

背景
有機合成成分と情報分野の関わり
3. ケモインフォマティクスとテキストマイニング
提案手法
今後の予定

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

背景

2/20

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発展してきている。

新型コロナウイルスの流行等によって、創薬のための新規化合物開発のニーズが高まり、ケモインフォマティクスの需要は拡大している。

合成時の化学反応を予測するソフトウェアが増加している一方で、環境面とコストを考慮して、使用する触媒を酵素に置き換える動きが出てきている。

1.2 本研究の目的

- 化学反応が与えられたとき、効率の良い生体触媒として作用する酵素を予測する
- 酵素を分類した EC 番号 (Enzyme Commission numbers) に記載された反応式の化合物と、与えられた反応式の化合物を比較し、類似度が高い反応式の EC 番号を予測する

背景

有機合成分野と情報分野の関わり

3. ケモインフォマティクスとテキストマイニング

提案手法

今後の予定

有機合成と酵素

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)

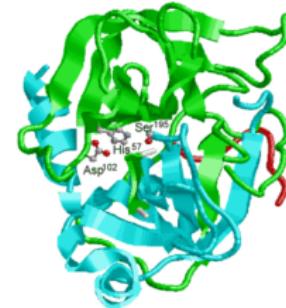


図 1: 酵素の構造

背景

有機合成分野と情報分野の関わり

3. ケモインフォマティクスとテキストマイニング

提案手法

今後の予定

2.1 有機合成とケモインフォマティクス

4/20

ケモインフォマティクス (化学情報学)

- 化学に関するデータを情報技術を使って収集・分析する学問
- 化学は観測された現象、実験から得られた経験などから新たなルール・規則を見出す
→機械学習などと相性がよい
- 情報分野の問題に適用できる内容が多い
例：化合物中の原子をノード、結合をエッジとみなす→グラフ理論
分子軌道の計算→数値解析などのシミュレーション

背景

有機合成分野と情報分野の関わり

3. ケモインフォマティクスとテキストマイニング

提案手法

今後の予定

2.2 酵素と EC 番号

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

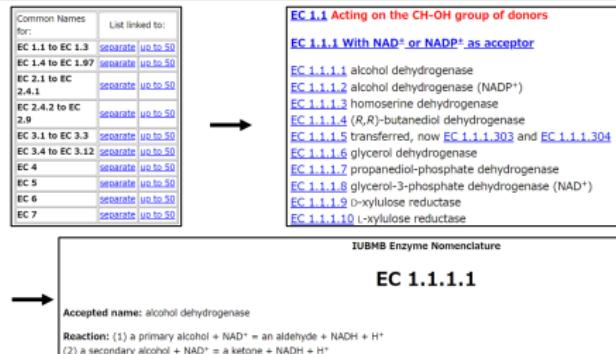


図 2: データベース (Enzyme Nomenclature) で EC.1.1.1.1 を参照した場合

背景

有機成分野と情報分野の関わり

3. ケモインフォマティクスとテキストマイニング

提案手法

今後の予定

2.3 化学データベース

6/20

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, 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, 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

