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

1815070 武藤 克弥

富山県立大学 電子・情報工学科

February 4, 2022

背景

1.1 本研究の背景

近年, ケモインフォマティクスと呼ばれる，化学に関するデータを情報技術を用いて分析する分野が，新型コロナウイルスによる新薬開発の必要性に伴い，需要が高まっている．

合成時の化学反応をケモインフォマティクスや機械学習を用いて予測する研究が増加している一方で，環境面と効率性を考慮して，目的物を合成する際に酵素を触媒として用いることが世界の風潮となってきている．

しかし，酵素に関する情報は生物分野に関わっており，有機合成の知識だけでは，目的の合成反応に対して何の酵素を用いるべきか判断が難しい場合がある．

1.2 本研究の目的

有機合成に用いる最適な酵素の目途をつけるシステムがあれば，有機合成の研究者自身で，酵素候補を探索し，次の実験のステップをスムーズ進めることができる．そのため，本研究では，与えられた反応式に対して，用いるべき酵素を予測するシステムの開発を目的とする

2.1 有機合成と情報技術

有機合成の発展

- 病の治療として古くから，天然の有機化合物が用いられてきた.
- 天然の化合物から，薬効成分のみを取り出したり，天然の化合物を人工的に作り出してきた歴史とともに，有機合成分野は発展していった.
- コンピュータの発達によって，実験データがデータベースに蓄積されるようになる
- 化学の現象をコンピュータ上で上手く表現することによって，高速データ処理，反応設計・予測が容易になってきた.

2.2 酵素と EC 番号

酵素について

- 生体内で重要な化学反応を触媒 (= タンパク質 = 生体触媒)
- ある化合物を変化させるが自分は変化しない→何回でも使える
- 基質特異性という，特定の化合物 (基質) のみ作用するという性質がある→鍵 (基質)，鍵穴 (酵素) に例えられる
- メリット 1：特定の基質にしか作用しないが，化学反応をより早く進める
- メリット 2：化学触媒は高温や高圧といった条件で使用⇔生体触媒 (酵素) は常温→環境にやさしい

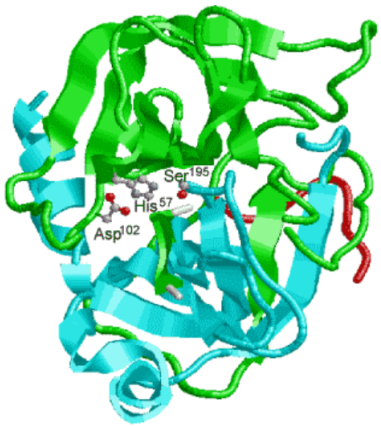


図 1: 酵素の構造の例

2.2 酵素と EC 番号 2

EC 番号 (Enzyme Commission numbers)

- 酵素を 4 組の数字の組み合わせからなる番号で分類したもの (例 EC ○. ○. ○. ○)
- １番目の数字はどの化学反応の触媒となるかで 7 つに分類 (例 1 = 酸化還元酵素，3 = 加水分解酵素，7 = 輸送酵素
- 2 番目，3 番目の数字はどの化学結合・基質 (化合物) に作用するかで分類　　→ 4 番目で 3 番目に属する酵素を指す

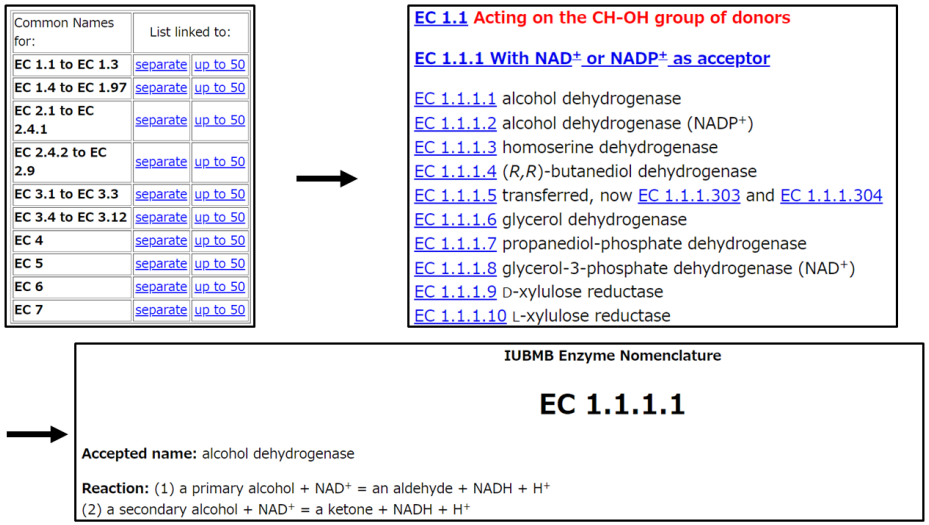


図 2: データベース (Enzyme Nomenclature) で EC.1.1.1.1 を参照した場合

2.3 化学データベース

使用するデータベース

1. Kyoto Encyclopedia of Genes and Genomes(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，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，その番号の酵素を使った反応式，反応式 ID，KEGG と PubChem との化合物 ID 対応表を KEGG 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構造を記した Mol ファイル (text) を API で取得する

