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
有機合成と酵素
提案手法
今後の予定

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

有機合成と酵素

背景
有機合成と酵素
提案手法
今後の予定

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)

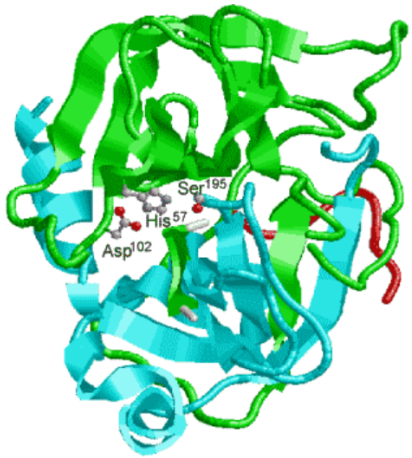


図 1: 酵素の構造

有機合成と酵素 2

背景
有機合成と酵素
提案手法
今後の予定

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

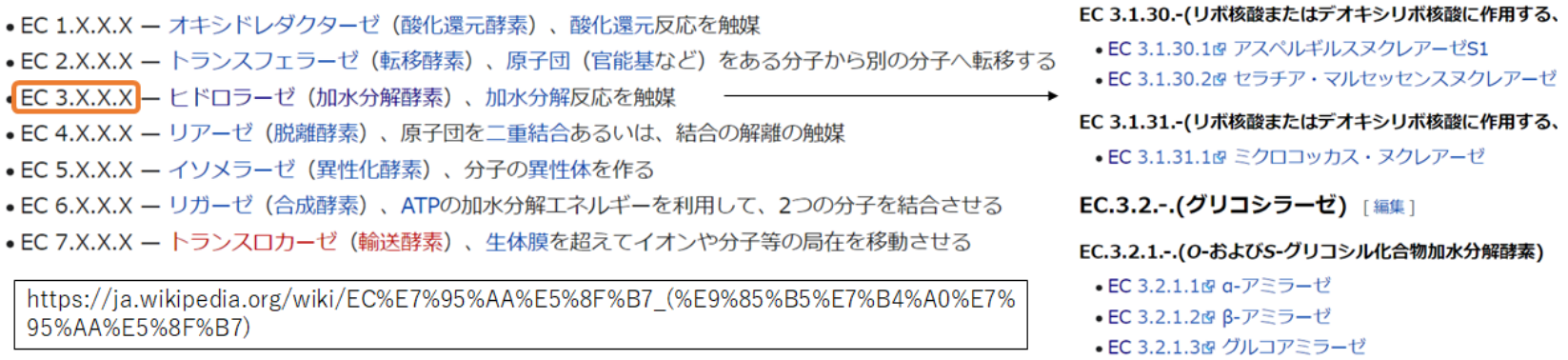


図 2: EC 番号

提案手法

背景
有機合成と酵素
提案手法
今後の予定

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
