1. はじめに
2. 有機合成分野と情報分野の関わり
4. 提案手法
5. 実験結果並びに考察
6. おわりに

# 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

Chemoinformatics Using Feature Selection and Clustering for Enzyme Commission Number Prediction in Organic Synthesis

武藤 克弥 (Katsuya Muto)
t815070@st.pu-toyama.ac.jp

富山県立大学 工学部 電子・情報工学科 情報基盤工学講座

Teams, 14:20-14:35 Wednedday, February 16, 2022.

# 1. はじめに



1. はじめに
2. 有機合成分野と情報分野の関わり
4. 提案手法
5. 実験結果並びに考察
6. おわりに

近年，新型コロナウイルスなどの影響で新薬開発の需要が高まり，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の設計や予測を行う研究が増加している．

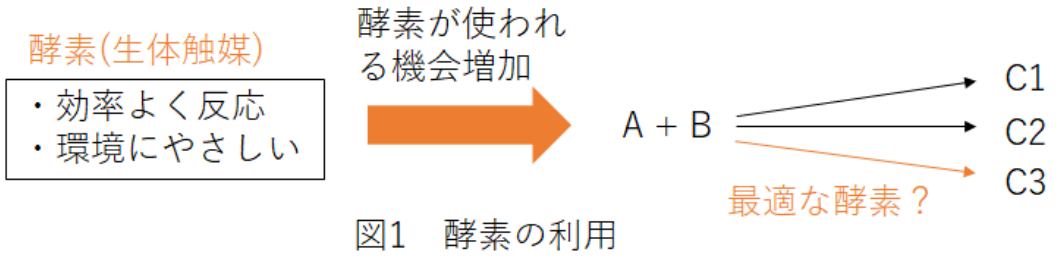



図1　酵素の利用

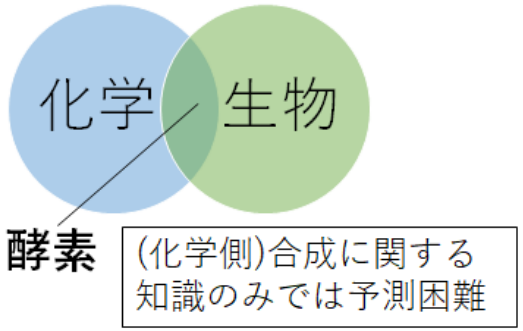



図2　酵素予測の複雑性

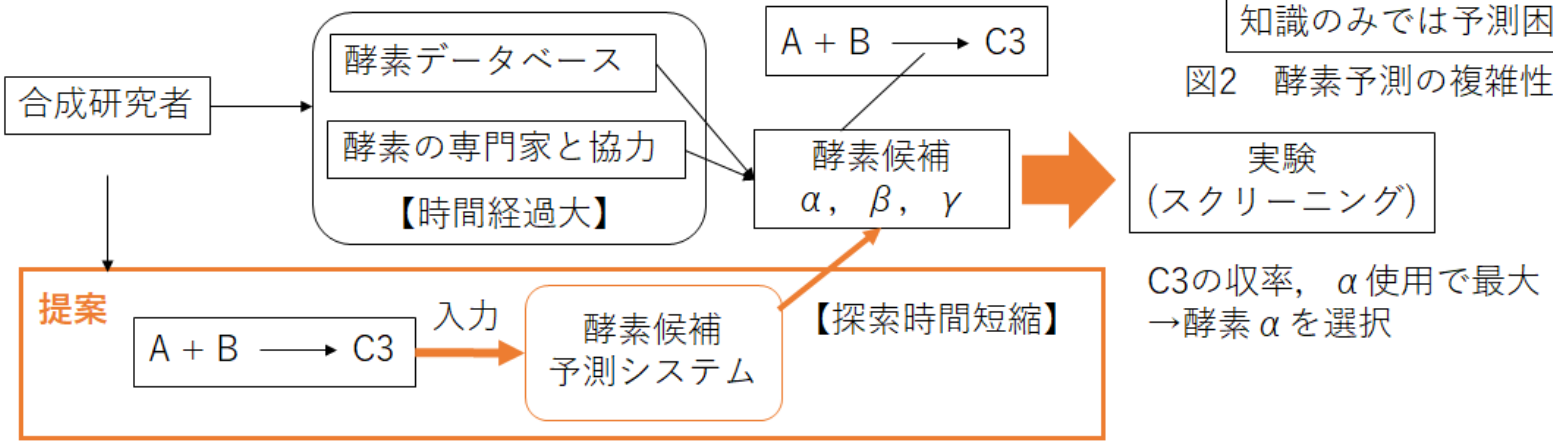



図3　酵素選択の流れ

※本研究を進めるにあたり，酵素分野に関して生物工学科酵素化学工学講座の浅野泰久教授，有機合成分野に関してくすりのシリコンバレーTOYAMA研究拠点化プロジェクトディレクター岩﨑源司博士(薬学)よりご指導いただきました．

# 2.1 有機合成と酵素および情報分野



1. はじめに
2. 有機合成分野と情報分野の関わり
4. 提案手法
5. 実験結果並びに考察
6. おわりに

有機合成では生体内に存在する有機物などを人工的に生成し，医薬品や必要となる化合物を作り出す.

