卒業論文

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

富山県立大学 工学部 電子・情報工学科
1815070 武藤 克弥
指導教員 奥原 浩之 教授
提出年月: 2022年2月

目 次

図一覧

表一覧

記号一覧

以下に本論文において用いられる用語と記号の対応表を示す.

用語	記号
クラスタ i	C_i
C_i に属するデータ集合	$\boldsymbol{X}_i$
クラスタ間の距離	$d(C_1, C_2)$
クラスタ内の要素間の距離	$d(\boldsymbol{X}_1, \boldsymbol{X}_2)$
個数	n
次元数	p
p 次元観測ベクトル	$\boldsymbol{x}_j$
i 番目のユニット，ユニット数	m_i k
i 番目ユニットの重心	$\boldsymbol{r}_i$
i 番目ユニットの重みベクトル	$\boldsymbol{\xi}_i$
$\boldsymbol{x}_j$ と $\boldsymbol{\xi}_i$ のユークリッド距離	$\|\|\boldsymbol{x}_j - \boldsymbol{\xi}_i\|\|$
$\|\|\boldsymbol{x}_j - \boldsymbol{\xi}_i\|\|$ を最小化する $\boldsymbol{\xi}_i$	$\boldsymbol{\xi}_c$
$\boldsymbol{\xi}_c$ を持つ勝者ユニット	m_c
学習率係数	$\alpha(t)$
ユニット m_c の近傍領域	N_c
N_c の散らばりに関する調整関数	$h(t)$
反応物 i の特性値	RT_i
生成物 i の特性値	PD_i
記述子 j の特性値変化量	cv_j
i 番目の反応式の特徴ベクトル	$\boldsymbol{DF}_i$
記述子 u，v 間の相関係数	s_{uv}

第1章

はじめに

§ 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．化合物の特性や構造を分析したり，化合物の特徴を抽出し，機械学習における分類や化学反応の設計や予測といったことが行われている．

現在，新型コロナウイルスの世界的な流行をはじめとする多種の影響によって，新薬開発のニーズが高まっている．2026 年までの間に，ケモインフォマティクス業界は，年平均成長率 13 %で市場が成長すると予想されていることからも [1]，ケモインフォマティクスの需要は日々拡大している．

有機合成分野においては，ケモインフォマティクスや機械学習などの技術を取り入れて，化学反応の設計や予測をする研究が増加している．一方で，目的の生成物を得るために使用する反応触媒に，グリーンケミストリーの観点から，環境適応型の酵素を用いることが世界の風潮となってきている．酵素に代表される生体触媒は，人工的な化学触媒に比べて環境にやさしく，化学反応をより効率的に進めることから，化学触媒の代わりに生体触媒を用いて合成を行う取り組みが増加している．実際，目的の化合物を生成するために従来では 10 ステップの合成を行っていたものを，生体触媒を取り入れることで 3 ステップまで短縮したという研究事例もある [2]．これらのことから，目的物生成のために酵素反応を取り入れたうえで，反応設計を行うこと，あるいは，特定の反応に対して生体触媒として最適な酵素を予測することも重要な要素の一つとなってきている．

情報科学の観点からとらえると，酵素を触媒として取り入れる際，反応物 (基質) に対して特定の生体酵素を加えれば目的の生成物が得られる．つまり，基質と生成物が決まった場合，それに対して最適な酵素を予測するというのは容易に見えるかもしれない．ところが，実際には基質特異性と呼ばれる，酵素が基質に対して高い反応性を示すかどうかという酵素の特性によって，問題が複雑になる．有機合成化学を研究していて，酵素に関する知識を持ち合わせていたり，経験が豊富であれば，どの酵素が使えそうかある程度予測ができるかもしれない．しかし，先ほど述べた基質特異性に加えて，酵素のタンパク質配列を参照したりと，遺伝子分野にかかわる部分もあり，有機合成の知識だけでは解決が難しい場合がある．