2.3 化学データベース2

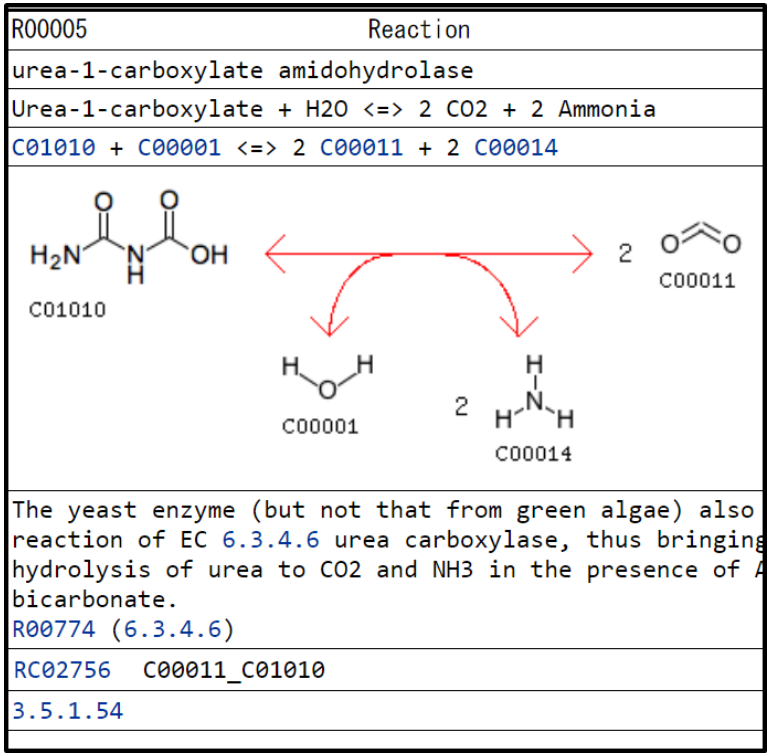


図 3: KEGG の例

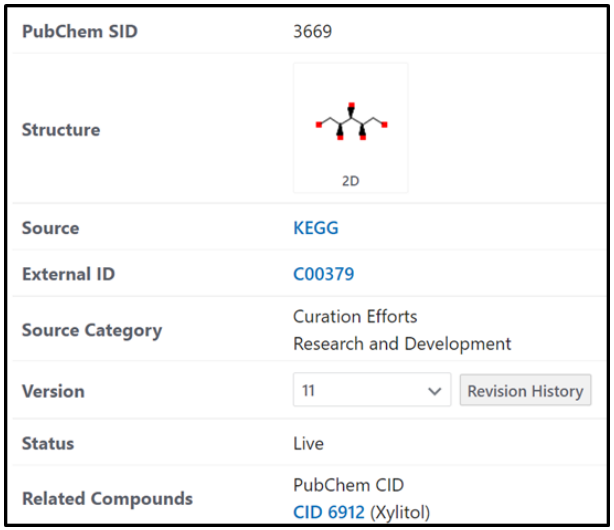


図 4: PubChem の例

3.1 化学データベースからの情報抽出

Web サイトからの情報収集

- API(Application Programming Interface: API)
  Web サイトの管理者などが，特定のデータを自動的に取得できるように定められたプログラミング形式

  →定められた関数・クラスメソッド，引数などを指定することで，目的のデータを簡単に取得できる

KEGG・PubChem の API

- KEGG API
  - 「http://rest.kegg.jp/」の下層に入っているデータの url を取得
  - それぞれのプログラミング言語における，url の中身を取得するメソッドによってデータを得る
- PubChempy
  - python に実装されている PubChem 内のデータを取得するライブラリ
  - 引数に化合物名や ID などを指定してデータを得る

3.2 化学物の構造表現法と EC 番号予測手法

ケモインフォマティクスでの化学構造表現

化学構造をコンピュータ上で扱うための，様々な表現法がある．

1. SMILES： 化合物を文字列で表現 (元素 C,O,N,H を大文字，2 重結合は =，3 重結合は#など)
2. フィンガープリント： 化合物の特徴的な構造をビット列で表現
→構造の類似度比較で使用

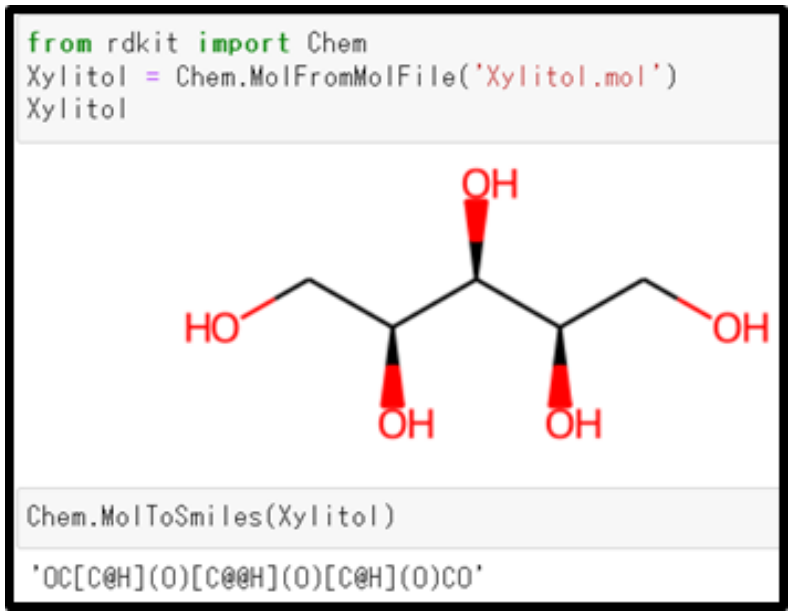


図 5: キシリトールの構造式と SMILES

化合物の数値化

機械学習で様々な予測をするためには，数値化が必須.