**【酵素の性質】**
- **生体内の化学反応を触媒する生体触媒**
- **特定の化合物(基質)のみ作用し，反応を効率に進める.**
- **化学触媒に比べて，環境にやさしい.**

} **有機合成で用いられる機会が増加**

**化学は情報工学に適用しやすい**

**化合物構造 = グラフ理論**
**反応経路 = 経路最適化問題**

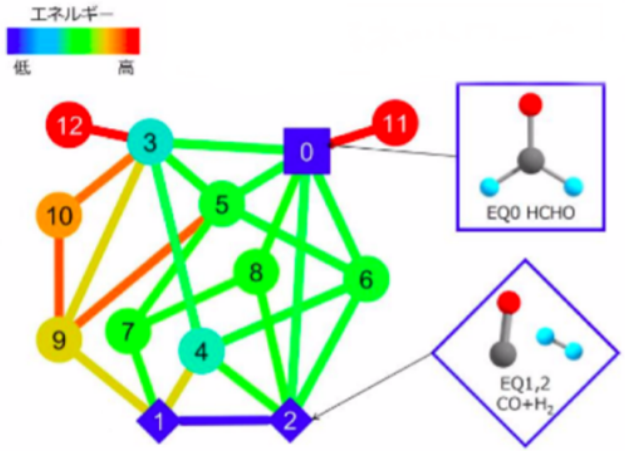



図4　最短経路探索の例[1]

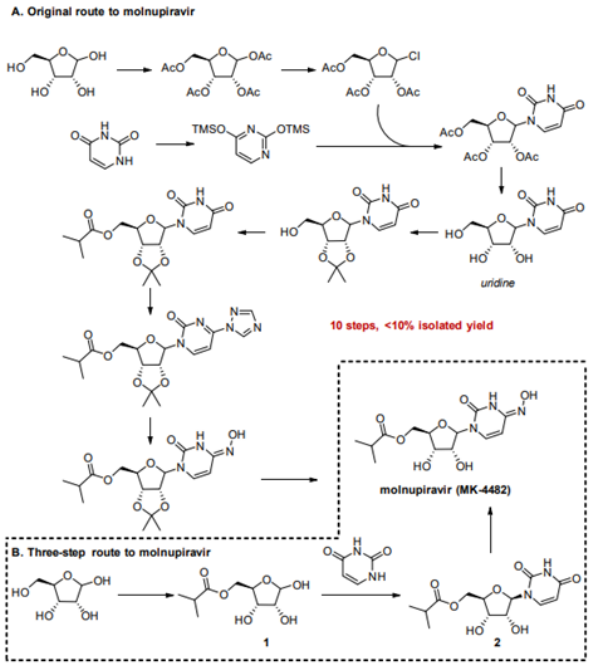



図5　酵素を用いた合成ステップ短縮化[2]

[1] 中野裕太, 瀧川一学, “化学反応ネットワークにおける最適反応経路候補の列挙”, 情報処理学会研究報告, Vol. 122, No. 16, 2019.

[2] Tamas Benkovics, John A. McIntosh, Steven M. Silverman, Jongrock Kong, Peter Maligres, Tetsuji Itoh, Hao Yang, et al., “Evolving to an Ideal Synthesis of Molnupiravir, an Investigational Treatment for COVID‐19”, *ChemRxiv*, 2020.

# 2.2 酵素番号 (EC 番号)



1. はじめに
2. 有機合成分野と情報分野の関わり
4. 提案手法
5. 実験結果並びに考察
6. おわりに

酵素を 4 組の数字 (EC ○. ○. ○. ○) の組み合わせで分類したもの.

