

AlphaFold2 特徴量による生物学的配列予測
Improving Biological Sequence Prediction with Alphafold2
Representation

京都大学 化学研究所 バイオインフォマティクスセンター 生命知識工学 Jiang Zhiqian

研究成果概要

生物学的配列から機能部位を正確に予測することは、生命活動や疾患のメカニズムを理解する上で不可欠である。しかしながら、高精度な予測に必要とされる複雑な構造的コンテキストを捉えることは、従来の配列ベースの手法にとって依然として課題である。AlphaFold2 (AF2) の登場は、この分野にパラダイムシフトをもたらした。構造予測における AF2 の卓越した性能に着目し、本研究では、より広範な下流タスク (downstream tasks) の課題に対処するために、その Evoformer 層から抽出された中間表現 (intermediate representation) を活用する可能性を探究する。具体的には、このモジュールは複雑な相互作用パターンを潜在的な単一表現 (single representation) へとエンコードする。このベクトルはコンパクトな情報キャリアとして機能し、進化の歴史と暗黙的な構造的セマンティクスの両方を捉えることで、人手で設計された (handcrafted) 配列特徴量よりも豊富なデータソースを提供する。

本研究では、AlphaFold2 (AF2) から得られる高次元の単一表現を、生物学的配列予測タスクに統合することの有効性を体系的に評価する。事前学習済みの AF2 モデルを堅牢な特徴抽出器として扱い、これらの潜在的な構造埋め込み (structural embeddings) を抽出して表現を取得し、それをサポートベクターマシン (SVM)、決定木 (DT)、ランダムフォレスト (RF) といった機械学習アルゴリズムから、畳み込みニューラルネットワーク (CNN)、多層パーセプトロン (MLP)、グラフニューラルネットワーク (GNN) を含む深層学習アーキテクチャに至るまで、様々な予測モデルに組み込んだ。我々は、カルパイン (Calpain) 切断部位予測および核酸結合部位予測という下流タスクを通じて、AF2 表現の性能を評価した。その結果、AF2 表現は、位置特異的スコア行列 (PSSM) や隠れマルコフモデル (HMM) プロファイルなどの従来の特徴量を常に上回るか、あるいは大幅に強化することを示し、カルパイン予測においては AUC 0.891 を達成し、グラフベースのモデルにおいても結合残基を効果的に識別した。さらに、SHapley Additive exPlanations (SHAP) および Uniform Manifold Approximation and Projection (UMAP) を用いた解釈可能性分析により、この性能向上が、単に局所的な配列モチーフに依存するのではなく、暗黙的な構造コンテキストを利用するモデルの能力に由来することが明らかになった。これらの知見は、AF2 表現が、幅広いバイオインフォマティクス応用に向けて配列データと構造的洞察の間のギャップを埋める、特徴的かつ有益な特徴量として機能することを裏付けるものである。