- 物性・化学特性値： 化合物の物理的・化学的特性を数値化
  →化合物の特徴量となる (例：分子量，電荷密度，溶解度など)

  ※特徴量を計算する特性値のことを記述子と呼ぶ
- フィンガープリントも化合物の構造的特徴を表現する記述子
- RDKit： 構造式・SMILES，フィンガープリントなどを扱う Python ライブラリ，208 種類の記述子を備えている.

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

```
152.14600000000002
```

```
Descriptors.TPSA(Xylitol)
```

```
101.15
```

図 6: キシリトールの MolWt と TPSA を RDKit で計算した例

3.2 化学物の構造表現法と EC 番号予測手法 3

EC 番号予測

- EC 番号にはその酵素を使った代表的な反応式が数種類登録されている
  →代表的な反応式： 生物の体内で起こっている反応
- EC 番号の反応式約 7 千種に対して，ある特徴量を用いて機械学習させ，この反応式はどの EC 番号かを再分類 (学習データ EC 番号，テスト用 EC 番号に分ける)
  →自前で考えた特徴量に対して，EC 番号の分類精度を検証する研究が多くなされている.

フィンガープリントを用いた EC 番号予測

次のような反応式： 反応物 1 + 反応物 2 →生成物 1 + 生成物 2 を考える.

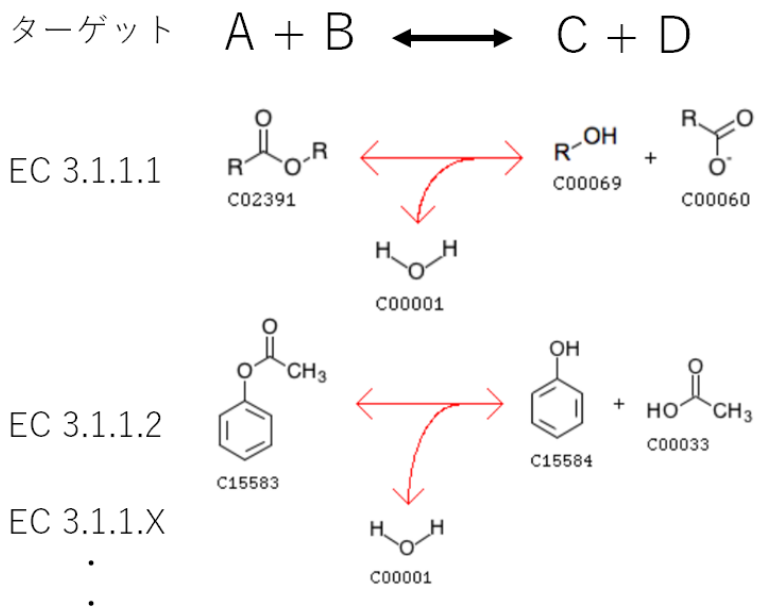
1. 反応式に出てくる化合物 (反応物・生成物) の構造特性を示すフィンガープリント FP を計算
2. 反応差分フィンガープリント RFP を計算

4. 提案手法

4.1 EC 番号の予測

ターゲットなる反応式に対し，生体触媒として最適な酵素の EC 番号を予測する

- EC 番号の反応式 (EC 反応式) は自然界で見られる，酵素を触媒として使った代表的な反応
- ターゲット反応式と EC 反応式を比較し，類似した特徴がある
→ターゲット反応式でも同じ酵素を使えば目的の生成物を作れる可能性がある．(類似性の概念)



4. 提案手法 2

4.1 反応式比較の基準

反応物から生成物変化時の，記述子の特性値変化量で比較

- 特性値の変化 = 生体触媒を加えることで生じると仮定
- ターゲット反応式と EC 反応式の特性値変化が類似→その生体触媒をターゲットで使えば目的の反応が起こると予測

(各反応) 反応物と生成物の個数をそれぞれ 2 個の場合を仮定．反応物の特性値を RT_i，生成物の物性値を PD_i として，各反応式の特性値変化量 DF を定義.

$$DF = (PD_1 + PD_2) - (RT_1 + RT_2) \tag{2}$$

n 個の記述子に対して同じ計算をするため，各反応式は n 種類の特性値変化を持った，n 次元の特徴ベクトル $\boldsymbol{DF}$ となる.

$$\boldsymbol{DF} = (DF_1, DF2, \cdots, DF_n) \tag{3}$$

4. 提案手法 (3)

