4ZK-04

タンパク質分子のトポロジーと立体構造に基づくGCNによるEC番号の推定

竹内　啓†　　青柳　詠美†　丸山　直道†　小島　正樹‡

東京薬科大学大学院 生命科学研究科†　　東京薬科大学生命科学部‡

1. はじめに

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、連続的な3次元空間を探索対象とするため、従来の実験的手法を完全に代替するまでには至っていない。本研究室では、分子の位相幾何学的特徴に基づく立体構造をグラフとして表現し、構造予測から進化計算や創薬探求まで目指すVOLTES（Virtual Optimization of Local Tertiary Structures）プログラムパッケージを開発した。VOLTESは、立体構造を系の自由度に等しい二面角座標を用いて木構造で表現すること、および二面角座標を6値化し構造空間を離散化することにより、上記の課題に対応している。今回、VOLTESの機械学習への応用を目指し、AIの特徴量としての有用性を確認するため、Pytorch Geometric[1]を用いて立体構造とトポロジー情報からEC番号の推定を行った。

2. VOLTESの原理

一般にタンパク質分子のトポロジーがグラフ理論の「木」として表現できる（Abe et al., 1983）ことを利用し、VOLTESでは二面角座標をN末端から順に並べた「木」構造のデータを、計算機上で入れ子の（nested）リストとして実装した。ここで単結合BCのまわりの二面角は、原子AとBとCを含む平面と、原子BとCとDを含む平面のなす角として定義される。このとき各二面角の値に応じて6種類の配座異性体が存在するため、これを巡回6進数（0〜5の数値）で表してVOLTESの座標とした。全ての二面角のVOLTES座標を、上記の分子トポロジーに対応する入れ子のリストとして表現したものをVOLTESフォーマットと呼ぶ。本研究室ではこれまでVOLTESを用いてタンパク質の論理的構造設計や分子進化とトポロジーとの相関解析[2]などを行ってきた。また、VOLTES形式に基づくデータベースの構築や機械学習への応用を目指した研究を継続している。

3. 先行研究

本研究室では、VOLTESフォーマットを直接グラフ構造に変換し、GCN(Graph Convolutional Network)を用いてEC番号を分類した[3]。しかし、結果は数の多いEC番号に推測結果が偏ってしまい、分子のトポロジーや立体構造の情報を有するVOLTES形式のデータの特徴量を、機械学習によって十分に抽出できなかった。

4. 提案手法

VOLTESフォーマットを直接入力とせず、巡回6進数の扱いと、立体構造の特徴に着目し、以下の2点の特徴量抽出手法を考案した。まず巡回6進数を数値として扱うのではなく、6次元特徴量にパケット化することにより、不当な大小関係を排除した（カテゴリ化モデル、例えば3は[0,0,0,1,0,0]）。また、タンパク質分子は残基単位で構成されているという考えに基づき、主鎖・側鎖により構成される残基ごとに、6次元特徴量をグループ化した（グループ化モデル）。グループ化の際は、グループ内の主鎖・側鎖の各カテゴリ値を加算する方式とした。

5. 実験方法

pdbj[4]にBioUnitとしてエントリーされているタンパク質[5]のうち、1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VOLTESフォーマットに変換したタンパク質データを対象とした。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を表1に挙げる。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を用いた。以上のデータに基づいてGCNConv[6]によりEC番号を推定し、分類精度により特徴量の有用性を判断した。

Estimation of EC number based on topology and conformation of proteins

† Kei Takeuchi, Eimi Aoyagi, Naomichi Maruyama, 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

表 1　本実験での対象酵素数

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

以下の 3 つのモデルを対象として表 2 の条件にて学習を行い、分類精度を比較した。
モデル 1：先行研究
モデル 2：カテゴリ化モデル
モデル 3：カテゴリ化＋グループ化モデル

表 2　実験条件

項目	内容
環境	Pytorch Geometric
モデル構成（GCN） 入力層×隠れ層 1×隠れ層 2	モデル 1：1×6×100 モデル 2：6×30×150 モデル 3：6×30×150
分類	線形パーセプトロン
活性化関数	ReLU
最適化	ADAM
損失関数	CrossEntropyLoss （クラス間不均衡対策）
ハイパーパラメータ	学習率：0.001, バッチサイズ：50, Weight Decay：0, エポック数：20000

6.　結果・考察

学習曲線を図 1 に、分類精度を表 3 に示す。

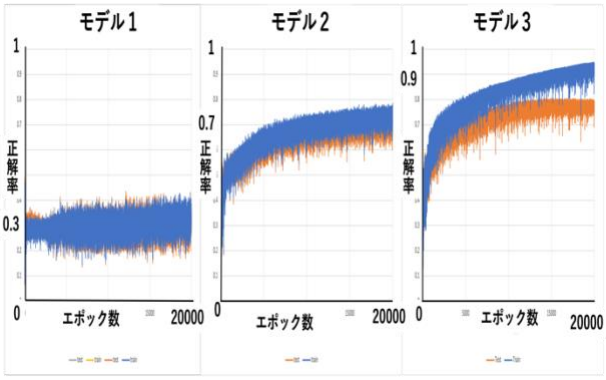

図 1　学習曲線

（左からモデル 1、モデル 2、モデル 3
それぞれ青：学習データ、橙：テストデータ）

表 3　学習精度の結果

モデル	学習精度(テスト精度)
モデル 1	32.2%（30.1%）
モデル 2	78.4%（74.2%）
モデル 3	93.5%（78.6%）

モデル 1 とモデル 2 の比較から、カテゴリ化の効果が高いことがわかった。モデル 2 とモデル 3 の比較から、グループ化の効果が有意に認められた。以上のことから、カテゴリ化とグループ化は、分子構造を表現する特徴量として有効であると考えられる。

7.　まとめ・今後の課題

先行研究の結果を踏まえて、新たな特徴量の抽出方法を考案し、カテゴリ化とグループ化の有効性を確認した。

今回は主鎖側鎖の 1 つの残基のカテゴリ化した特徴量を単純に加算してグループ化したが、主鎖側鎖の重み付けや、隣接残基のグループ化等など、さらに望ましいグループ化法について現在検討している。また、特徴量の有効性の検証を主目的としたため、汎化性の向上の検討は未実施である。データ拡張やパラメータチューニング等学習プロセスの改善や、本提案で用いた GCN 以外のモデルの検討も課題である。

8.　参考文献

[1] https://pytorch-geometric.readthedocs.io/en/latest/index.html
[2] 寺林杏理, 東海林暁貴, 小島正樹 “LISP プログラミングによるグラフ理論の諸問題の効率的なアルゴリズム” 「東京薬科大学研究紀要」24, 33-37 (2021)
[3] 青柳詠美, 小島正樹 “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” 情報処理学会第 84 回全国大会 1ZM-01 (2022)
[4] https://pdbj.org
[5] 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)
[6] Thomas N. Kipf, Max Welling ”Semi-Supervised Classification with Graph Convolutional Networks” ICLR, (2017)