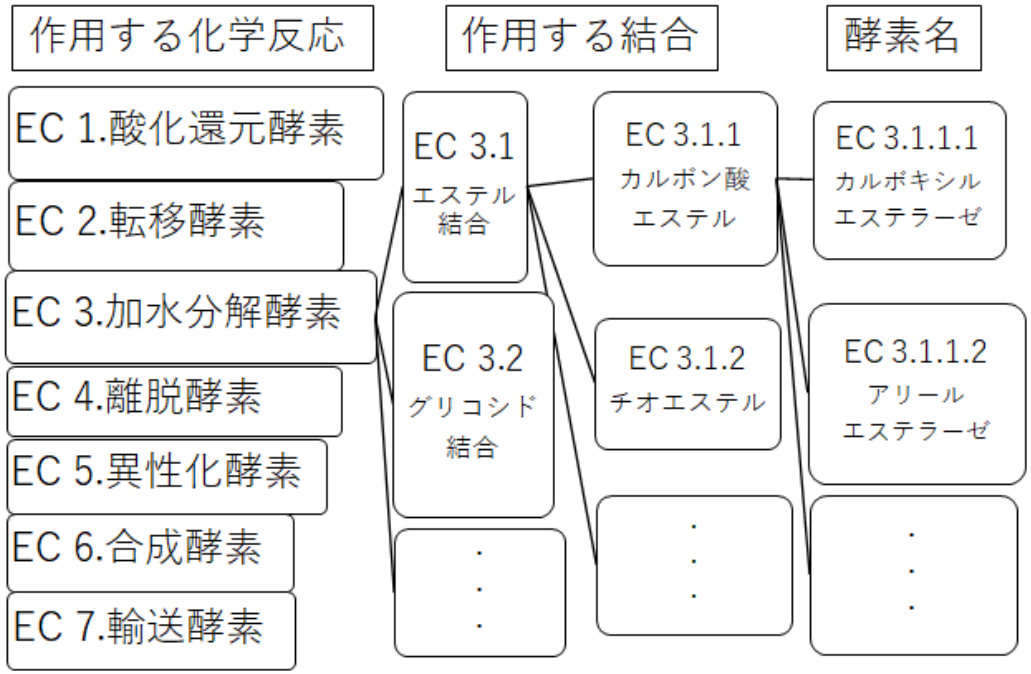



図6　EC番号による酵素分類

EC 3.1.1.1
(カルボキシルエステラーゼ)
を用いた**代表的な反応式**
=**<u>自然界で観測される反応</u>**

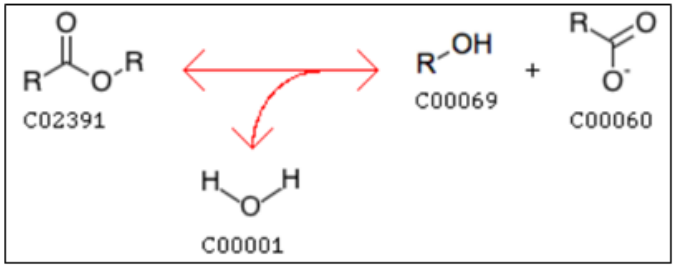



図7　EC 3.1.1.1の代表的な反応式

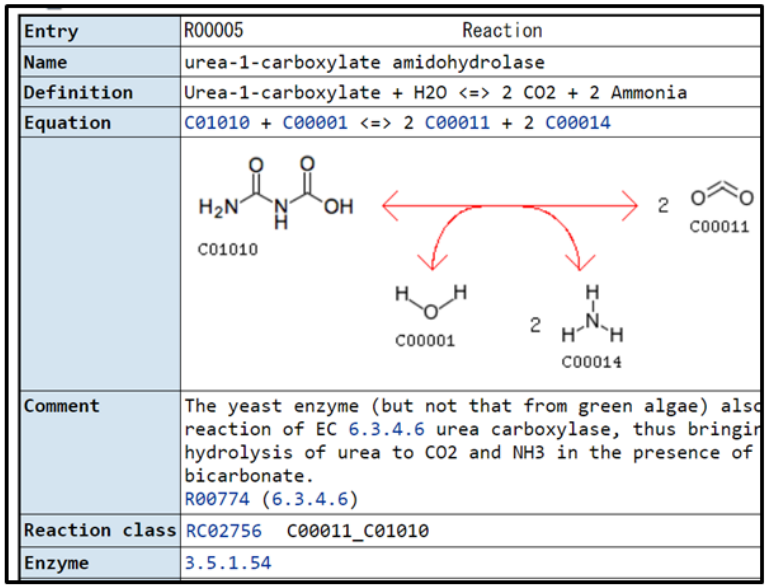
# 2.3 使用する化学データベース



1. はじめに
2. 有機合成分野と情報分野の関わり
4. 提案手法
5. 実験結果並びに考察
6. おわりに

Kyoto Encyclopedia of Genes and Genomes(KEGG) と PubChem から必要となる情報を取得する.

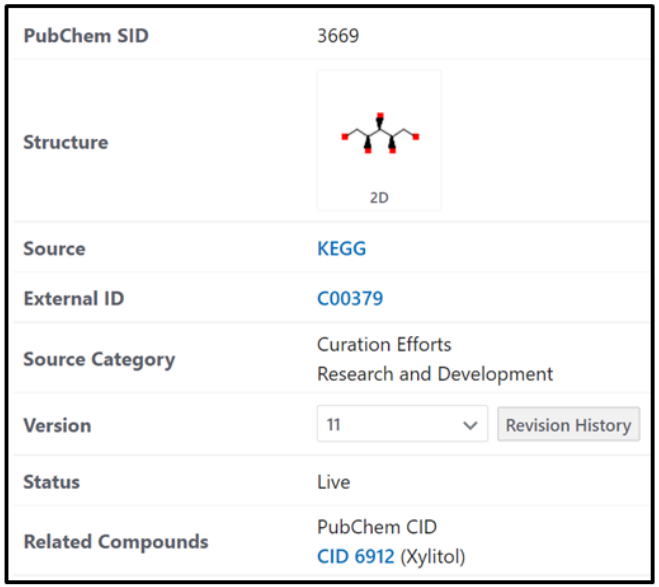
**KEGG: 酵素情報(遺伝子情報, タンパク質相互作用), 酵素の化学反応情報などを記載.**



| | |
|---|---|
| Entry | R00005 Reaction |
| Name | urea-1-carboxylate amidohydrolase |
| Definition | Urea-1-carboxylate + H2O <=> 2 CO2 + 2 Ammonia |
| Equation | C01010 + C00001 <=> 2 C00011 + 2 C00014 |
| | C01010 C00001 2 C00011 2 C00014 |
| Comment | The yeast enzyme (but not that from green algae) also reaction of EC 6.3.4.6 urea carboxylase, thus bringing hydrolysis of urea to CO2 and NH3 in the presence of bicarbonate. R00774 (6.3.4.6) |
| Reaction class | RC02756 C00011_C01010 |
| Enzyme | 3.5.1.54 |

図8　KEGGのWebサイト[3]

**PubChem: 化合物の化学・物理特性, 毒性情報, 文献情報等を収録.**



| | |
|---|---|
| PubChem SID | 3669 |
| Structure | 2D |
| Source | KEGG |
| External ID | C00379 |
| Source Category | Curation Efforts<br>Research and Development |
| Version | 11 Revision History |
| Status | Live |
| Related Compounds | PubChem CID<br>CID 6912 (Xylitol) |

図9　PubChemのWebサイト[4]

**それぞれ実装されているAPIを用いてデータを取得**

[3] "KEGG: Kyoto Encyclopedia of Genes and Genome", https://www.genome.jp/kegg/kegg_ja.html, 閲覧日 2022.1.17.

[4] "PubChem", https://pubchem.ncbi.nlm.nih.gov/, 閲覧日 2022.1.17.

# 3.2 計算機での化学構造表現，EC 番号予測



1. はじめに
2. 有機合成分野と情報分野の関わり
4. 提案手法
5. 実験結果並びに考察
6. おわりに

ケモインフォマティクスでは化学構造を計算機で扱うための，様々な表現法が開発されている．

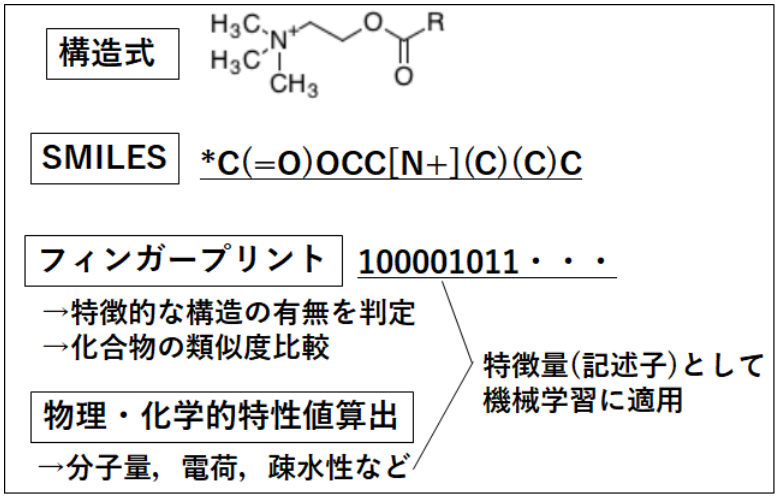
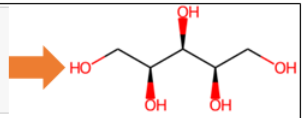



