

背景

有機合成と酵素

提案手法

今後の予定

有機合成における酵素選択のための テキストマイニング

1815070 武藤 克弥

富山県立大学 電子・情報工学科

January 7, 2022

背景

近年、有機合成に生体触媒として酵素を用いることが、高い反応性、グリーンケミカルの面で良いことから積極的に行われている。

酵素に関する情報は生物分野に広く展開しているが、合成熟練者なら、ある反応に用いるべき酵素をある程度把握できる。

しかし、酵素の複雑な特性ゆえ、合成の知識内では対処できない場合が多く、酵素の専門家に依頼し、酵素候補を絞ってもらう必要があった。

目的

- ① 有機合成側のユーザに対し、適切な酵素候補を提示するシステム設計を目指す

酵素とは

- 生体内で代謝を起こすための生物に必要となるもの
- らせん構造で、分子的にはかなり大きい
- ところどころにある穴に化合物 (タンパク質) が入り、化学反応を起こす
→化合物を変化させるための「生体触媒」として作用する
- 「基質特異性」があり、穴に入る化合物は限られている
→その代わり入れば反応が速く進む (化学触媒に比べて速い)

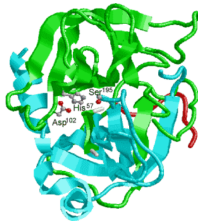


図 1: 酵素の構造

EC 番号 (Enzyme Commission numbers)

- 使われる化学反応ごとに、酵素を分類したもの
→ 4 組の番号の組み合わせ (X.X.X.X)
- 結合の種類, 作用などで 4 層に細分化されている

背景

有機合成と酵素

提案手法

今後の予定

- EC 1.X.X.X — オキシドレダクターゼ (酸化還元酵素)、酸化還元反応を触媒
- EC 2.X.X.X — トランスフェラーゼ (転移酵素)、原子団 (官能基など) をある分子から別の分子へ転移する
- **EC 3.X.X.X** — ヒドロラーゼ (加水分解酵素)、加水分解反応を触媒
- EC 4.X.X.X — リアーゼ (脱離酵素)、原子団を二重結合あるいは、結合の解離の触媒
- EC 5.X.X.X — イソメラーゼ (異性化酵素)、分子の異性体を作る
- EC 6.X.X.X — リガーゼ (合成酵素)、ATP の加水分解エネルギーを利用して、2 つの分子を結合させる
- EC 7.X.X.X — トランスロカーゼ (輸送酵素)、生体膜を超えてイオンや分子等の局在を移動させる

[https://ja.wikipedia.org/wiki/EC%E7%95%AA%E5%8F%B7_\(%E9%85%B5%E7%B4%A0%E7%95%AA%E5%8F%B7\)](https://ja.wikipedia.org/wiki/EC%E7%95%AA%E5%8F%B7_(%E9%85%B5%E7%B4%A0%E7%95%AA%E5%8F%B7))

EC 3.1.30.- (リボ核酸またはデオキシリボ核酸に作用する、

- EC 3.1.30.1 ϕ アスベルギルスヌクレアーゼ S1
- EC 3.1.30.2 ϕ セラチア・マルセッセンスヌクレアーゼ

EC 3.1.31.- (リボ核酸またはデオキシリボ核酸に作用する、

- EC 3.1.31.1 ϕ ミクロコッカス・ヌクレアーゼ

EC.3.2.- (グリコシラーゼ) [編集]

EC.3.2.1.- (O-およびS-グリコシル化合物加水分解酵素)

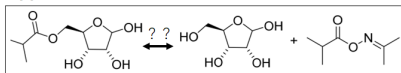
- EC 3.2.1.1 ϕ α -アミラーゼ
- EC 3.2.1.2 ϕ β -アミラーゼ
- EC 3.2.1.3 ϕ グルコアミラーゼ

図 2: EC 番号

EC 番号提示システム

- ある反応を与えた際、使うべき酵素の EC 番号を提示してくれるシステム
- ある反応式と EC 番号酵素 (の代表的な) 反応式の「同じ項にある化合物」の類似度を比較する
- 類似的距離に近い酵素反応の EC 番号を提案する