- 類似的距離が近い酵素反応の EC 番号を提案する

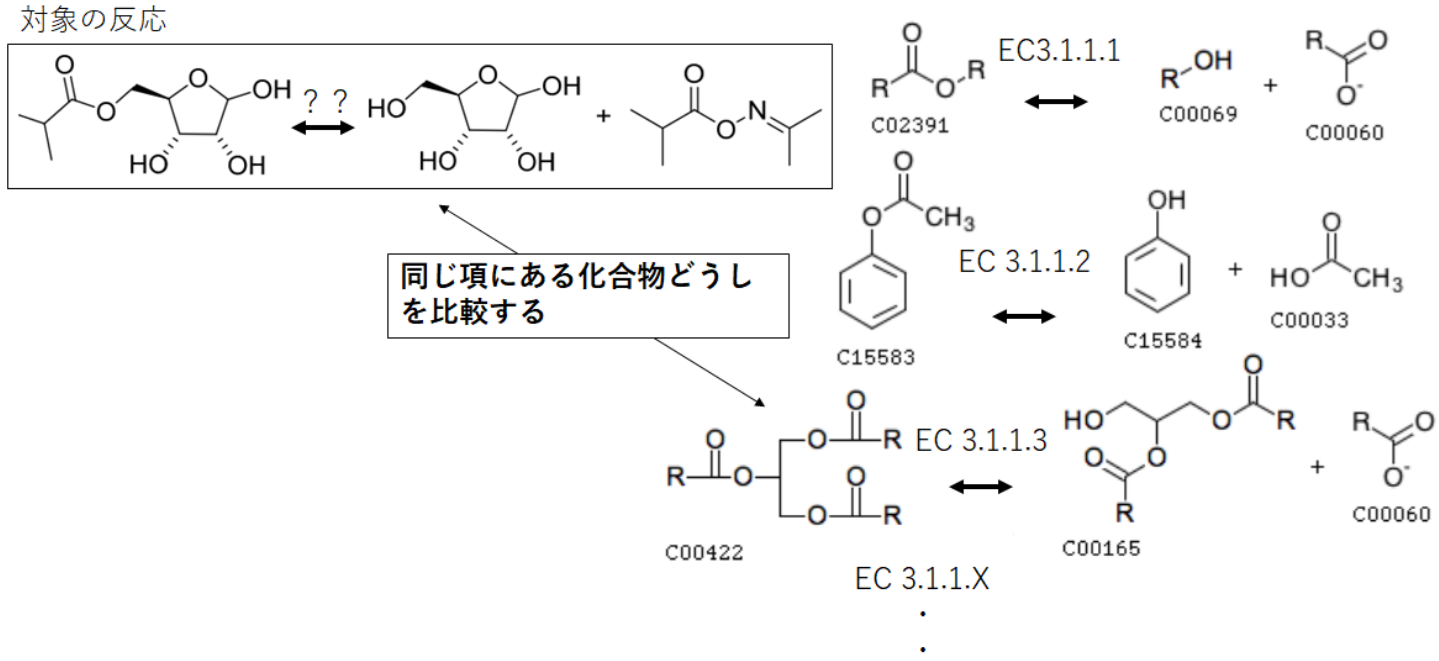


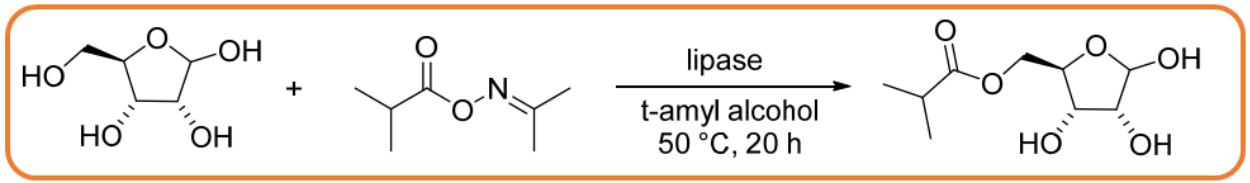
図 3: 対象とする反応式と EC 番号酵素の反応式

提案手法

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きてほしい

Enzyme Name		5-isobutyryl ribose assay yield%
Novozym 435 (Candida antarctica lipase B)	**Novozym社製**	65
IMMTLL-T2-150 (Thermomyces lanuginosus lipase)		40
IMMRES-T2-150 (Resinase HT lipase)	**Novozym社製**	38
IMMLIPX-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