図6　計算機上における化学構造表現

**RDKit(Python)：プログラム上で化学構造を扱う**
- **SMILES**
- **フィンガープリント(構造記述子)**
- **特性値(記述子)**

```
from rdkit import Chem
Xylitol = Chem.MolFromMolFile('Xylitol.mol')
Xylitol
```



```
from rdkit.Chem import Descriptors
print("SMILES: " + Chem.MolToSmiles(Xylitol))
print("分子量: " + str(Descriptors.MolWt(Xylitol)))
print("LogP: "+ str(Descriptors.MolLogP(Xylitol)))
print("TPSA: "+ str(Descriptors.TPSA(Xylitol)))
```

```
SMILES: OC[C@H](O)[C@@H](O)[C@H](O)CO
分子量: 152.14600000000002
LogP: -2.9462999999999995
TPSA: 101.15
```

図7　RDKitを用いたSMILES・数値表現

**EC番号予測**

**EC反応式の反応物から生成物の「構造変化」を学習**
**→EC反応式の分類精度を検証**

EC 番号反応式： 反応物 1 + 反応物 2 →生成物 1 + 生成物 2

$$\rightarrow RFP = FP_{\text{生成物 1+生成物 2}} - FP_{\text{反応物 1+反応物 2}}$$

**FP：各化合物フィンガープリント**
**RFP：反応差分フィンガープリント≒構造変化**

# 4.1 提案手法(EC 番号の予測)



1. はじめに
2. 有機合成分野と情報分野の関わり
4. 提案手法
5. 実験結果並びに考察
6. おわりに

ターゲット反応式と EC 反応式の類似性を比較し，最も類似する EC 反応式の EC 番号を最適な酵素候補として予測する．

・反応物から生成物が生成されるときの構造変化をもとに反応式の類似性を比較
・ターゲットとEC反応式の構造変化が類似
＝そのEC番号の酵素を使えば目的生成物を得られる可能性がある(類似性の概念)

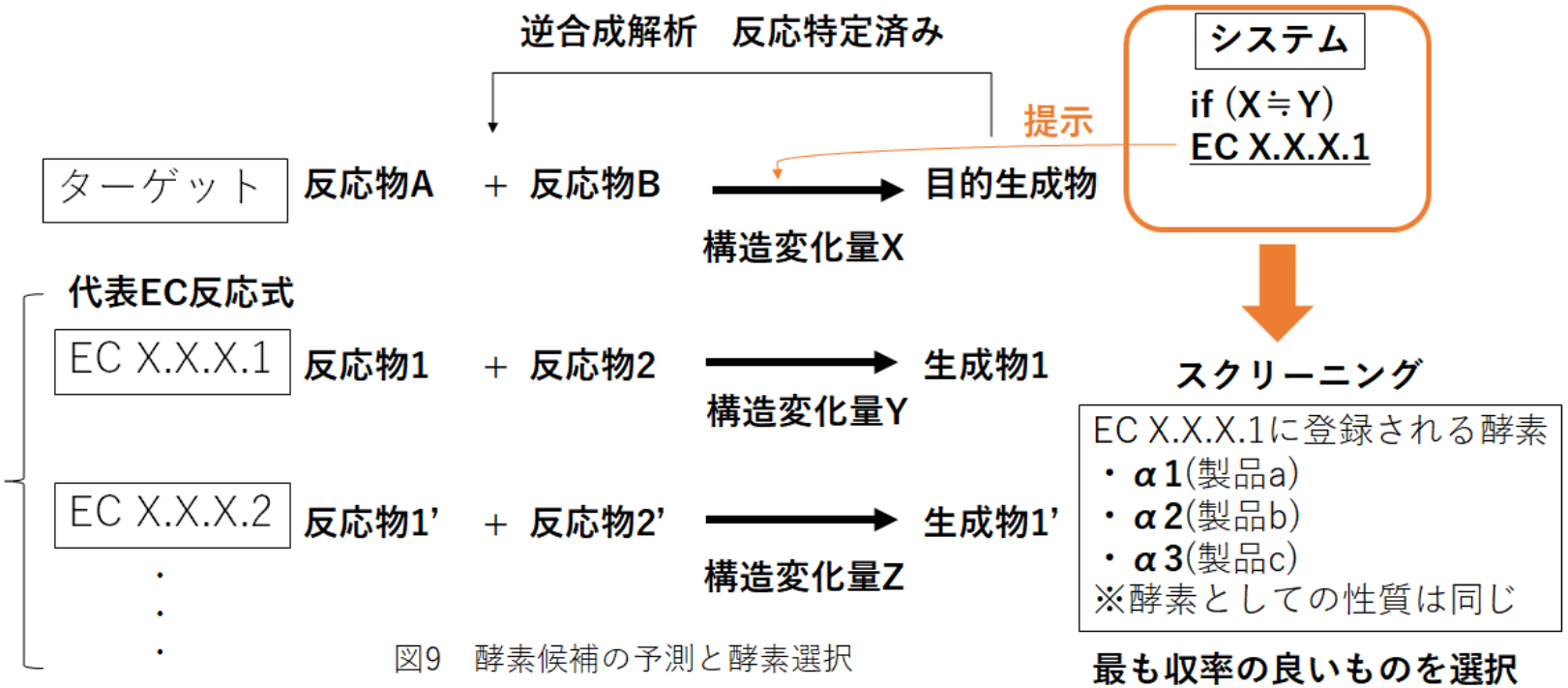



図9　酵素候補の予測と酵素選択

# 4.1 構造変化を捉える特徴



1. はじめに
2. 有機合成分野と情報分野の関わり
4. 提案手法
5. 実験結果並びに考察
6. おわりに

反応物から生成物に変化する際の「特性値変化量」で類似性比較

**先行研究：反応差分フィンガープリント**

EC 番号反応式： 反応物 1 + 反応物 2 →生成物 1 + 生成物 2

$$\rightarrow RFP = FP_{\text{生成物 1+生成物 2}} - FP_{\text{反応物 1+反応物 2}}$$

1種のフィンガープリントで
捉えられる構造変化には限界がある

(多数のフィンガープリントが
開発されている)