§ 1.2　本研究の目的

生体触媒（Biocatalyst）を用いた有機合成化学において，目的とする生成物を効率よく得るために，酵素のデータベースを参照したり，酵素の研究を行っている専門家と協力するなどして，最適な酵素候補の目途をつけるという手法が取られる場合がある．実際は，酵素にも同様の性質を持っていたり，複数の企業製品が存在していたりと，触媒候補が複数存在する場合があるため，スクリーニングなどの実験の試行錯誤を繰り返しながら，最終的に 1 つに絞られていく．ここで，酵素の候補を探索したり，新たな酵素を設計する際に，有機合成化学の研究者自身で，酵素候補を探索することができれば，次の実験のステップまでスムーズに進めることができると考えられる．つまり，目的生成物を得るために，最適な酵素を迅速に予測・設計してくれるようなツールが存在すればよい．

本研究では，反応式を与えた際，その反応を触媒するのに必要な酵素を予測するシステムを考える．前述のとおり，1 つの酵素に絞り込むためには，様々な条件が絡む実験を必要とするため，おおまかな予測という形になる．しかし，有機合成化学の知識内で手順を進めていくことが可能となるため，十分に有効性があると考えられる．

酵素は酵素番号 (Enzyme Commission numbers: EC 番号) とよばれる，4 組の数字の組み合わせからなる番号が割り振られており，どの反応を触媒し，どの結合・基質に反応するかによって分類されている [11] [12]．与えられた反応に対して，酵素 (EC 番号) を予測できれば，その EC 番号の酵素から何を選択するかという次のステップに進むことができる．

EC 番号の情報の中には，反応物から生成物への，その酵素を使った代表的な反応が記載されている．そこで，本研究では，酵素を予測するターゲットとなる反応式内の反応物から生成物，また，EC 番号の代表的な反応式内の反応物から生成物，それぞれの物理・化学的特性値の変化を比較し，類似性が最も高い反応の酵素番号を提示して，最適な酵素を予測する．

主な流れとして，化学・酵素データベースから酵素の EC 番号および，代表的な反応式の情報を取得し，EC 番号と反応式の対応表を作成する．次に各反応式を，反応物と生成物に分解する．ターゲットの反応式も同様に分解し，各化合物の構造をコンピュータ上で扱うための表現に変換する．その後，複数の化合物の物理・化学特性値を計算し，各反応式において反応物から生成物への特性値の変化量を求める．この複数の特性値変化量を要素にもつ多次元ベクトルを，反応式の特徴ベクトルとして表現する．最終的に特徴ベクトルの次元削減を行い，クラスタリングによって反応式の特徴ベクトルを 2 次元平面上に出力する．得られた結果から，ターゲットの反応式に対して，最も近い場所に位置する，反応式の EC 番号に登録されている酵素を，用いるべき最適な酵素として予測する．

§ 1.3　本論文の概要

本論文は次のように構成される．

第 1 章 本研究の背景と目的について説明した．背景では，ケモインフォマティクスの概要，有機合成において，生体触媒を用いることのメリットとその課題について述べた．目的では，目的の生成物を得る際に用いる，最適な酵素を予測するための，EC 番号を予測するシステムの概要について述べた．

第 2 章 有機合成，ケモインフォマティクス，および酵素の概要を述べる．また，本研究で用いるデータベースについて述べる．

第 3 章 化学データベースからの情報抽出，ケモインフォマティクスでにおける化合物の構造表現法，EC 番号予測の概要を述べる．また，クラスタリング手法について述べる．