各化合物の特徴抽出

- ターゲット & EC 番号化合物に対して 208 項目の物性値を算出
  →特徴量 208

	MaxEStateIndex	MinEStateIndex	MaxAbsEStateIndex	MinAbsEStateIndex	qed	MolWt	HeavyAtomMolWt	ExactMolWt
右辺第一項	11.084849	-1.428007	11.084849	0.174306	0.507460	220.221	204.093	220.094688
3.1.1.33	10.486662	-1.607446	10.486662	0.294306	0.376897	222.193	208.081	222.073953
3.1.1.6	9.652778	-0.136574	9.652778	0.134259	0.392730	59.044	56.020	59.013304
3.1.1.1	9.763889	-0.111111	9.763889	0.074074	0.386221	44.009	44.009	43.989829
3.1.1.7 3.1.1.8	10.300428	-0.200509	10.300428	0.200509	0.421378	146.210	130.082	146.117555
...	...	...	...	...	...	...	...	...
3.1.1.106	12.257782	-5.347078	12.257782	0.051001	0.107483	601.355	576.155	601.082238
3.1.1.113	9.821528	-0.210648	9.821528	0.210648	0.437851	88.106	80.042	88.052429
3.1.1.112	10.203017	-0.184583	10.203017	0.184583	0.542210	130.187	116.075	130.099380
3.1.1.111	11.239615	-4.515564	11.239615	0.350225	0.369149	286.153	273.049	286.032793
3.1.1.118	10.737761	-4.606142	10.737761	0.221991	0.598934	226.077	219.021	225.987854

図 9: 各化合物の 208 種の物性値 (特徴量)

提案手法 3

4.2 凝集型クラスタリングによる次元削減

反応変化の特徴を捉えるには，なるべく多く説明変数を用いる

→無駄に多すぎると，多重共線性と次元の呪いが発生する．

1. 多重共線性： 説明変数同士で高い相関関係があると発生→同じ説明変数があると同義 = 過学習・分類精度低下の原因
2. 次元の呪い： 説明変数が多すぎることで生じる (同様に過学習の原因)

なるべく化学変化の特徴を多く説明し，必要量だけの記述子の組み合わせを考える

多重共線性・次元の呪いの対策

変数同士の相関が高い→ (通常) どっちか徐外
△特徴説明に重要な変数を誤消去する可能性

→クラスタリングで相関の高い変数どうしを 1 つにまとめて合成変数を作る

4.2 階層的クラスタリングと主成分分析による特徴選択 2

クラスタリング

- 特定の基準に従わず，類似するもの同士でクラスタを形成し，分類する教師無し学習
- (階層的) 凝集性クラスタリング： 距離の近いデータ同士をまとめてクラスタを形成した後，クラスタ同士も段階的にまとめていき，最終的に 1 つのクラスタを形成する手法

→クラスタ間距離が最短のクラスタから結合 (定義は手法による)

クラスタ C_1，C_2 に属するデータの集合：$\mathbf{x_1}$,$\mathbf{x_2}$
$\mathbf{x_1}$ と $\mathbf{x_2}$ の距離 $d(\mathbf{x_1}, \mathbf{x_2})$，クラスタ間の距離：$d(C_1, C_2)$

最長距離法

クラスタ内のデータで最も遠いもの同士の距離をクラスタ間距離としたもの．外れ値には弱いが，クラスタサイズが一定になる傾向がある．

$$d(C_1, C_2) = \max_{\mathbf{x_1} \in C_1, \mathbf{x_2} \in C_2} \{d(\mathbf{x_1}, \mathbf{x_2})\} \tag{4}$$

4.2 凝集型クラスタリングによる次元削減 3

相関係数によるクラスタリング

相関係数が 0.9 以上の記述子同士をクラスタリングしていく.

1. 記述子 i, j 間の相関係数を s_{ij} としたとき，逆数を取った，$1/s_{ij}$ を記述子間の距離として最長距離法を用いる.
2. 初めの 1 つ 1 つのクラスタリングでは，$s_{ij} >= 0.9$ すなわち，$1/s_{ij} <= 1/0.9 \approx 1.11$ の距離となる記述子間でクラスタリング
3. クラスタ同士になったときは，最長距離法のクラスタ間距離の中で最短のクラスタ同士でマージ

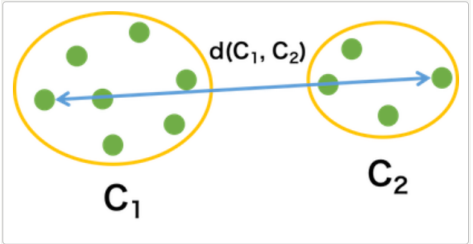


図 10: 完全連結法のイメージ

自己組織化マップ (Self-Organizing Map: SOM)

多次元データを低次元にマッピングし，可視化する非階層的クラスタリング手法.

n 個の p 次元観測ベクトル $\boldsymbol{x}_j = (x_{j1}, x_{j2}, ..., x_{jp}), (j = 1, 2, ..., n)$ を，ユニット $m_i (i = 1, 2, ..., k)$ からなる 2 次元平面上に写像する. 各ユニットの重心：$\boldsymbol{r}_i = (r_{i1}, r_{i2})$，各ユニットの重みベクトル：$\boldsymbol{\xi}_i = (\xi_{i1}, \xi_{i2}, ..., \xi_{ip})$，$\boldsymbol{x}_j$，$m_i$ をそれぞれ，入力層，出力層として，次の手順によって出力層を更新する.