類似度比較手法

背景
有機合成と酵素
提案手法
今後の予定

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

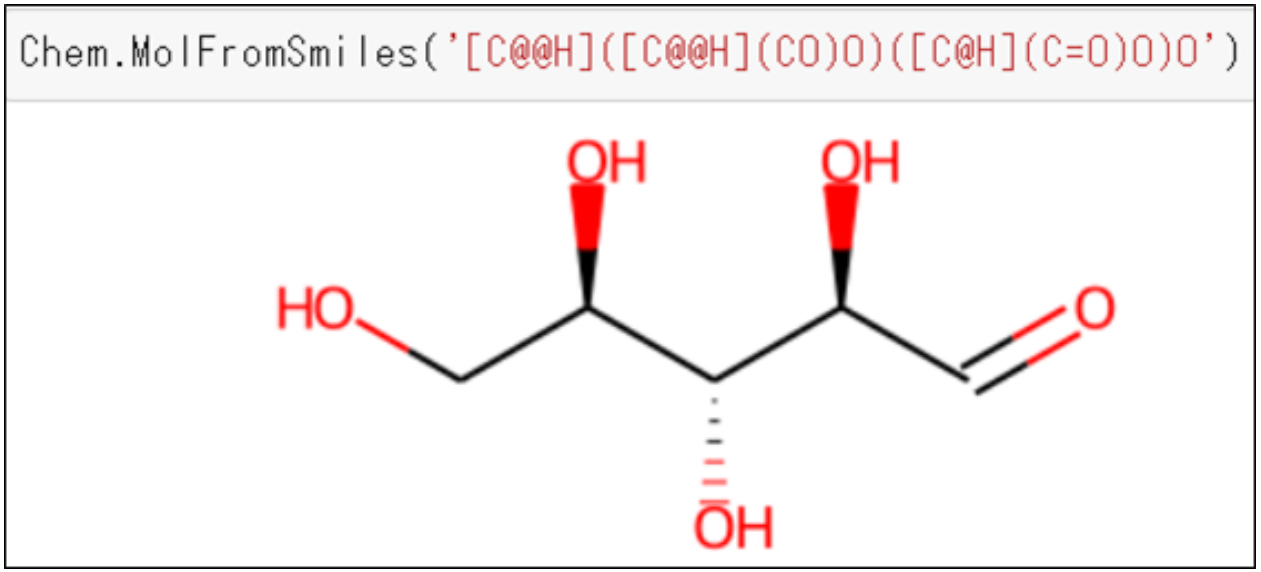


図 5: SMILES 表記と構造式

類似度比較手法

背景
有機合成と酵素
提案手法
今後の予定

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

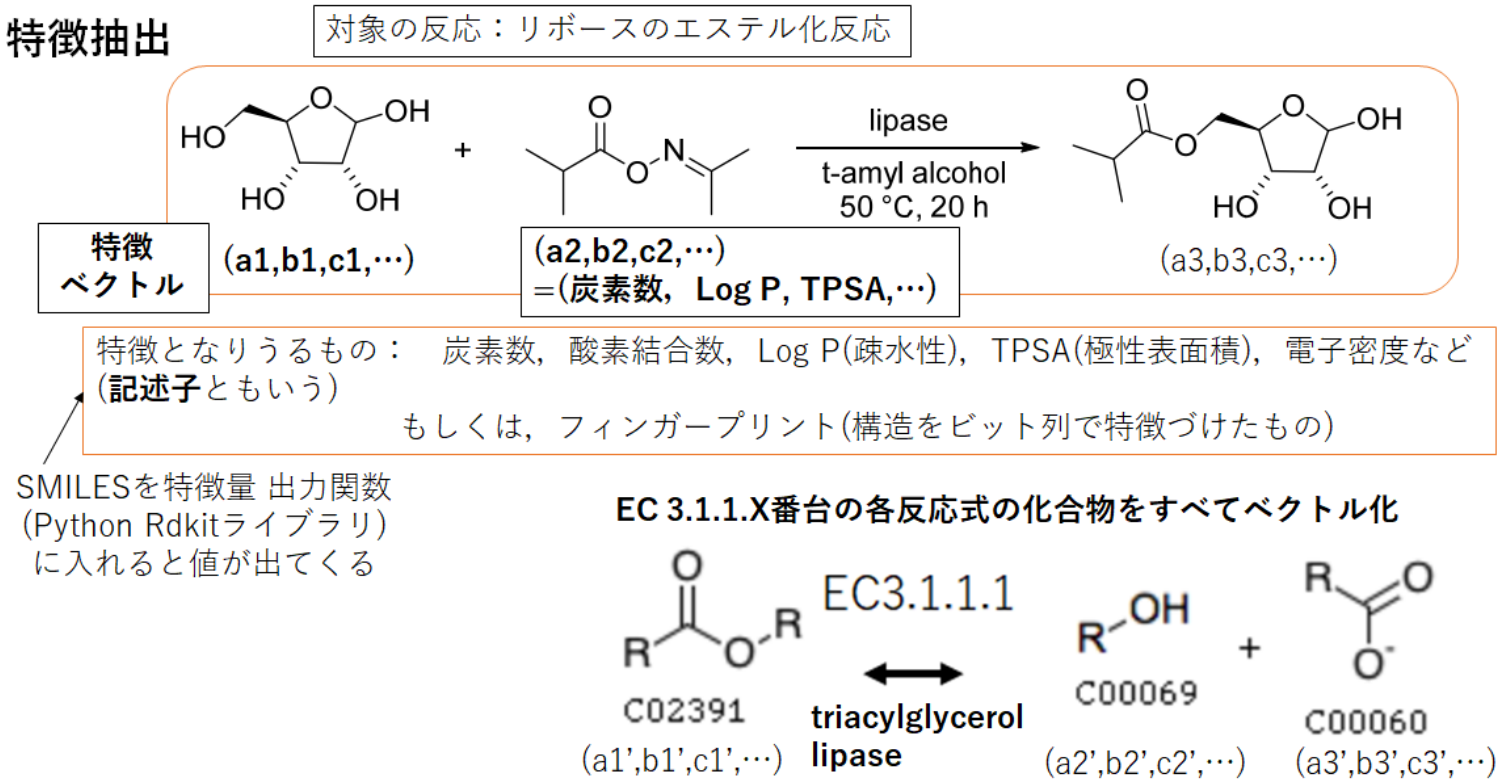


図 6: 特徴ベクトルの作成

特徴量選択と類似クラスタリング

背景
有機合成と酵素
提案手法
今後の予定

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

$$\mathrm{AIC} = -2\ln L + 2k$$