背景

有機合成成分野と情報分野の関わり

3. ケモインフォマティクスとテキストマイニング

提案手法

今後の予定

2.3 化学データベース2

7/20

背景

有機合成分野と情報分野の関わり

3. ケモインフォマティクスとテキストマイニング

提案手法

今後の予定

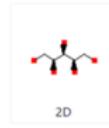
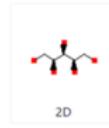
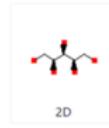
<table border="1"> <tr><td>Entry</td><td>EC 1.1.1.10</td><td>Enzyme</td></tr> <tr><td>Name</td><td>L-xylitol reductase; xylitol dehydrogenase (ambiguous)</td><td></td></tr> <tr><td>Class</td><td>Oxidoreductases; Acting on the CH-OH group of donors; with NAD⁺ or NADP⁺ as acceptor BRITE hierarchy</td><td></td></tr> <tr><td>Sysname</td><td>xylitol:NADP+ 4-oxidoreductase (L-xylulose-forming)</td><td></td></tr> <tr><td>Reaction(IUBMB)</td><td>xylitol + NADP+ = L-xylulose + NADPH + H⁺ [RN:R01904]</td><td></td></tr> <tr><td>Reaction(KEGG)</td><td>R01904 REACTION</td><td></td></tr> <tr><td>Substrate</td><td>xylitol [CPD:C00379]; NADP+ [CPD:C00006]</td><td></td></tr> <tr><td>Product</td><td>L-xylulose [CPD:C00312]; NADPH [CPD:C00005]; H⁺ [CPD:C00008]</td><td></td></tr> </table> <table border="1"> <tr><td>Entry</td><td>R01904</td><td>Reaction</td></tr> <tr><td>Name</td><td>Xylitol:NADP+ 4-oxidoreductase (L-xy)</td><td></td></tr> <tr><td>Definition</td><td>Xylitol + NADP+ => L-Xylulose + NADP</td><td></td></tr> <tr><td>Equation</td><td>C00379 + C00006 => C00312 + C00005</td><td></td></tr> <tr><td colspan="3"> </td></tr> <tr><td>Reaction class</td><td>RC00001 C00005_C00006 RC00102 C00312_C00379</td><td></td></tr> <tr><td>Enzyme</td><td>1.1.1.10</td><td></td></tr> </table>	Entry	EC 1.1.1.10	Enzyme	Name	L-xylitol reductase; xylitol dehydrogenase (ambiguous)		Class	Oxidoreductases; Acting on the CH-OH group of donors; with NAD ⁺ or NADP ⁺ as acceptor BRITE hierarchy		Sysname	xylitol:NADP+ 4-oxidoreductase (L-xylulose-forming)		Reaction(IUBMB)	xylitol + NADP+ = L-xylulose + NADPH + H ⁺ [RN:R01904]		Reaction(KEGG)	R01904 REACTION		Substrate	xylitol [CPD:C00379]; NADP+ [CPD:C00006]		Product	L-xylulose [CPD:C00312]; NADPH [CPD:C00005]; H ⁺ [CPD:C00008]		Entry	R01904	Reaction	Name	Xylitol:NADP+ 4-oxidoreductase (L-xy)		Definition	Xylitol + NADP+ => L-Xylulose + NADP		Equation	C00379 + C00006 => C00312 + C00005					Reaction class	RC00001 C00005_C00006 RC00102 C00312_C00379		Enzyme	1.1.1.10		<table border="1"> <tr><td>Entry</td><td>C00379</td><td>Compound</td></tr> <tr><td>Name</td><td>Xylitol</td><td></td></tr> <tr><td>PubChem SID</td><td>3669</td><td></td></tr> <tr><td>Structure</td><td colspan="2">  2D </td></tr> <tr><td>Source</td><td colspan="2">KEGG</td></tr> <tr><td>External ID</td><td colspan="2">C00379</td></tr> <tr><td>Source Category</td><td colspan="2">Curation Efforts Research and Development</td></tr> <tr><td>Version</td><td>11</td><td>Revision History</td></tr> <tr><td>Status</td><td colspan="2">Live</td></tr> <tr><td>Related Compounds</td><td colspan="2">PubChem CID CID 6912 (Xylitol)</td></tr> </table>	Entry	C00379	Compound	Name	Xylitol		PubChem SID	3669		Structure	 2D		Source	KEGG		External ID	C00379		Source Category	Curation Efforts Research and Development		Version	11	Revision History	Status	Live		Related Compounds	PubChem CID CID 6912 (Xylitol)	
Entry	EC 1.1.1.10	Enzyme																																																																										
Name	L-xylitol reductase; xylitol dehydrogenase (ambiguous)																																																																											
Class	Oxidoreductases; Acting on the CH-OH group of donors; with NAD ⁺ or NADP ⁺ as acceptor BRITE hierarchy																																																																											
Sysname	xylitol:NADP+ 4-oxidoreductase (L-xylulose-forming)																																																																											
Reaction(IUBMB)	xylitol + NADP+ = L-xylulose + NADPH + H ⁺ [RN:R01904]																																																																											
Reaction(KEGG)	R01904 REACTION																																																																											
Substrate	xylitol [CPD:C00379]; NADP+ [CPD:C00006]																																																																											
Product	L-xylulose [CPD:C00312]; NADPH [CPD:C00005]; H ⁺ [CPD:C00008]																																																																											
Entry	R01904	Reaction																																																																										
Name	Xylitol:NADP+ 4-oxidoreductase (L-xy)																																																																											
Definition	Xylitol + NADP+ => L-Xylulose + NADP																																																																											
Equation	C00379 + C00006 => C00312 + C00005																																																																											
Reaction class	RC00001 C00005_C00006 RC00102 C00312_C00379																																																																											
Enzyme	1.1.1.10																																																																											
Entry	C00379	Compound																																																																										
Name	Xylitol																																																																											
PubChem SID	3669																																																																											
Structure	 2D																																																																											
Source	KEGG																																																																											
External ID	C00379																																																																											
Source Category	Curation Efforts Research and Development																																																																											
Version	11	Revision History																																																																										
Status	Live																																																																											
Related Compounds	PubChem CID CID 6912 (Xylitol)																																																																											