多数の「**物理・化学的特性値の変化量**」を用いて構造変化を捉える
(→RDKit: 特性値を表現する記述子208種)

**提案手法: 特性値変化量**

各反応式に対する記述子(n種)の特性値: $cv_j(j = 1, 2, ..., n)$

$$cv_j = (PD_1 + PD_2) - (RT_1 + RT_2)$$

($RT_i$： 反応物 $i$ の特性値, $PD_i$： 生成物 $i$ の特性値)

各反応式の特徴ベクトル: $\boldsymbol{DF}_i(i = 1, 2, ..., m)$

$$\boldsymbol{DF}_i = (cv_{i1}, cv_{i2}, \cdots, cv_{ij}, \cdots, cv_{in})$$

| | 記述子 1 | 記述子 2 | ⋯ | 記述子 $n$ |
|---|---|---|---|---|
| $\boldsymbol{DF}_1$ | $cv_{11}$ | $cv_{12}$ | ⋯ | $cv_{1n}$ |
| $\boldsymbol{DF}_2$ | $cv_{21}$ | $cv_{22}$ | ⋯ | $cv_{2n}$ |
| ⋮ | ⋮ | ⋮ | ⋱ | ⋮ |
| $\boldsymbol{DF}_m$ | $cv_{m1}$ | $cv_{m2}$ | ⋯ | $cv_{mn}$ |

表1　各反応式の特性値変化量

$\boldsymbol{DF}_T$ (ターゲット)に最も類似する$\boldsymbol{DF}_i$ のEC番号を提示

# 4.2 凝集型クラスタリングによる次元削減



1. はじめに
2. 有機合成分野と情報分野の関わり
4. 提案手法
5. 実験結果並びに考察
6. おわりに

用いる記述子が多い場合，多重共線性が発生する可能性がある．

**多重共線性:** 変数(記述子)間に高い相関関係があると発生
→同じ変数があると同義 ＝ 過学習・分類精度低下の原因

**多重共線性の対策1**
△相関の高い記述子のうち一方を削除
→重要な変数を削除する可能性
→複数の記述子間の相関関係を考慮しない

**多重共線性の対策2**
相関の高い記述子どうしをクラスタリング

**最長距離法を用いたクラスタリング**
各クラスタ内の記述子どうしで最長となる距離
(クラスタ間距離)が最短となるクラスタどうしをマージ

(相関係数0.9以上でマージ)
手順1: 相関係数が最も高いペアを順次マージ
手順2: 最長距離法でクラスタどうしをマージ
手順3: 同クラスタ内記述子の物性値を標準化・平均化
→合成記述子の作成

相関が高い=記述子間の距離が近い

**クラスタ間の相関が高い**

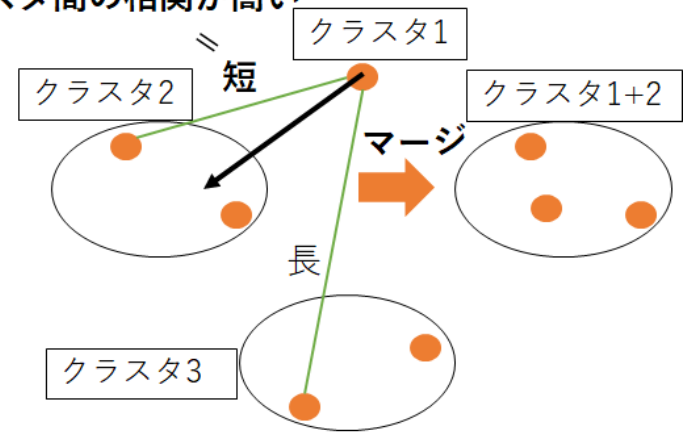


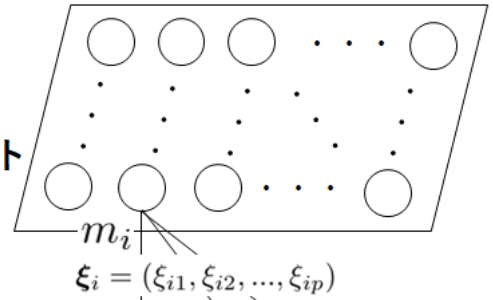

図10　最長距離法のイメージ

# 4.3 SOMによる反応式のクラスタリング



1. はじめに
2. 有機合成分野と情報分野の関わり
4. 提案手法
5. 実験結果並びに考察
6. おわりに

自己組織化マップ (Self-Organizing Map: SOM) を用いて，多次元の特徴ベクトルを 2 次元上に可視化し，クラスタリングを行う




→n個の反応式，p種の記述子からなる特徴ベクトル

ユニット $m_i (i = 1, 2, ..., k)$

ユニット $i$ の重心：$\boldsymbol{r}_i = (r_{i1}, r_{i2})$

ユニットの重みベクトル：$\boldsymbol{\xi}_i = (\xi_{i1}, \xi_{i2}, ..., \xi_{ip})$

ランダムに初期化

$\boldsymbol{x}_j$ と $\boldsymbol{\xi}_i$ のユークリッド距離: $||\boldsymbol{x}_j - \boldsymbol{\xi}_i||$

手順 (1): $||\boldsymbol{x}_j - \boldsymbol{\xi}_i||$ が最小となる勝者ユニット $m_c$ を求める

手順 (2): $\boldsymbol{x}_j$ に近くなるように $m_c$ の重みを更新

$$\boldsymbol{\xi}_i = \boldsymbol{\xi}_i + \alpha(t)(\boldsymbol{x}_j - \boldsymbol{\xi}_i)$$

