研究内容

タンパク質と他の分子との相互作用の解析は、薬剤、有用酵素、機能性食品などの開発において、基盤となる非常に重要な技術である。これらの開発の最初の段階では、特定のタンパク質に結合する分子、あるいは、特定の分子に結合するタンパク質の探索がよく行われるが、こうしたスクリーニングを行うには、生化学実験や構造解析は多くの時間と労力を必要とするため、生物情報科学による予測が期待されている。本研究では、タンパク質のアミノ酸配列を入力として、タンパク質と脂質, DNA, RNA との相互作用部位を予測する手法を開発した。これらの予測では、ホモロジー検索を用いた方法と機械学習を用いた方法の双方を利用する。ホモロジー検索を用いた方法では、ターゲットのタンパク質の配列をクエリとして、BLAST 検索を行い、得られた配列が相互作用部位を含んでいるとき、その部位を相互作用部位と予測する方式を用いている。

タンパク質－脂質相互作用部位予測では、脂質リガンドを、fatty acyls, polyketide, prenol, sterol の 4 種の脂質リガンドクラスごとに予測器を作成し、脂質結合残基の特徴を個別に学習させることによって、予測性能の向上を図った。機械学習の手法としては、support vector machine（SVM）を適用し、結合残基およびその周辺残基の position specific matrix（PSSM） を特徴量として予測を行った。機械学習を用いた手法の fatty acyls, polyketide, prenol, sterol の 4 種の脂質リガンドクラスの AUC 値は、それぞれ 0.773, 0.790, 0.835, 0.759 で、脂質全体の AUC 値は 0.793 であった。prenol 以外は、クラスごとの特徴を学習させる効果よりも、脂質全体の多数の配列の特徴を学習させることで予測精度

が向上したものと考えられる。ホモロジー検索を用いた方法では、E-value のしきい値を 10^{-4} としたとき、accuracy として 0.945 という予測値を得ることができた。

タンパク質－DNA およびタンパク質－RNA 相互作用部位予測についても、ホモロジー検索を用いた方法と機械学習を用いた方法を実現した。機械学習は SVM を用い、配列特徴としては、PSSM、アミノ酸種、アミノ酸の側鎖のタイプによる 10 分類を採用した。それぞれの予測性能を表 1 および表 2 に示す。

表 1　タンパク質－DNA 相互作用部位予測の性能

方法	Accuracy	Sensitivity	Specificity
ホモロジー検索を用いた方法	0.926	0.192	0.990
機械学習を用いた方法	0.757	0.695	0.762

表 2　タンパク質－RNA 相互作用部位予測の性能

方法	Accuracy	Sensitivity	Specificity
ホモロジー検索を用いた方法	0.928	0.618	0.948
機械学習を用いた方法	0.810	0.429	0.835

また、タンパク質－モノヌクレオチドの相互作用部位予測では、相互作用部位だけでなく、相互作用部位の場合、15 種類のモノヌクレオチド AMP, ADP, ATP, CMP, CDP, CTP, GMP, GDP, GTP, TMP, TDP, TTP, UMP, UDP, UTP のどれと相互作用するかを予測する手法を開発した。深層学習 convolutional neural network（CNN）と feature embedding の手法を用い、非結合部位を含めた 16 種類の残基の分類を行った。表 3 に予測性能を示す。予測対象をアデニンヌクレオチド AMP, ADP, ATP に限って予測した場合、全体で予測するよりも高い予測精度を達成しているが、これは、アデニンヌクレオチドの結合部位の特徴をより効果的に学習できていると考えられる。

表 3　タンパク質－モノヌクレオチド相互作用部位予測（分類）の性能

予測対象	Accuracy	Precision	Recall
モノニクレオチド結合タンパク質全体を対象としたとき	0.733	0.681	0.889
AMP, ADP, ATP 結合タンパク質全体を対象としたとき	0.901	0.542	0.798

そのほか、天然変性領域のアミノ酸配列から、他分子との相互作用により、disorder 状態から order 状態に遷移する領域 MoRFs（molecular recognition features）を同定するシステムを開発した。深層学習 CNN を利用し、PSSM と 13 種の AAIndex（疎水性、二次構造の出現傾向、溶媒和自由エネルギーなど）を特徴量とする。その結果、AUC 値 0.778 という高い予測性能を達成した。

今年度 2018 年 10 月（採択）～2019 年 3 月の実績は上記の通りであるが、今後、さらに成果を論文発表する予定である。

研究業績著

(1) Chun Fang, Yoshitaka Moriwaki, Caihong Li, and Kentaro Shimizu: Prediction of MoRFs Based on n-gram Convolutional Neural Network, 2019 7th International Conference on Bioinformatics and Computational Biology, accepted.

(2) Chun Fang, Yoshitaka Moriwaki, Aikui Tian, Caihong Li, and Kentaro Shimizu: Identifying short disorder-to-order binding regions in disordered proteins with a deep convolutional neural network method, Journal of Bioinformatics and Computational Biology, 10, 1142 (2018).