図 4: PubChem の例

図 3: KEGG の例

3.1 化学データベースからの情報抽出

8/20

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

背景

有機合成分野と情報分野の関わり

3. ケモインフォマティクスとテキストマイニング

提案手法

今後の予定

4. 提案手法

9/20

4.1 使うべき酵素(EC番号)の予測

ある反応を与え、触媒として使用できる酵素の EC 番号を予測する

- 全ての化合物をクラスタリングし、「目的の反応式の化合物」に近い「化合物」の反応式の EC 番号を提示する
- EC 番号の反応式は生物の体内で起こっている反応
→似たような化合物が生体反応で使われている
= そこで使われている酵素を同じように使えば、目的のものを作れる可能性がある。

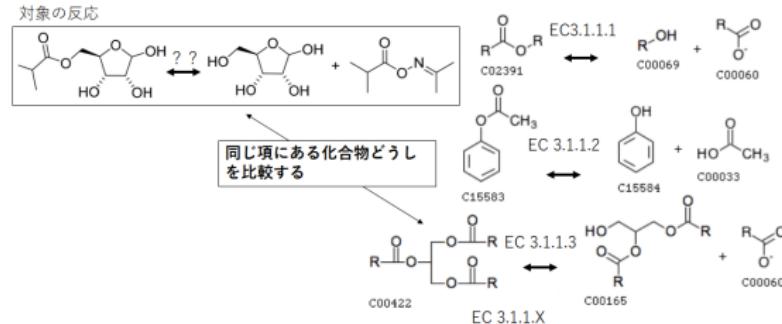


図 5: 対象とする反応式と EC 番号酵素の反応式

背景

有機合成分野と情報分野の関わり

3. ケモインフォマティクスとテキストマイニング

提案手法

今後の予定

4. 提案手法(2)

10/20

使うべき酵素の EC 番号の予測

背景

有機合成分野と情報分野の関わり

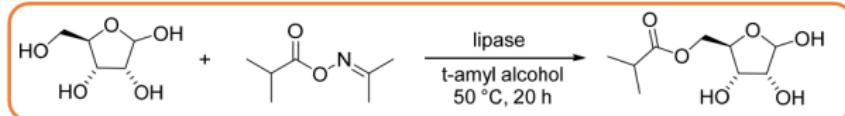
3. ケモインフォマティクスとテキストマイニング

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
- X に属する酵素の反応式(約 100 種)の中からクラスタリングで最も近いものとして EC3.1.1.3 の反応式が予測されると想定