→ $\alpha(t)$： 学習率係数 (学習回数 $t$ の増加で減少)

手順 (3): $m_c$ の付近の $m_i$ も多少 $\boldsymbol{x}_j$ に近づくように更新

$$\boldsymbol{\xi}_i = \boldsymbol{\xi}_i + h(t)(\boldsymbol{x}_j - \boldsymbol{\xi}_i)$$

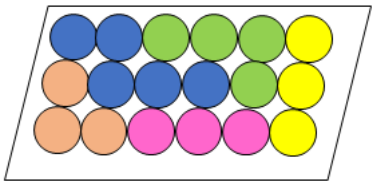
→ h(t)： 近傍関数 (勝者ユニットから離れるほど影響力が弱まる)

$$h(t) = \alpha(t) \exp\left[\frac{-||\boldsymbol{r}_c - \boldsymbol{r}_i||}{2\sigma^2(t)}\right] \quad (\sigma^2(t)：\text{調整関数})$$

手順 (4): $\boldsymbol{x}_{j+1}$ から $\boldsymbol{x}_n$ まで 1,2 を繰り返す

指定した学習回数に達するまで，(1)～(4)を繰り返す

→最終的に確定した各勝者ユニットに反応式の特徴ベクトルをマッピング



**各ユニットを色分けでクラスタリング**

→ 各ユニット間のマージ: $\boldsymbol{\xi}_i$ どうしのユークリッド距離

各クラスタ間のマージ: ウォード法

図10　各ユニットのクラスタリング

# 5.1 数値実験の概要 (ターゲット・EC 反応式)



1. はじめに
2. 有機合成分野と情報分野の関わり
4. 提案手法
5. 実験結果並びに考察
6. おわりに

## リボースのエステル化反応式と EC3.1.1 反応式を比較する

**ターゲット：リボースのエステル化反応(モルヌピラビル生成の1ステップ目)**

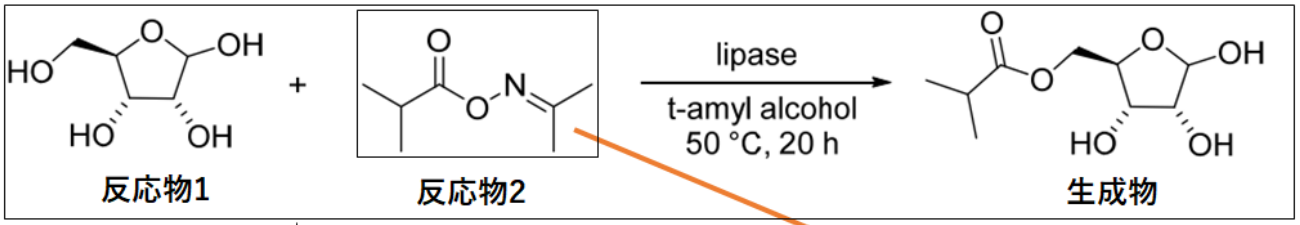
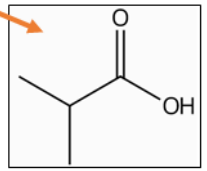



図10　ターゲット反応式

**スクリーニングでEC3.1.1.3**
**(トリアシルグリセロールリパーゼ)を選択**

**→SOMでEC 3.1.1.3がターゲット付近に**
**位置するかで，提案手法を評価**




**対象とする反応は通常起こりえない**
**←を初めは使用**
**収率向上のため等価体(反応物2)を選択**

**置き換えた反応式=ターゲット1**
**置き換え前の反応式=ターゲット2**

**比較対象のEC反応式： EC3.1.1反応式113種類**

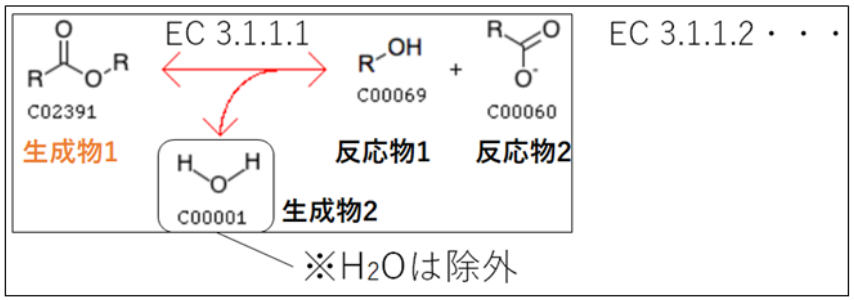



図11　EC3.1.1反応式

**EC反応式の条件：**

**エステル化**
**=加水分解酵素(EC 3)** 認識できた場合を仮定

**加水分解=可逆的** 左方向(エステル化)で比較

**$H_2O$を除外(変化度合いを合わせるため)**

# 5.1 数値実験の概要 (動画)



数値実験のデモ動画を提示します.

1. はじめに
2. 有機合成分野と情報分野の関わり
4. 提案手法
5. 実験結果並びに考察
6. おわりに

# 5.2 実験結果 (クラスタリングによる次元削減)



1. はじめに
2. 有機合成分野と情報分野の関わり
4. 提案手法
5. 実験結果並びに考察
6. おわりに

18 個のクラスタが形成され，80 次元の特徴ベクトルとなった.