1. $j = 1$ から n で，各 $\boldsymbol{x}_j$ に対してユークリッド距離 $||\boldsymbol{x}_j - \boldsymbol{\xi}_i||, (i = 1, 2, ..., k)$ を求め，最小値にする $\boldsymbol{\xi}_i$ を $\boldsymbol{\xi}_c$ と置く. この $\boldsymbol{\xi}_c$ を持つユニットを勝者ユニット m_c と呼ぶ.
2. 勝者ユニット m_c とその近傍のユニットが持つ重みベクトルを以下のように更新する.

$$\begin{cases} \boldsymbol{\xi}_i \leftarrow \boldsymbol{\xi}_i + h(t)\{\boldsymbol{x}_j - \boldsymbol{\xi}_i\} & i \in N_c \\ \boldsymbol{\xi}_i \leftarrow \boldsymbol{\xi}_i & i \notin N_c \end{cases} \tag{5}$$

SOM　続き

このとき，h(t) は以下で定義される近傍関数である．ただし，$\alpha(t)$ を学習率係数 (学習回数を変数とした単調減少関数)，$\sigma^2(t)$ はユニット m_c の近傍領域 N_c の散らばりに関する調整関数とする．

$$h(t) = \alpha(t) \exp\left[\frac{-||\boldsymbol{r}_c - \boldsymbol{r}_i||}{2\sigma^2(t)}\right] \tag{6}$$

3. j で更新した $\boldsymbol{\xi}_i$ を保存し，$\boldsymbol{x}_{j+1}$ から $\boldsymbol{x}_n$ まで 1,2 を繰り返す．
4. 3 までを 1 回の学習とし，指定した回数まで学習を行う
5. 学習後，ユークリッド距離 $min||\boldsymbol{x}_j - \boldsymbol{\xi}_i||$ を満たす $\boldsymbol{\xi}_c$ を持つユニット m_c に $\boldsymbol{x}_i$ をマッピングする