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-isobutyl ribose assay yield%
Novozym 435 (Candida antarctica lipase B)	Novozym 社製 65
IMMTLL-T2-150 (Thermomyces lanuginosus lipase)	40
IMMRRES-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 6: 対象とする反応と酵素製品

4. 提案手法(3)

11/20

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

	MaxESStateIndex	MinESStateIndex	MaxAbsESStateIndex	MinAbsESStateIndex	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 7: 各化合物の 208 種の物性値 (特徴量)

背景

有機合成成分野と情報分野の関わり

3. ケモインフォマティクスとテキストマイニング

提案手法

今後の予定

特徴選択

208 種の特徴量から分析に使う特徴量を減らす

◎特徴選択の必要性◎

- ① 多重共線性の問題解決 (1)
- ② 次元の削減 (2)

(1) 多重共線性への対策

多重共線性 = 非常に相関の高い (> 0.9 程度) 特徴量同士が存在するときに起こる

→相関が高い = ほぼ同じ特徴量なので重複していることになる
(過学習の原因)

→重複しそうな特徴量を 1 つに絞る

- 類似度の高い特徴量同士をクラスタリングして、その中から 1 つだけ選ぶようにする

背景

有機合成分野と情報分野の関わり

3. ケモインフォマティクスとテキストマイニング

提案手法

今後の予定

完全連結法を用いた階層的クラスタリング

- ① 最も距離の遠い変数同士から順番に連結し、クラスタを形成する
- ② クラスタ同士の連結ではクラスタ内の変数で最も遠いもの同士を基準として連結

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

C_i : クラスター*i*
 x_i : クラスター*i*内で最も遠い変数

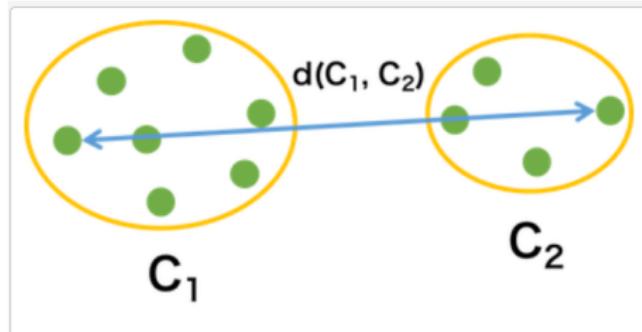


図 8: 完全連結法のイメージ

背景

有機合成成分野と情報分野の関わり

3. ケモインフォマティクスとテキストマイニング

提案手法

今後の予定

(2) 次元の削減

- 特微量を多く用いるほど、その化合物の物性をより説明できる
⇒過適合を起こしやすくなる
- 特微量が少ないと、類似性比較の精度があまりよくない
→適切な特微量を必要な数だけ選ぶようにする

主成分分析による変数選択

- 特微量を合成して新たな特微量を作る
→第1主成分、第2主成分、…第n主成分という新たな特微量にする（※上位の主成分ほど化合物の特徴をよく表している）
- 第1主成分に元の特微量がそれぞれどれくらい影響しているか調べる（第i成分まで調べる）
→主成分負荷量で見る