**表4　クラスタ番号と所属する記述子の対応表**

| 0 | 0 | 1 | 1 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Kappa2 | Kappa3 | fr_Al_COO | fr_COO | NumVale | Chi0n | Chi0v | Chi1n | Chi1v | Chi2n | Chi2v | Kappa1 | LabuteAS | SMR_VSA | SlogP_VS | NumRota | MolMR |
| 3 | 3 | 4 | 4 | 5 | 5 | 6 | 6 | 6 | 7 | 7 | 8 | 8 | 8 | 9 | 9 | 9 |
| NumAliph | RingCoun | FpDensit | FpDensit | SMR_VSA | SlogP_VS | SMR_VSA | VSA_ESta | fr_C_O | NumAliph | NumSatu | fr_Ar_OH | fr_phenol | fr_pheno | fr_alkyl_h | fr_ketone | fr_lactone |
| 10 | 11 | 12 | 12 | 13 | 14 | 14 | 14 | 15 | 15 | 16 | 16 | 16 | 16 | 17 | 17 | 17 |
| fr_ester | fr_ether | MaxEStat | MaxAbsE | NumSatu | fr_NH1 | fr_NH2 | fr_amide | VSA_ESta | fr_allylic_ | VSA_ESta | NumAron | NumAron | fr_benzen | MolWt | HeavyAtc | ExactMolWt |
| 17 | 17 | 17 | 17 | 17 | 17 | 17 | 18 | 18 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 |
| Chi0 | Chi1 | Chi3n | Chi3v | Chi4n | Chi4v | HeavyAtc | SMR_VSA | TPSA | NOCount | NHOHCo | EState_V | EState_V | NumHete | fr_COO2 | FractionC | VSA_EState4 |
| 26 | 27 | 28 | 29 | 30 | 31 | 32 | 33 | 34 | 35 | 36 | 36 | 36 | 36 | 36 | 37 | 38 |
| VSA_ESta | VSA_ESta | NumHDo | fr_bicycli | fr_C_O_n | fr_metho | fr_Ar_CO | VSA_ESta | SlogP_VS | fr_unbrch | SlogP_VS | NumAliph | NumSatu | fr_NH0 | fr_piperd | EState_V | fr_Al_OH |
| 38 | 39 | 40 | 41 | 42 | 43 | 44 | 45 | 46 | 47 | 48 | 49 | 50 | 51 | 52 | 53 | 54 |
| fr_Al_OH | VSA_ESta | Ipc | PEOE_VS | SlogP_VS | PEOE_VS | HallKierA | PEOE_VS | EState_V | PEOE_VS | fr_ArN | PEOE_VS | EState_V | SMR_VSA | EState_V | PEOE_VS | EState_VSA10 |
| 55 | 56 | 57 | 58 | 59 | 60 | 61 | 62 | 63 | 64 | 65 | 66 | 67 | 68 | 69 | 70 | 71 |
| EState_V | PEOE_VS | qed | VSA_ESta | PEOE_VS | fr_aldehy | EState_V | FpDensit | MinAbsE | SlogP_VS | VSA_ESta | EState_V | BalabanJ | PEOE_VS | PEOE_VS | PEOE_VS | SMR_VSA6 |
| 72 | 73 | 74 | 75 | 76 | 77 | 78 | 79 | | | | | | | | | |
| SlogP_VS | PEOE_VS | BertzCT | PEOE_VS | SMR_VSA | MolLogP | NumHAc | MinEStateIndex | | | | | | | | | |

**表5　次元削減後の特徴ベクトル(clusterX: 合成記述子)**

| | cluster0 | cluster1 | cluster2 | cluster3 | cluster4 | cluster5 | cluster6 | cluster7 | cluster8 | cluster9 | ... | PEOE_VSA10 | SMR_VSA6 | SlogP_VSA10 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| T | -0.358029 | -0.842985 | 0.204247 | 4.571328 | 0.622585 | 0.207469 | -1.696273 | 2.341995 | 0.173683 | -0.133043 | ... | 0.796801 | 0.043121 | -0.1269 |
| 33 | -0.223667 | -0.842985 | 0.219299 | -0.103504 | 0.568775 | 0.207469 | 0.045521 | -0.141173 | 0.173683 | -0.133043 | ... | 0.796801 | 0.043121 | -0.1269 |
| 6 | 3.495807 | -0.842985 | 0.161480 | -0.103504 | -0.360529 | 0.207469 | 0.004312 | -0.141173 | 0.173683 | -0.133043 | ... | 0.046341 | 0.043121 | -0.1269 |
| 1 | 3.518113 | 1.079686 | 0.207055 | -0.103504 | -0.497256 | 0.207469 | 0.009803 | -0.141173 | 0.173683 | -0.133043 | ... | 0.046341 | 0.043121 | -0.1269 |
| 7_8 | -0.214983 | -0.842985 | 0.219101 | -0.103504 | 0.596503 | 0.207469 | 0.036318 | -0.141173 | 0.173683 | -0.133043 | ... | 0.796801 | 0.043121 | -0.1269 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 106.1 | -0.233937 | -0.842985 | 0.213692 | -0.103504 | 0.310413 | 0.207469 | 0.083963 | -0.141173 | 0.173683 | -0.133043 | ... | -0.646994 | 0.043121 | -0.1269 |
| 113 | -0.251825 | -0.842985 | 0.217973 | -0.103504 | 0.596561 | 0.207469 | 0.012651 | -0.141173 | 0.173683 | -0.133043 | ... | 0.046341 | 0.043121 | -0.1269 |
| 112 | -0.182730 | -0.842985 | 0.219101 | -0.103504 | 0.501004 | 0.207469 | 0.031504 | -0.141173 | 0.173683 | -0.133043 | ... | 0.046341 | 0.043121 | -0.1269 |
| 111 | -0.280504 | 1.079686 | 0.215345 | -0.103504 | -0.970931 | 0.207469 | 0.055048 | -0.141173 | 0.173683 | -0.133043 | ... | -1.333272 | 0.043121 | -0.1269 |
| 118 | -0.266600 | 1.079686 | 0.223487 | -0.103504 | -1.377476 | 0.207469 | 0.075928 | -0.141173 | 0.173683 | -0.133043 | ... | 0.046341 | 0.043121 | -0.1269 |

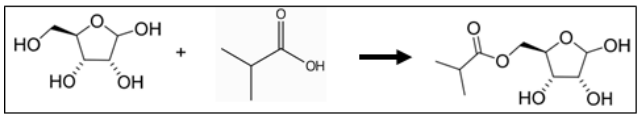
# 5.2 実験結果(SOMによる反応式クラスタリング)



1. はじめに
2. 有機合成分野と情報分野の関わり
4. 提案手法
5. 実験結果並びに考察
6. おわりに

R 言語の som ライブラリを元に作られたプログラムを使用し，反応式のクラスタリングを行った.

