1ZM-01

タンパク質立体構造の木表現とGCNに基づくEC番号の推定

青柳 詠美†　　小島　正樹‡

東京薬科大学院　生命科学研究科†　　東京薬科大学　生命科学部‡

1. はじめに

タンパク質の立体構造は、その生理機能と密接な関係がある。近年ビッグデータを用いたタンパク質の立体構造の予測が成功しつつあるが、分子機能や生理作用との相関については謎の点も多い。私たちは、タンパク質分子の位相幾何学的特徴に注目して、立体構造まで含めてグラフ構造で表現し、構造予測から進化計算、創薬探求まで行うVOLTES（Virtual Optimization of Local Tertiary Structures）プログラムパッケージを開発した。VOLTESでは、タンパク質の立体構造を系の自由度に等しい二面角座標を用いて、分子のトポロジーを反映した木構造で表現する。さらに二面角座標データを6値化することにより、探索対象の構造空間を離散化している。今回、VOLTESの機械学習への応用を目指して、まずはAIで有用な特徴量となるのかを調べるため、Pytorch Geometric[1]を用いて立体構造情報のみからEC番号の推定を行なった。

2. VOLTESの開発背景

VOLTESでは、タンパク質分子のトポロジーは、グラフ理論の「木」として表現できる（Abe et al., 1983）ことを利用し、二面角座標を、N末端から順に「木」の形に並べることで「木」構造のデータを計算機上でリストの入れ子として実装した。ここで言う二面角とは、任意の単結合BCのまわりの二面角を平面ABCの法線ベクトルと、平面BCDの法線ベクトルのなす角として計算するものを指す。二面角の値には、6種類の配座異性体が存在し、VOLTESでは巡回6進数（0～5の数字）で表すものとする。以降、各巡回6進数を該当する二面角のVOLTES座標と呼ぶ。本研究室ではこれまでVOLTESをリストの入れ子として実装したアルゴリズムを用いてタンパク質の論理的構造設計(2019)や分子進化とトポロジーとの相関解析（2020）に用いてきた。

Tree representation of protein conformation and estimation of EC number based on GCN

† Eimi AOYAGI, Graduate School of Life Sciences, Tokyo University of Pharmacy and Life Sciences

‡ Masaki KOJIMA, School of Life Sciences, Tokyo University of Pharmacy and Life Sciences

3. 実験方法

pdbj[2]にBioUnitとしてエントリーされているタンパク質[3]のうち、一本のペプチド鎖で構成されている酵素タンパク質6107個を対象に、VOLTESのgraph構造に変換されたタンパク質データからEC番号を推定するGCN(Graph Convolutional Network:グラフ畳み込みニューラルネットワーク)を作成した。グラフの構造はN末端からC末端に向かって結合間にノード番号を振った有向グラフである。今回使用した酵素群のうち各EC番号にエントリーされている酵素の数は表1の通りである。

EC番号	酵素数[個]
2	2879
3	2184
1	433
4	349
5	201
6	61

表 1 本実験での対象酵素数

GCNネットワークにGCNConv[4]を使用した。上述の通り、EC番号に酵素をわけると、個数に大きな差があり、一つのモデルで処理を行うと、下図のように困難である事が判明した。

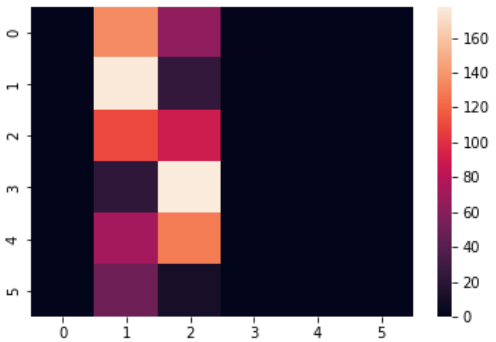

図1　単独モデルでのEC番号推定結果。縦軸が正解ラベル、横軸が推測結果。各軸共に[0:EC番号1番、1:EC番号2番、2:EC番号3番、3:EC番号4番、4:EC番号5番、5:EC番号6番]。数の多いEC番号2番および3番という推測結

Copyright ©2022 Information Processing Society of Japan.
All Rights Reserved.

