

早稲田大学大学院情報生産システム研究科

博士論文審査結果報告書

論 文 題 目

**Study on Deep Transfer Learning Methods for
the Predictions of Protein Functions**

申 請 者

Xin YUAN

情報生産システム工学専攻
ニューロコンピューティング研究

2022 年 6 月

タンパク質は生物全ての生命現象を担う基本的な物質であり、タンパク質の機能予測は、多くの生物学的研究にとって非常に重要である。従来、タンパク質の機能予測は全て実験的な手法で行われていたが、次世代シーケンサー等の DNA 配列解析技術が向上し、多様な生物種のゲノム配列が解読されることにより、配列データの解析に基づいてタンパク質の機能を予測する手法が近年注目されている。配列データを解析してどこにどのような遺伝子があるかを注釈づけることがアノテーションである。アノテーションには、構造アノテーションと機能アノテーションがある。構造アノテーションとはゲノム配列から遺伝子として機能する配列部分を見つけ出すことであり、機能アノテーションとは見つけ出されたタンパク質配列がどのような機能を持っているのかを予測することである。本研究では、機能アノテーションとして、タンパク質配列からの機械学習法による Gene Ontology (GO)アノテーション等の予測を行っている。

機械学習による機能アノテーションは、タンパク質配列から直接的に機能を予測できるため近年多くの注目を集めており、様々な成果を挙げているが、多くの課題も残されている。GO で定義されている用語 (GO term) の数が 4 万個以上あるため、GO アノテーションの予測は非常に複雑なマルチラベル分類問題となっている。そのため、深層ニューラルネットワークのような複雑なモデルが必要となり、大量の学習データが必要になる。しかし、学習データとして利用する実験によるアノテーションデータは一部のタイプ種に限定され、多くの生物種では学習データの数が非常に少ない。さらに、バイオインフォマティクスのタスクでは、過学習になりやすいため転移学習法が適用されにくいという問題がある。そこで本研究では、まず、複数の予測タスクを一つのモデルで実装できる深層 Multi-Head Multi-End (MHME) Convolutional Neural Network (CNN)モデルを提案する。提案する深層 MHME CNN モデルでは、関連する複数の予測タスクに対して深層 CNN 特徴抽出器を共有することで、各タスクの限られた学習データを有効活用するだけでなく、それぞれの共通特徴を抽出し、転移学習の実施を容易にする。次に、提案した深層 MHME CNN モデルを適用し異なる生物種および異なるタスクの転移学習により、高性能な GO アノテーション予測と、タンパク質細胞内局在予測、およびタンパク質相互作用 (PPI: Protein-Protein Interaction) 予測という 3 つのタンパク質の機能アノテーション予測タスクを実現している。

以下に、本論文の構成と各章の概略について述べ、評価を与える。

第 1 章では、タンパク質とタンパク質の機能を紹介し、本研究の 3 つのタンパク質の機能アノテーション予測タスクの背景・目的を明らかにしている。

第 2 章では、GO アノテーション予測のための階層型深層 MHME CNN モデルの構築について述べている。まず、莫大な数かつお互いに関連するラベルに対処するため、複数のレベルを持つ階層的に組織化された局所分類器の集合からなる階層型マルチラベル GO アノテーション予測器を構築する。次に、一つのモデルで全ての局所分類器が実装できる深層 MHME CNN モデルを提案し、深層 CNN 特徴抽出器を各局所分類器で共有することで、各レベルに限られた

学習サンプルを有効活用することで強力なマルチラベル局所分類器を構築する。提案した深層 MHME CNN モデルは、特徴抽出器として機能する局所分類器間に共有される深層 CNN モデルの本体部分と、各々局所分類器に特徴抽出器を共有するための特徴融合変換を行うオートエンコーダ集合のマルチエンド部分、および線形マルチラベル分類器集合のマルチヘッド部分の 3 つの部分から構成されている。このように、深層 MHME CNN モデルでは、限られた学習データで各レベルの強力なマルチラベル局所分類器を構築して高い分類性能を実現している。性能評価のための数値例では、提案した深層 MHME CNN モデルを複数のベンチマークデータセットに適用して、最先端のアノテーション予測器 DeepGO (M. Kulmanov et al. 2017)、HMC-LMLP (R.Cerri et al. 2016) と比較し、総合評価 F-score では、全てのデータセットにおいて提案した深層 MHME CNN モデルが従来法より優れていることが示されている。例えば、生物過程 “BP” というデータセットでは、DeepGO と HMC-LMLP はそれぞれ 0.3600 と 0.4921 であるのに対して、提案法は 0.6317 である。また、提案 GO アノテーション予測器を再学習なしで異なる生物種に適用した場合も良い予測結果が得られており、優れた転移学習能力があることも確認されている。よって、提案した深層 MHME CNN モデルの有効性が明らかとなっている。