ターゲット1



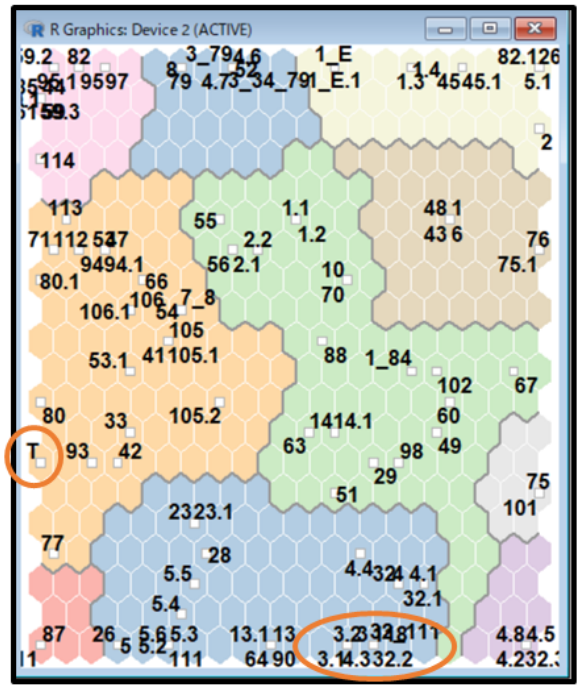
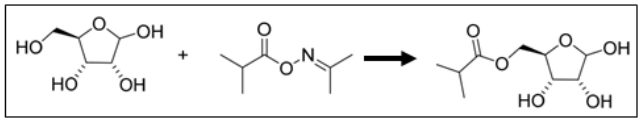



図11　SOMの結果(ターゲット1)

ターゲット2



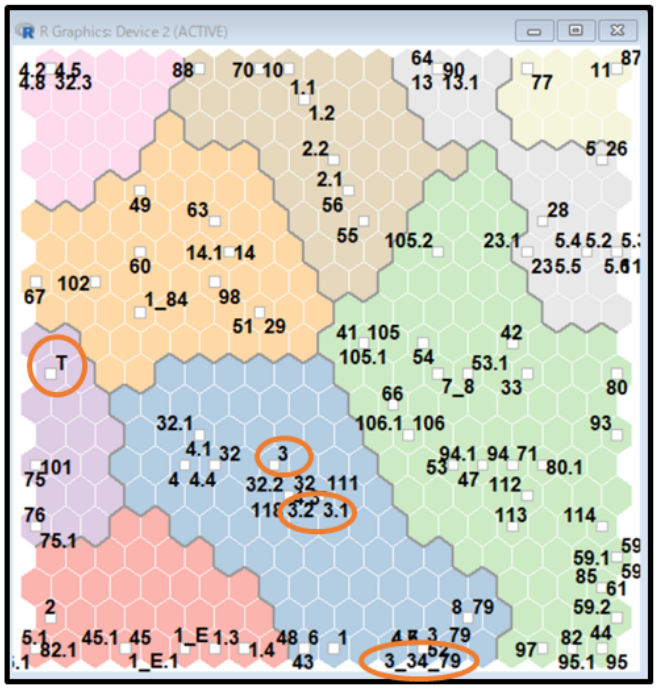



図12　SOMの結果(ターゲット2)

**ユニット数:**
**400(20×20)**

**学習回数:**
**1,000(1段階目)**
**→大まかな順序付け**

**200,000(2段階目)**
**→収束段階**

**T:　ターゲット**

**アンダーバー:**
**EC番号重複**

**ピリオド:**
**EC代表反応式が複数ある場合の区別**

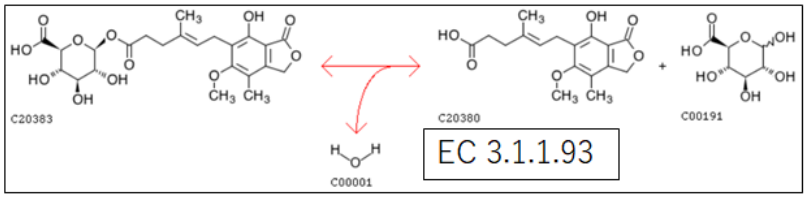
# 5.2 実験の考察



1. はじめに
2. 有機合成分野と情報分野の関わり
4. 提案手法
5. 実験結果並びに考察
6. おわりに

同クラスタ内でターゲット (T) の付近にある EC 反応式に，いくつかの構造的共通点が見られたが，予測されるべき EC3.1.1.3 は異なるクラスタに属していた.

**【結果】**
**・ターゲット1の付近にあるEC 3.1.1.93の反応物にはグルコース(リボースと同様に糖類)の構造の一部が含まれていた.**




**【予測が上手くいかなかった原因の考察】**
**(1)反応式中の係数を反映していない**
**→化合物の比率も考慮し，より化学的な変化を捉える**

**(2)特性値変化量では捉え切れない要因が多く影響した**
**(ターゲット合成するための試薬や溶媒，実験環境など)**
**→特性値変化量以外の要因も特徴に含めて再検証**

**(3)合成変数にした際に重要な変数の影響力を弱めた**
**→クラスタ内の記述子の重要度を評価して，重みづけする方法を提案**

1. はじめに
2. 有機合成分野と情報分野の関わり
4. 提案手法
5. 実験結果並びに考察
6. おわりに

## おわりに

化学反応の効率化や環境面から，化学合成時に酵素の生体触媒を用いる機会が増え，最適な酵素候補の予測が重要となってきている．本研究では，酵素候補の探索時間短縮を目的としたEC番号の予測手法を提案した．結果として，予測されるべき酵素候補を予測することには至らなかった．

## 今後の課題

・化学反応時の特徴変化をより詳細に捉えるため，重要な記述子を残しつつ，さらに次元削減していく手法の検討

・EC反応式のクラス分類問題に着目した記述子選択手法の提案
→最も精度よく分類できる記述子の組み合わせを特定し，ターゲット反応式に対する酵素予測において再検証