対象の反応



同じ項にある化合物どうし
を比較する

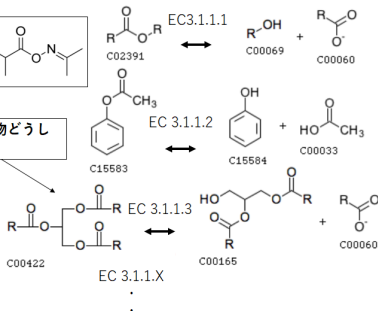


図 3: 対象とする反応式と EC 番号酵素の反応式

背景

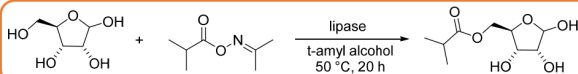
有機合成と酵素

提案手法

今後の予定

- 対象の反応 = リボースのエステル化反応 → EC 3.1.1.X(カルボン酸エステル加水分解酵素) に該当する
- 【論文内】対象の反応は EC 3.1.1.3 の酵素製品を使っている(答え)
- 残り一桁にあたる酵素(約 100 個)の中から最も高い類似性として EC3.1.1.3 の酵素をシステム側で提示できればよい

対象の反応：リボースのエステル化反応 → EC 3.1.1.X に該当する



EC 3.1.1.3に分類される酵素製品を使えば一番収率がいいことは分かっている
→システムで3.1.1.3が最も類似度が高く出てほしい

Enzyme Name	5-isobutyryl ribose assay yield%
Novozym 435 (Candida antarctica lipase B)	Novozym社製 65
IMMILL-T2-150 (Thermomyces lanuginosus lipase)	40
IMMRES-T2-150 (Resinase HT lipase)	Novozym社製 38
IMMLPX-T2-150 (Lipex 100 L lipase)	Novozym社製 56
IMML51-T2-150 (Novozymes 51032)	Novozym社製 61
IMMP6-T2-250 (protease from Bacillus licheniformis)	11
Lipozyme RM IM	Novozym社製 10
CDX IMB-103	33

図 4: 対象とする反応と酵素製品

SMILES

- 構造式を文字列に変換したもの
- アルファベットが元素, `[],()` や`@`が形状を判断している
- Python Rdkit ライブラリに SMILES から構造解析を行う関数が数多くある

→ EC 番号反応式の化合物を全て SMILES に変換して類似度比較を行う

```
Chem.MolFromSmiles('[C@H]([C@H](CO)O)([C@H](C=O)O)O')
```

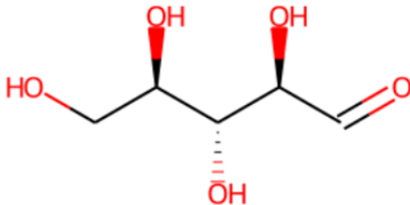


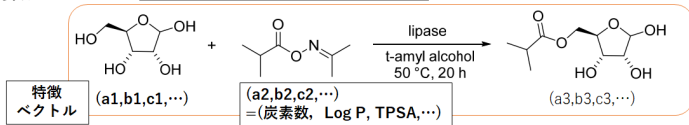
図 5: SMILES 表記と構造式

SMILES からの特徴量抽出

- ① 化合物 (SMILES) の各特性値を特徴として数値化
- ② 特徴値を何個か選んで多次元ベクトルにする

特徴抽出

対象の反応：リボースのエステル化反応



特徴となりうるもの：炭素数、酸素結合数、Log P(疎水性)、TPSA(極性表面積)、電子密度など
(記述子ともいう)

もしくは、フィンガープリント(構造をビット列で特徴づけたもの)

SMILESを特徴量 出力関数
(Python Rdkitライブラリ)
に入れると値が出てくる

EC 3.1.1.X番台の各反応式の化合物をすべてベクトル化

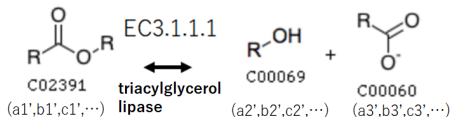


図 6: 特徴ベクトルの作成

特徴量選択と類似クラスタリング

9/13

特徴量 (変数) 選択に用いる手法 (AIC)