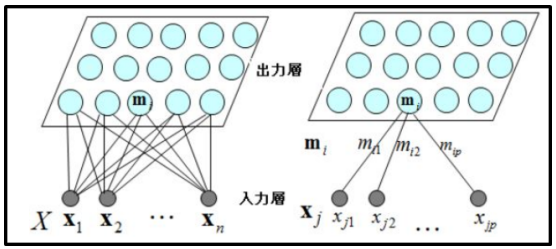


図 11: SOM のイメージ

5.1 数値実験の内容

KEGG と PubChem で取得したデータを整理する流れを作成

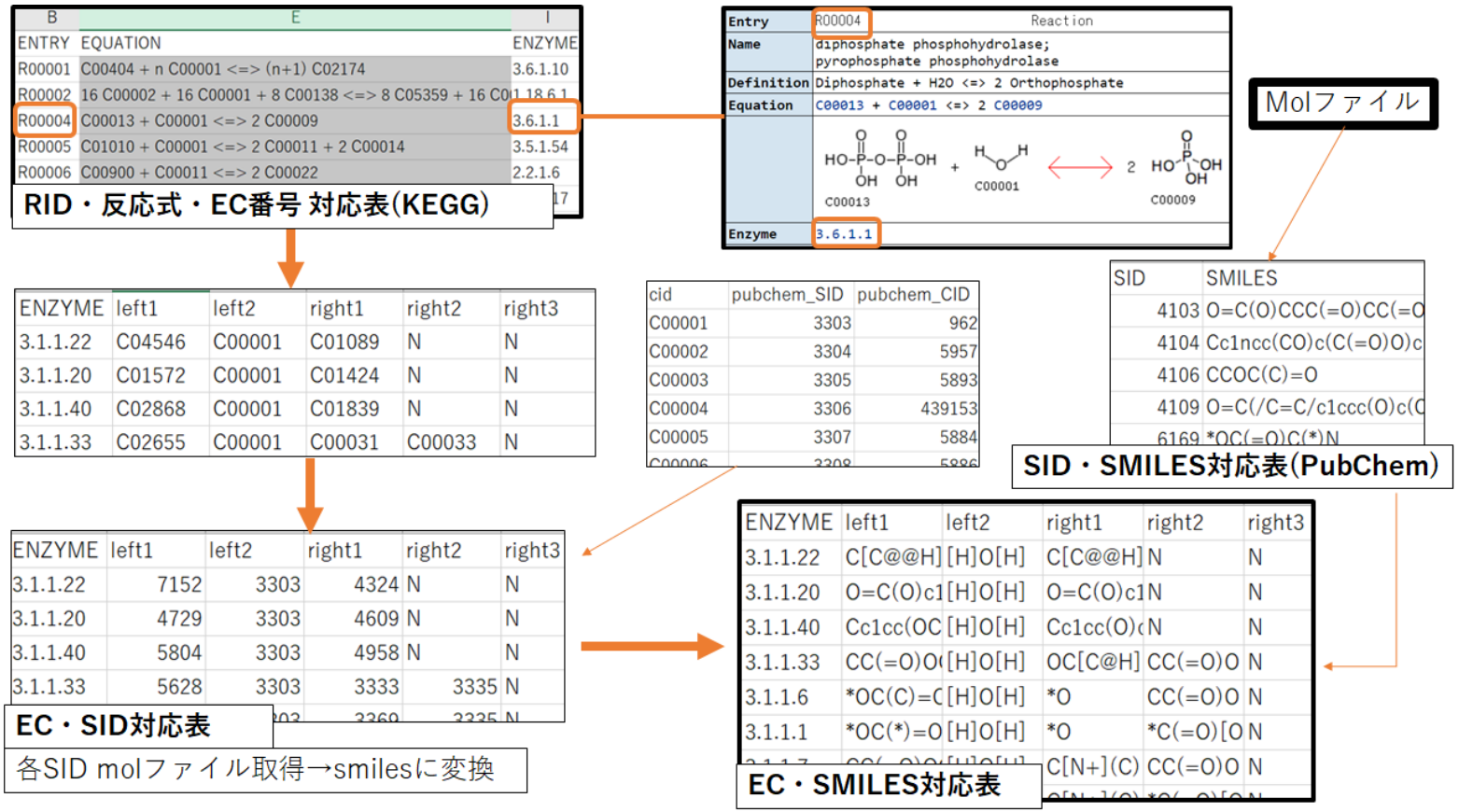


図 12: EC 番号・SMILES 対応表

5.1 数値実験の内容 2

比較対象となる EC 反応式

- 比較する EC 反応式を EC3.1.1 番号内に酵素 113 種類に絞る
  → (上位ディレクトリは予測できたものと仮定する)
- ターゲットの反応「エステル加水分解の逆反応 (エステル化反応)」
  → EC3.1.1 の加水分解酵素が適当
- 加水分解は可逆的な反応 = 反応物・生成物を入れ替えられる
  → EC3.1.1 反応式もエステル化する方向 (右辺を反応物，左辺を生成物とする) でターゲットと比較

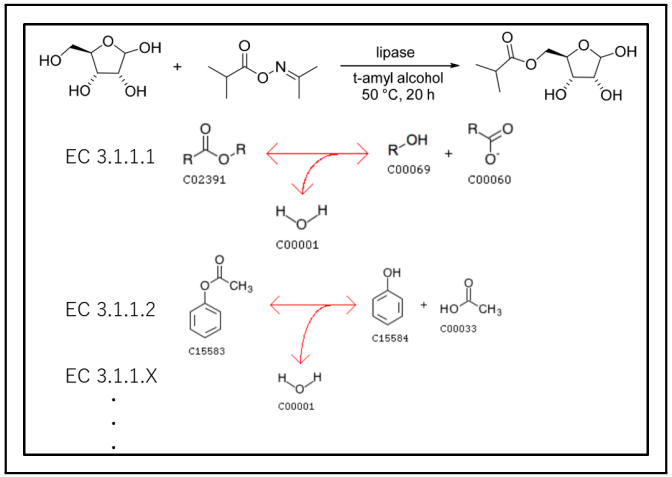


図 13: ターゲットと EC 反応式

5.2 実験結果と考察

0	0	0	0	0	0	1	1	2	2	2	2
PEOE_VSA6	EState_VSA4	EState_VSA8	VSA_EState7	fr_allylic_oxid	fr_unbrch_alkane	NumAliphat	RingCount	SMR_VSA4	SlogP_VSA4	NumAliphaticCa	NumSaturatedCarbo
2	2	3	3	4	4	4	4	4	4	4	4
fr_NH0	fr_piperdine	FpDensityMo	FpDensityMo	Chi1n	Chi1v	Kappa1	SMR_VSA5	SMR_VSA7	SlogP_VSA5	SlogP_VSA6	VSA_EState8
4	4	4	5	5	6	6	7	7	8	8	8
NumRotatal	MolLogP	MolMR	fr_Al_COO	fr_COO	PEOE_VSA11	SlogP_VSA7	BertzCT	HallKierAlph	VSA_EState6	NumAromaticCa	NumAromaticRings
8	9	10	10	10	11	11	12	12	13	14	14
fr_benzene	fr_bicyclic	PEOE_VSA1	PEOE_VSA10	PEOE_VSA14	VSA_EState1	fr_ether	NumAliphat	NumSaturat	fr_C_O_noCOO	NumHAcceptors	NumHeteroatoms
15	16	16	16	17	17	17	18	18	18	19	20
fr_ester	fr_alkyl_halid	fr_ketone	fr_lactone	SMR_VSA10	VSA_EState2	fr_C_O	fr_Ar_OH	fr_phenol	fr_phenol_noO	EState_VSA3	FractionCSP3
21	22	22	22	23	23	24	24	25	25	25	26
VSA_EState	SMR_VSA1	TPSA	NOCount	MaxEStateInde	MaxAbsEStateInde	NHOHCount	NumHDono	fr_NH1	fr_NH2	fr_amide	fr_COO2
27	28	29	30	30	30	30	30	30	30	30	30
fr_Ar_COO	NumSaturate	EState_VSA7	MolWt	HeavyAtomMol	ExactMolWt	NumValenc	Chi0	Chi0n	Chi0v	Chi1	Chi2n
30	30	30	30	30	30	30	31	32	33	33	34
Chi2v	Chi3n	Chi3v	Chi4n	Chi4v	LabuteASA	HeavyAtom	fr_ArN	fr_aldehyde	fr_Al_OH	fr_Al_OH_noTert	fr_methoxy
35	36	37	38	39	40	41	42	43	44	45	46
Ipc	SlogP_VSA1	SMR_VSA3	VSA_EState5	PEOE_VSA12	MinAbsEStateIndex	PEOE_VSA8	PEOE_VSA	EState_VSA	PEOE_VSA3	PEOE_VSA13	EState_VSA1
47	48	49	50	51	52	53	54	55	56	57	58
VSA_EState	EState_VSA2	qed	MinEStateInd	PEOE_VSA7	EState_VSA10	VSA_EState	Kappa3	EState_VSA	BalabanJ	FpDensityMorga	SlogP_VSA2
59	60	61	62	63	64	65					
Kappa2	SlogP_VSA10	PEOE_VSA2	EState_VSA5	VSA_EState3	SMR_VSA6	PEOE_VSA9					