5.1 数値実験 今やっていること

15/20

階層性クラスタリングによる特徴合成

背景

有機合成分野と情
報分野の関わり

3. ケモインフォマ
ティクスとテキス
トマイニング

提案手法

今後の予定

5.1 数値実験 今やっていること 2

背景

有機合成分野と情報分野の関わり

3. ケモインフォマティクスとテキストマイニング

提案手法

今後の予定

	mean_in_cluster_0	mean_in_cluster_1	mean_in_cluster_2	RingCount	mean_in_cluster_4	mean_in_cluster_5	mean_in_cluster_6	NumSaturatedHeterocy		
PC1	0.016678	0.036788	0.038271	0.027602	0.004460	0.007662	0.015697	0.006		
PC2	0.002925	0.001416	0.002265	0.007932	0.012147	0.024411	0.037752	0.003		
PC3	0.000352	0.002905	0.003072	0.007244	0.010231	0.015342	0.002710	0.008		
PC4	0.001228	0.002524	0.000933	0.004092	0.000361	0.022387	0.000027	0.038		
PC5	0.018238	0.005772	0.001538	0.001207	0.004406	0.000610	0.001201	0.006		
PC6	0.004439	0.003933	0.000359	0.028148	0.002582	0.005580	0.000052	0.038		
	BertzCT	mean_in_cluster_9	mean_in_cluster_2	mean_in_cluster_1	mean_in_cluster_13	HallKierAlpha	Kappa2	RingCount	SlogP_VSA5	SMR_VSA5
PC1	0.197305	0.195726	0.19563	0.191802	0.178185	0.174671	0.168351	0.166139	0.160269	0.159599
	SMR_VSA1	NumHeteroatoms	mean_in_cluster_6	VSA_Estate1	SlogP_VSA2	MinEstateIndex	SlogP_VSA3	NumHAcceptors	EState_VSA1	EState_VSA1
PC2	0.218784	0.210405	0.194298	0.193636	0.193618	0.192712	0.185314	0.184671	0.183545	0.18135
	PEOE_VSA1	NumAromaticCarbocycles	SMR_VSA9	SlogP_VSA11	VSA_Estate7	PEOE_VSA10	Kappa3	NumRotatableBonds	PEOE_VSA14	PEOE_VSA
PC3	0.219076	0.215066	0.214358	0.213327	0.209262	0.207787	0.203269	0.196695	0.18517	0.1780