- 特徴量を多く用いるほど、その化合物の特性をよく表す⇔その反面、過適合を起こしやすくなる
- 特徴量が少ないと、類似性比較の精度があまりよくない
→適切な特徴量を必要な数だけ選ぶようにする

$$AIC = -2 \ln L + 2k$$

- L: 最大尤度 モデルの確からしさ
- k: パラメータの数
- 各説明変数について、それぞれAICを計算、もっとも小さいAICが最適解

SOM

使うデータベース

- ① KEGG : 遺伝子・タンパク質情報, タンパク質相互作用を表した KEGG PATHWAY, 酵素情報を表した KEGG ENZYME, 主に酵素反応の反応式について記した KEGG REACTION, 生態系に関連する化合物を集めた KEGG COMPOUND 等のデータからなるデータベース

→ EC 番号, EC 番号の代表的な反応式, 反応式 ID, PubChem との化合物 ID 対応表を API で取得する

- ② PubChem : 化合物の化学・物理特性, 毒性情報, 引用された文献情報等を収録したデータベース

→ EC 番号に含まれる化合物の Mol ファイルを API で取得する

KEGG と PubChem で取得したデータを整理する流れを作成

B	E	I
ENTRY	EQUATION	ENZYME
R00001	C00404 + n C00001 <=> (n+1) C02174	3.6.1.10
R00002	16 C00002 + 16 C00001 <=> 8 C00138 <=> 8 C05359 + 16 C01186.1	3.6.1.1
R00004	C00013 + C00001 <=> 2 C00009	3.6.1.1
R00005	C01010 + C00001 <=> 2 C00011 + 2 C00014	3.5.1.54
R00006	C00900 + C00011 <=> 2 C00022	2.2.1.6

RID・反応式・EC番号 対応表(KEGG)

Entry	R00004	Reaction
Name	diphosphate phosphohydrolase; pyrophosphate phosphohydrolase	
Definition	Diphosphate + H2O <=> 2 Orthophosphate	
Equation	C00013 + C00001 <=> 2 C00009	
Enzyme	3.6.1.1	

Molファイル

ENZYME	left1	left2	right1	right2	right3
3.1.1.22	C04546	C00001	C01089	N	N
3.1.1.20	C01572	C00001	C01424	N	N
3.1.1.40	C02868	C00001	C01839	N	N
3.1.1.33	C02655	C00001	C00031	C00033	N

cid	pubchem_SID	pubchem_CID
C00001	3303	962
C00002	3304	5957
C00003	3305	5893
C00004	3306	439153
C00005	3307	5884
C00006	3308	5885

SID	SMILES
4103	O=C(O)CCC(=O)CC(=O)C
4104	Cc1ncc(CO)c(C(=O)O)c
4106	CCOC(C)=O
4109	O=C(/C=C/c1ccc(O)c(C
6169	*OC(=O)C(*)N

SID・SMILES対応表(PubChem)

ENZYME	left1	left2	right1	right2	right3
3.1.1.22	7152	3303	4324	N	N
3.1.1.20	4729	3303	4609	N	N
3.1.1.40	5804	3303	4958	N	N
3.1.1.33	5628	3303	3333	3335	N

EC・SID対応表

各SID molファイル取得→smilesに変換

ENZYME	left1	left2	right1	right2	right3
3.1.1.22	C[C@@H][H]O[H]	C[C@@H]N	N	N	N
3.1.1.20	O=C(O)c1[H]O[H]	O=C(O)c1N	N	N	N
3.1.1.40	Cc1cc(OC)[H]O[H]	Cc1cc(O)cN	N	N	N
3.1.1.33	CC(=O)O([H]O[H]	OC[C@H] CC(=O)O	N	N	N
3.1.1.6	*OC(C)=C[H]O[H]	*O CC(=O)O	N	N	N
3.1.1.1	*OC(*)=O[H]O[H]	*O *(C=O)[O]N	N	N	N

EC・SMILES対応表

図 7: EC 番号・SMILES 対応表

やってきたこと 2

- ① 特徴量ベクトルの作成
- ② AIC の例題「溶解度を説明する変数の選択」を python で解いた

今後の予定

- どのようにして特徴量を選択するか検討
- SOM の実装