第 3 章では、GO アノテーションからの転移学習による深層タンパク質細胞内局在予測器の構築について述べている。GO アノテーションは細胞内局在の予測に有用であることが知られている。しかし、多くの生物種では、実験による GO アノテーションデータが十分に入手できない。本研究では、タイプ種において GO アノテーションの深層学習を行い、他の生物種の細胞内局在予測に転移をさせることで高性能なタンパク質細胞内局在予測器を構築する。まず、深層 MHME CNN モデルを用いて、深層 CNN 特徴抽出器を深層 GO アノテーション予測器で共有して事前学習を行う。それから、事前学習させた深層 CNN 特徴抽出器を細胞内局在予測器に転移させ、タンパク質細胞内局在サンプルを用いて微調整を行う。このようにして、GO アノテーションの転移学習によって強化された高性能な深層タンパク質細胞内局在予測器を実現している。性能評価のための数値例では、提案した深層タンパク質細胞内局在予測器を Swiss-Prot の 4 個のベンチマークデータセットに適用して転移学習がある場合とない場合とを比較した結果、総合評価 F-score の平均値では、転移学習がない場合は 0.4969 であるのに対して、転移学習がある場合は 0.5271 である。また、最先端の従来法 Hum-mPLoc3.0 (H.Zhou et al. 2016) と比較した結果では、Hum-mPLoc3.0 は 0.5900 であるのに対して、提案法は 0.6207 である。これらのことから、提案した深層タンパク質細胞内局在予測器の有効性が明らかとなっている。

第 4 章では、タンパク質複合体検出のための転移学習に基づく深層 PPI 予測器の構築およびそれをベースにした PPI ネットワークの再構築やタンパク質複合体の検出について述べている。タンパク質複合体の検出では、PPI ネットワークの完全性が重要である。しかし、実験によって検出された PPI は非常に限

られているため、ほとんどの生物種では完全な PPI ネットワークが利用できないという問題点がある。本研究では、深層学習に基づいた PPI 予測器を構築して、未知の PPI を予測し完全な PPI ネットワークを再構築する。再構築された PPI ネットワークに対してスペクトルクラスタリング法を用いてタンパク質複合体の検出を行う。そのために、GO アノテーションの類似性がタンパク質相互作用に寄与し、細胞内局在の相違が負の相互作用に寄与するという事実を考慮し、深層 MHME CNN モデルを用いて、タイプ種のデータセットを用いて GO アノテーションおよび細胞内局在予測のタスクで深層 CNN 特徴抽出器を事前学習し、それを PPI 予測モデルに転移させ微調整を行う。このように、GO アノテーションおよび細胞内局在予測の転移学習で強化された高性能な深層 PPI 予測器を構築し PPI ネットワークの再構築を実現している。性能評価のための数値例では、まず、提案した深層 PPI 予測器を DIP というベンチマークデータセットに適用して PPI 予測の検証を行っている。そして、2 個のベンチマークデータセット (MIPS と CYC2008) に適用して、タンパク質複合体検出の検証を行っている。PPI 予測の検証では、提案した深層 PPI 予測器を 3 つの従来法 (Zeng et al. 2016)、(Sun et al. 2017)、(Bandyopadhyay et al. 2016) と比較した結果では、従来法の検出精度はそれぞれ 0.9594, 0.9377, 0.8700 であり、提案法は 0.9762 であった。また、タンパク質複合体検出の検証では、提案した PPI 予測器に基づいた手法を 3 つの最先端の従来法 MCODE (G.D. Bader et al. 2003), CFinder (B. Adamcsek et al. 2006), PCD-Bens (Z. Rehman et al. 2018) と比較し、総合評価 F-score の平均値では、従来法はそれぞれ 0.3496, 0.3168, 0.4218 であるのに対して、提案法は 0.6442 であった。提案した深層 PPI 予測器の有効性が明らかとなっている。

第 5 章では結論として、本研究の成果について総括し、今後の研究課題を論じている。

以上を要約すると、本論文では、複数の予測タスクを一つのモデルで実装できる深層 MHME CNN モデルを提案している。複数の予測タスクにおいて深層 CNN 特徴抽出器を共有して共通の特徴を抽出し活用することで、異なる生物種および異なるタスクからの転移学習に基づく GO アノテーション等の 3 つのタンパク質の機能予測を行っている。また、提案法を複数のベンチマークデータセットに適用しその有効性を明らかにしている。これらの成果は機械学習、バイオインフォマティクス分野の発展に寄与するところ大である。よって、本論文は、博士（工学）の学位論文として価値あるものと認める。

2022 年 5 月 18 日

審査員

主査	早稲田大学	教授	博士(情報工学)(九州工業大学)	古月 敬之
副査	早稲田大学	教授	工学博士(早稲田大学)	吉江 修
	早稲田大学	教授	博士(工学)(早稲田大学)	藤村 茂