3–18　有機合成における酵素番号予測のための特徴選択とクラスタリングを用いたケモインフォマティクス

奥原研究室
1815070 武藤克弥

1. はじめに

近年，新型コロナウイルスによる新薬需要の増加に伴い，ケモインフォマティクスと呼ばれる情報技術を用いて，化学反応を予測する研究が盛んに行われている．その際，酵素を触媒として使うことが，環境や反応効率の面から頻繁になってきている．目的の生成物を得るための，最適な酵素を予測するため，与えられた反応式と，酵素番号で分類される酵素の代表的な反応式をクラスタリングで比較する．最も類似性が高いと評価される反応式の酵素番号を予測する．

2. 有機合成分野と情報分野の関わり

化学は情報工学に似た特徴を持っていることから情報分野に適用できる分野が多く，有機合成においては，最適な反応経路設計や反応予測・分類等において情報技術が多く用いられている．

酵素は EC 番号という 4 組の数字の組み合わせで，性質ごとに分類されている．データベース KEGG では酵素の情報とその代表的な反応式，各反応式の化合物情報を，PubChem では化合物 ID の対応情報および化合物の SDF ファイルを取得する．

3. ケモインフォマティクスと情報技術

化合物は SMILES やフィンガープリントといった，文字列やビット列で表現することで，物性値の特徴量を持つ多次元ベクトルに表現することができ，機械学習を用いた構造比較や反応予測を円滑に行える．

反応式中の化合物を反応物と生成物のペアとみなし，特徴的な構造部分の反応変化に着目して最適な EC 番号の反応式に分類する研究が多く行われている [1].

4. 提案手法

KEGG と PubChem から必要とするデータテーブルを取得し，EC 番号と SMILES に変換した反応物，生成物からなる対応表を取得する．対応表内の各化合物に対して，Python の RDKit ライブラリにある 208 種の物性値を計算し，生成物と反応物の差分を取り各反応の 208 種の特徴変化量を取得する．

階層型クラスタリングを用いて，相関の高い特徴同士を標準化およびその平均をとって新たな合成特徴量を作成して次元削減を行う．最終的に SOM によるクラスタリングを行い．与えられた反応式と各 EC 番号の反応式の類似性を評価する．

5. 実験結果ならびに考察

208 種の特徴から全て 0 および，発散しているもの，またフラグメントに関する特徴量を除外し，106 種まで特徴を削減した．また，階層的クラスタリングによって 74 種まで減少した，R 言語で記述された SOM のプログラム [2] の実行結果は以下のようになった．

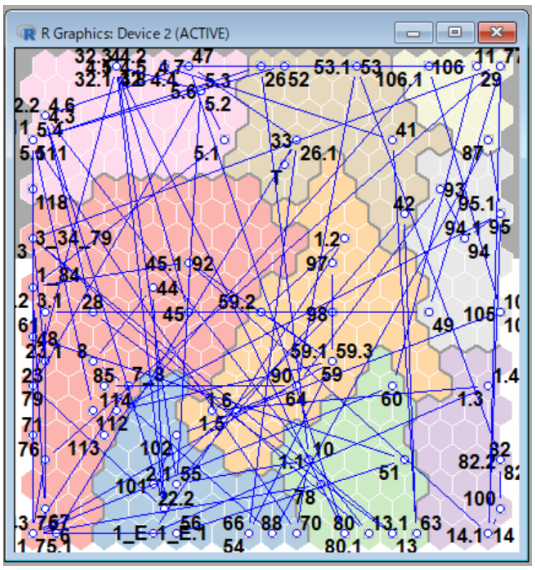

図 1:SOM によるクラスタリング結果

ラベルは EC3.1.1 の 4 番目の番号となっている．目的の反応式 (ラベル T) に対して EC3.1.1.33 が最も近くに分類している．

6. おわりに

生成物と反応物の特徴量の差分を取った特徴変化量を定義し，次元削減と SOM によって目的の反応式と各 EC 番号反応式の類似性を比較した．今後の課題として，全ての EC 番号に対して，特徴変化量を求めて，EC 番号予測の精度を検証していきたい．

参考文献

[1] Qian-Nam Hu, Hui Zhu, Xiaobing Li, Manman Zhang, Zhe Deng, Xiaoyan Yang, Zixin Deng, “Assignment of EC Numbers to Enzymatic Reactions with Reaction Difference Fingerprints”, *PLOS ONE*, Vol. 7, No. 12, 2012.

[2] 福嶋 瑞希 “環境認識ライフログからの行動パターン解析による類似性・イベント検出”, 富山県立大学学位論文 2018.