図 9: 主成分分析と主成分負荷量

やってきたこと

17/20

KEGG と PubChem で取得したデータを整理する流れを作成

背景

有機合成分野と情報分野の関わり

3. ケモインフォマティクスとテキストマイニング

提案手法

今後の予定

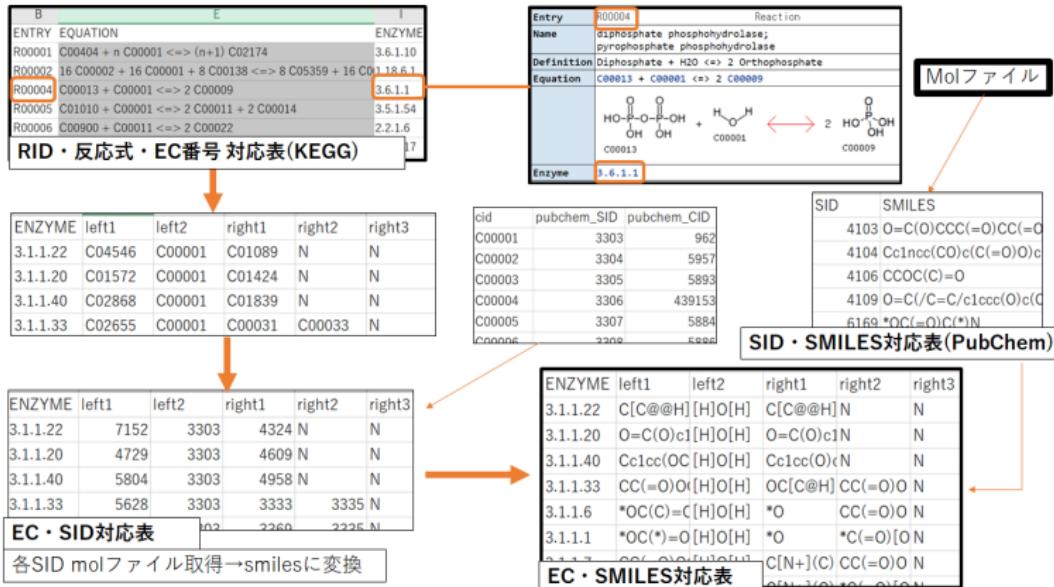


図 10: EC 番号・SMILES 対応表

やってきたこと

18/20

KEGG と PubChem で取得したデータを整理する流れを作成

背景

有機合成分野と情報分野の関わり

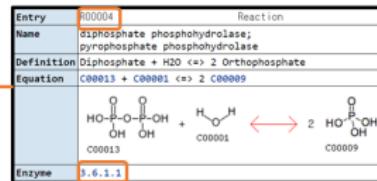
3. ケモインフォマティクスとテキストマイニング

提案手法

今後の予定

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 C00186.1		
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)



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

EC・SID対応表

cid	pubchem_SID	pubchem_CID
C00001	3303	962
C00002	3304	5957
C00003	3305	5893
C00004	3306	439153
C00005	3307	5884
C00006	3308	2260

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

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)[H]O[H]	C[C@H]N	N		
3.1.1.20	O=C(O)c1[H]O[H]	O=C(O)c1N	N		
3.1.1.40	Cc1cc(OC[H]O[H])Cc1cc(O)c1N	N			
3.1.1.33	CC(=O)O[H]O[H]	OC[C@H]CC(=O)O N			
3.1.1.6	*OC(C)=C[H]O[H]	*O	CC(=O)O N		
3.1.1.1	*OC(*)=O[H]O[H]	*O	*C(=O)[ON		
3.1.1.7	CC(=O)O[N+](C)CC(=O)O N				

EC・SMILES対応表

図 11: EC 番号・SMILES 対応表

今後の予定

19/20

今後の予定

- 第 10 主成分までの各主成分の因子負荷量から特徴選択する
- SOM の実装

背景

有機合成分野と情報分野の関わり

3. ケモインフォマティクスとテキストマイニング

提案手法

今後の予定

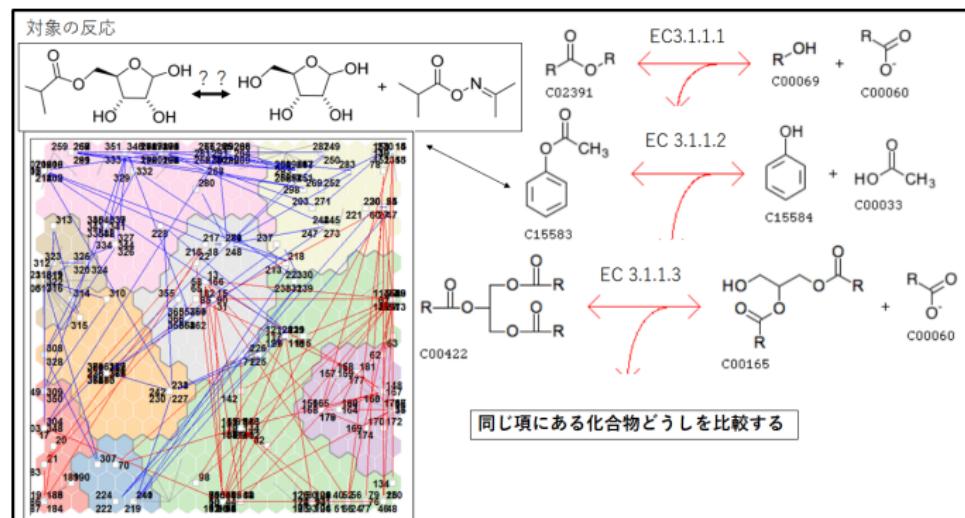


図 12: SOM のイメージ

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

背景

有機合成分野と情報分野の関わり

3. ケモインフォマティクスとテキストマイニング

提案手法

今後の予定