果が多い結果になった。

このため複数の判別モデルを用いて判断を行なった。各モデルの担当する問題は図の通りであり、各モデルとも同等のデータ量を 2〜3 値分類する問題へ分割を行った。

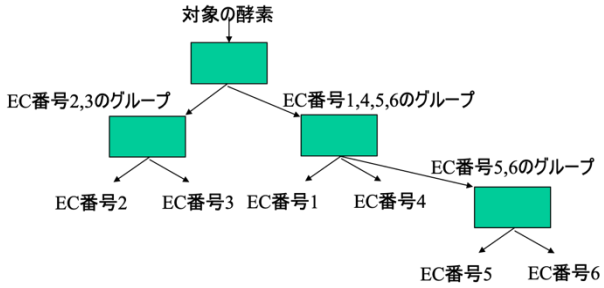

図 2　作成モデルの全容。長方形が各モデルの模式図。

ハイパーパラメーターはあらかじめ複数パターンで学習をし、各モデルで macro-f 値の最も高い条件を採用した。イテレーターは全て I=:0.01、学習回数は EC 番号 23,1456 の分類モデルで 100000 回、EC 番号 2,3 の分類モデルで 90000 回、EC 番号 1,4,56 の分類モデルで 20000 回、EC 番号 5,6 の分類モデルで 5000 回とした。

4. 結果・考察

以下の図が結果である。

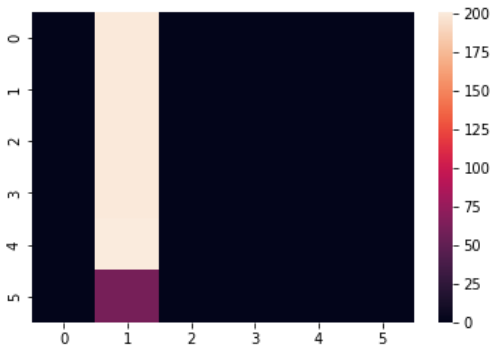

図3　EC 番号推定結果。縦軸が正解ラベル、横軸が推測結果。軸の定義は図 2 と同じ。

EC 番号 2 または 3 番か、それ以外かを判別する model で前者を、EC 番号 2 番か 3 番を判別する model で前者と判断される結果となった。このことから VOLTES データ構造から EC 番号を推測する困難さは各 EC 番号に分類されるタンパク質の数に起因するものではないことが分かった。似たようなタスクとして ENZYME の結果[5]があるが、今回はそれより芳しくない。この原因は ENZYME でのグラフデータの定義と VOLTES でのデータ定義の違いに起因する。ENZYME ではノードを既知の二次構造としているのに対し、VOLTES は原子レベルの立体構造とトポロジーを示し、原子の種類は陽に含まれていない。正しい立体構造では、原子座標のみから共有結合情報を再現できる(森本ら 2011)[6]ため、GCN で立体構造から原子の種類などの化学情報も含めて学習できることを期待したが、今回の学習条件では実現できなかったことが原因ではないかと考えている。また、今回対象とするタンパク質を単鎖に限定している。超グラフを用いられるもの[7]に変更し、対象タンパク質の限定が、今回のタスクに影響していないか検証する必要がある。

5. まとめ

今回、VOLTES の機械学習への応用を目指して Pytorch Geometric を用いて立体構造とトポロジー情報から EC 番号の推定を行なった。結果は数の多い EC 番号に推測結果が偏り、原因として VOLTES に内在しているデータを GCN で二次構造の学習またはそれに相当する情報まで学習できなかったことなどが挙げられる。

6. 参考文献

[1]https://pytorch-geometric.readthedocs.io/en/latest/index.html
[2] https://pdbj.org
[3] PDBj データベース内の Home/pub/pdb/data/biounit / 以下のディレクトリに存在する PDB ファイルを使用。(アクセス日時: 2021/12/05/17:57JST)
[4] Thomas N. Kipf, Max Welling ”Semi-Supervised Classification with Graph Convolutional Networks” ICLR, 2017
[5] Christopher Morris, Nils M. Kriege, Franka Bause, Kristian Kersting, Petra Mutzel, Marion Neumann ” TUDataset: A collection of benchmark datasets for learning with graphs ” ,(Graph Representation Learning and Beyond (GRL+), ICML 2020 Workshop.)
[6] 森本康幹, 中川隆司, 小島正樹 SAXS 法によるタンパク質立体構造の計算科学的解析, 「生物物理」 51, 88-91 (2011)
[7] Song Bai, Feihu Zhang, Philip H.S.Torr “ Hypergraph Convolution and Hypergraph Attention” Pattern Recognition October 13, 2020

Copyright ©2022 Information Processing Society of Japan.
All Rights Reserved.