図 14: 相関クラスタによる分類結果

5.2 実験結果と考察 2

	cluster0	cluster1	cluster2	cluster3	cluster4	cluster5	cluster6	cluster7	cluster8	fr_bicyclic	...	BalabanJ	FpDensityMorgan3	SlogP_VSA2
T	0.221492	4.470974	0.02782	0.468254	0.184763	-0.824063	0.031512	0.432294	0.075584	0.0	...	-4.679584	-1.600000	-10.209189
33	0.098625	-0.101231	0.02782	0.406077	0.220241	-0.824063	0.031010	-0.054880	0.075584	0.0	...	-2.896440	-1.666667	-10.213055
6	0.177826	-0.101231	0.02782	-0.630073	0.175482	-0.824063	0.031010	-0.263046	0.075584	0.0	...	-0.955661	-2.250000	-10.582719
1	0.305139	-0.101231	0.02782	-0.822865	0.387189	1.055450	0.031010	-0.272530	0.075584	0.0	...	-0.955661	-2.333333	-5.476192
7_8	0.103843	-0.101231	0.02782	0.186089	0.253276	-0.824063	0.031010	-0.131124	0.075584	0.0	...	-2.666912	-1.514286	-10.213055
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
106.1	0.303405	-0.101231	0.02782	0.148775	0.167701	-0.824063	0.824336	0.072527	0.075584	0.0	...	-2.822743	-1.824786	-10.213055
113	0.172470	-0.101231	0.02782	0.097692	0.257698	-0.824063	0.031010	-0.235343	0.075584	0.0	...	-1.604776	-1.666667	-10.213055
112	0.104844	-0.101231	0.02782	0.081002	0.252294	-0.824063	0.031010	-0.164011	0.075584	0.0	...	-2.424149	-1.555556	-10.213055
111	0.301378	-0.101231	0.02782	-1.116293	0.232535	1.055450	0.031010	-0.024475	0.067768	0.0	...	-2.773625	-2.590278	-10.213055
118	0.232751	-0.101231	0.02782	-1.532865	0.249663	1.055450	0.031010	-0.047983	0.075584	0.0	...	-2.512090	-2.821429	-10.213055