- 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

システムの流れ

背景
有機合成と酵素
提案手法
今後の予定

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

やってきたこと

背景
有機合成と酵素
提案手法
今後の予定

KEGG と PubChem で取得したデータを整理する流れを作成

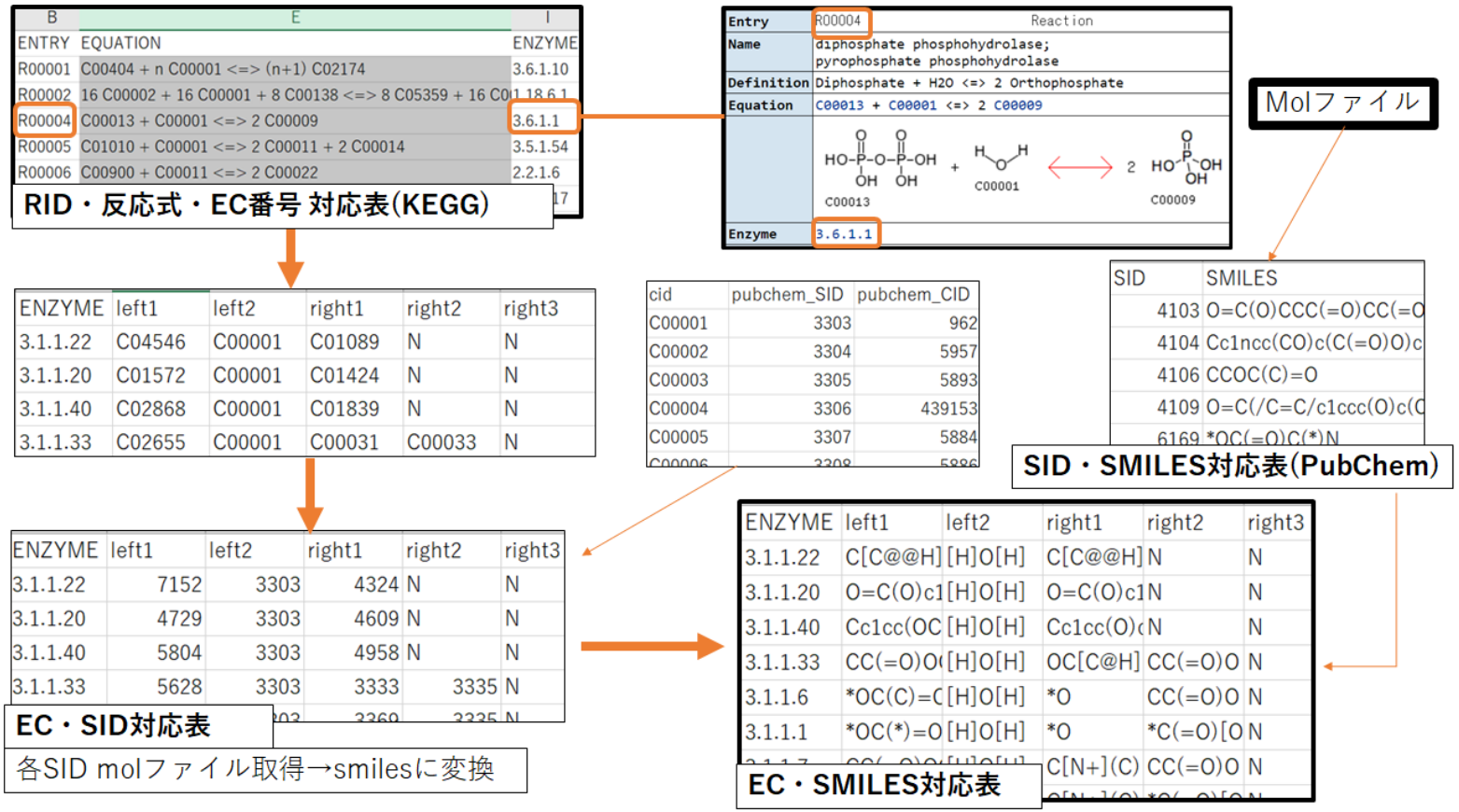


図 7: EC 番号・SMILES 対応表

やってきたこと 2

背景
有機合成と酵素
提案手法
今後の予定

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

背景
有機合成と酵素
提案手法
今後の予定

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