114 rows × 66 columns

図 15: 特徴量合成結果

5.2 実験結果と考察 3

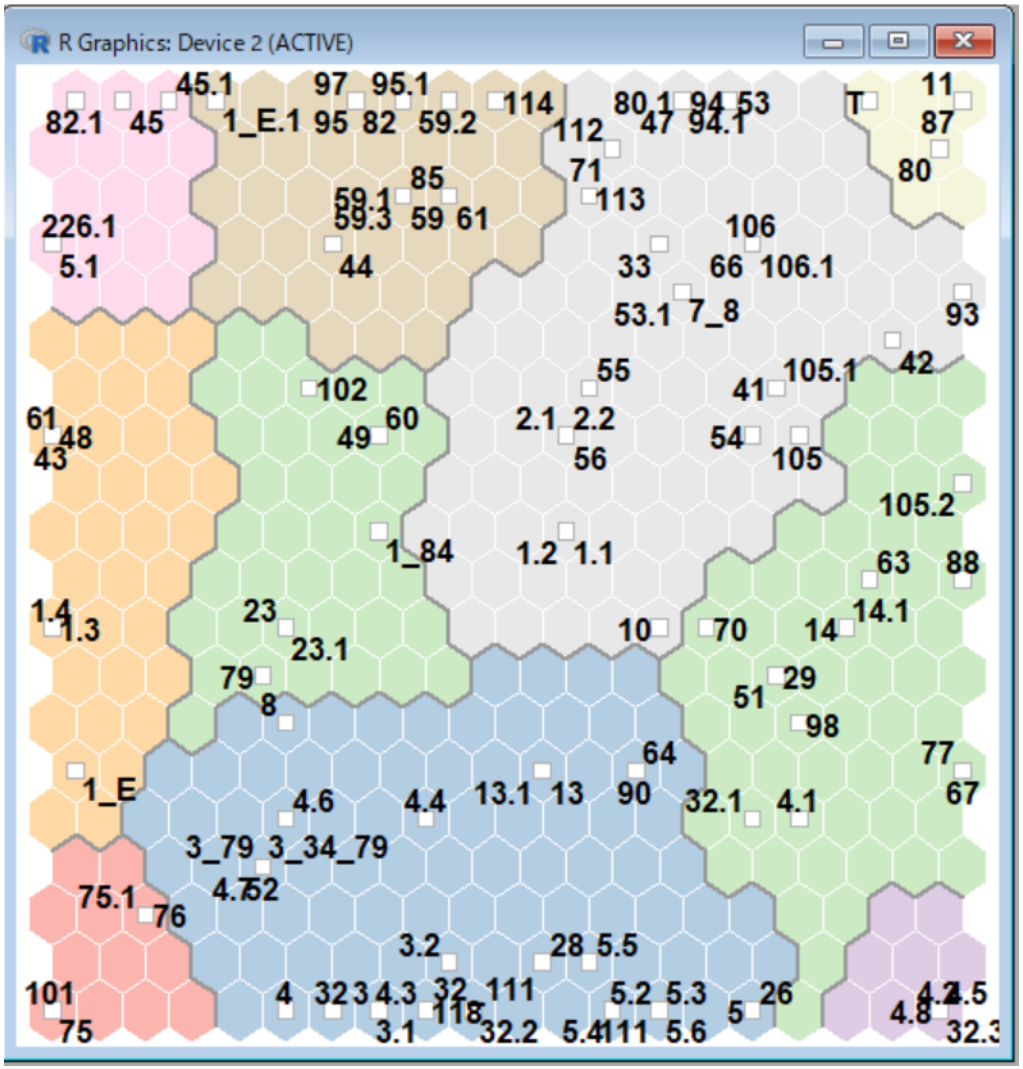


図 16: SOM の結果

おわりに

今後の課題

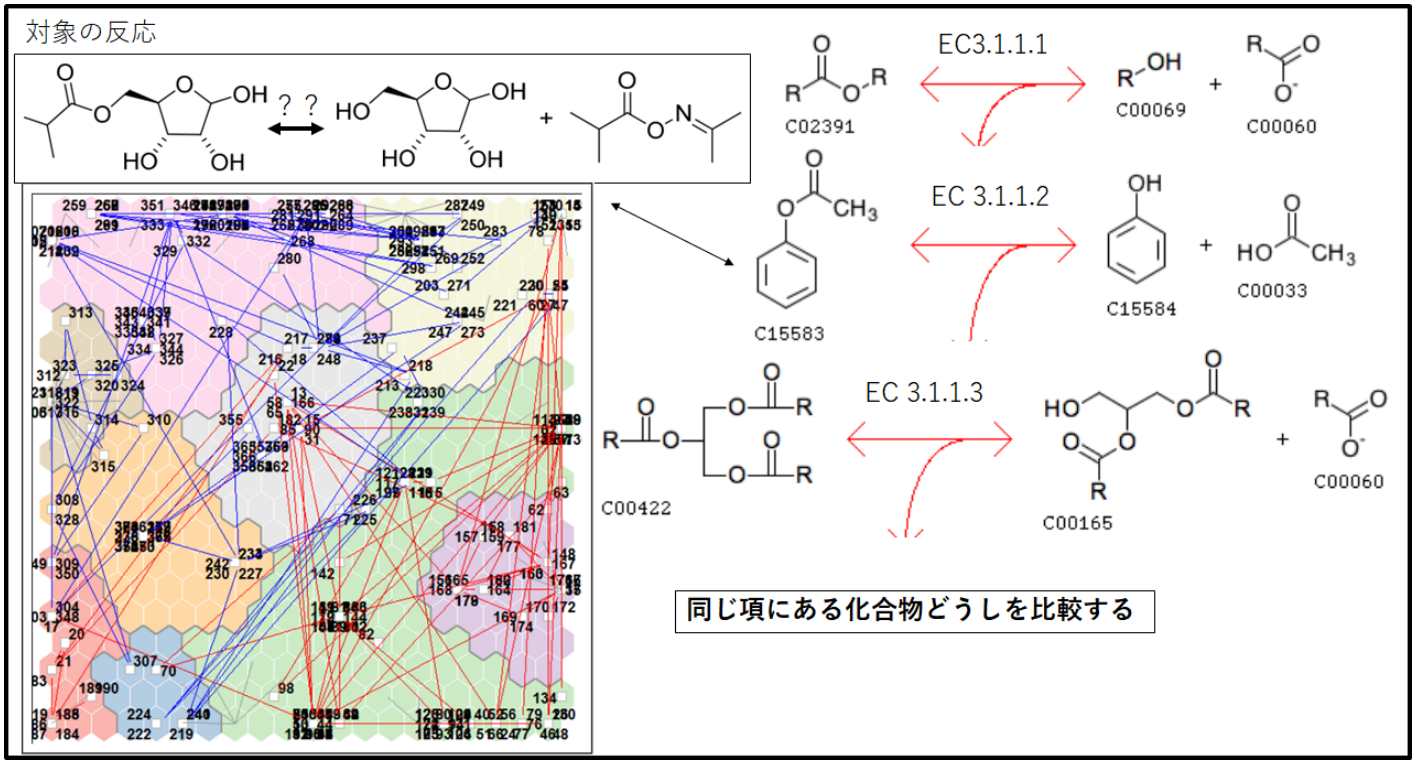


図 17: SOM のイメージ

システムの流れ

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