第 4 章 提案手法についての説明，および手順について説明する．

第 5 章 提案手法による数値実験の概要，実験結果と考察を述べる．

第 6 章 まとめと今後の課題について述べる．

第2章

有機合成分野と情報分野の関わり

§ 2.1　有機合成と情報技術

有機合成では人工的に有機化合物を作り出すことを目的としている．古くから病の治療として天然の有機化合物が用いられており，薬として有効な成分のみを取り出すことが近年行われてきた．1805 年，F.W.A.Serürner がアヘンから強い麻酔作用を持つ morphine を取り出すに成功したことを皮切りに, 薬効成分が次々に抽出されるようになっていき，その発展とともに有機化学が発展していった [3]．また，1828 年には，Wähler が有機化合物として初となる尿素の合成に成功し，その後 Liebig を筆頭として，有機化合物の扱い方，構造式での構造理解が明確化されていった [4]．現在まで比較的簡単に入手できる化合物から，天然に存在する薬の成分などを生成する全合成によって，様々なものが合成されてきた．コンピュータが発達するようになると，実験結果で得られた情報がデータベースに蓄積されていき，化学の現象をコンピュータ上で上手く表現することで，高速なデータ処理が可能となった．そして，データベースにある情報等から，情報技術によってデータを分類・予測する分野としてケモインフォマティクスは現在まで発展してきた．ケモインフォマティクスの研究分野として以下のものが挙げられている [5].

1. ケモインフォマティクス情報検索，データベース，グラフ理論，反応設計など
2. マテリアルズ・インフォマティクス構造物性相関など
3. バイオインフォマティクス
4. 計算機科学
5. 理論・計算科学 (量子科学，分子軌道法，分子科学)
6. コンビナトリアルケミストリー
7. 通信・システム (コンピュータネットワーク，並列化，専用機，コンピュータグラフィックスなど)
8. ラボラトリーオートメーション
9. 関連する化学教育・学習システム

化学分野では情報学に適用できる問題が多く存在する．例えば，化合物の構造に着目したとき，原子の部分を頂点，結合部分を辺とみなすことでグラフ理論の問題になる．合成時の反応経路の設計においては，どの化合物から出発し，いかにステップ数やコストなどを抑え，かつ効率的に目的物を生成していくかという最適化問題に帰着できる．例として，化学反応ネットワーク中の最適な反応経路に関する研究がある [6]．ここでは，化学反応における安定平衡構造を頂点，遷移状態を辺とした化学反応ネットワークにおける最短経路候

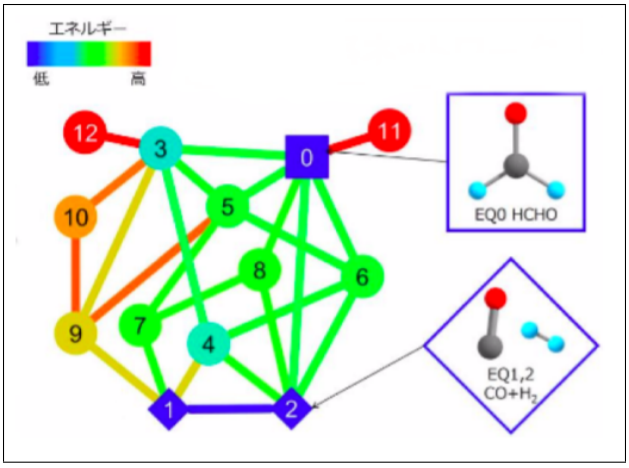


図 2.1: HCHO の化学反応ネットワーク [6]

補について検討している．莫大な通りの経路を調べる代わりに，反応が経由するそれぞれの状態における最大・最小エネルギー差が小さい経路に絞り，合成経路設計，反応予測，逆合成解析の 3 つの場合において，K 最短経路問題などを適用して，計算機実験による探索の性能評価を行っている．図 refRN は原子 H,C,H,O で構成される化合物の化学反応ネットワークを表している．化合物を合成する過程においては，化学合成経路設計と，化学反応予測の 2 つのアプローチがある [7]．化学合成経路設計では，目的とする最終的な生成物を設定し，それに対して，何の物質から出発してどのような合成経路をたどって，合成していくかという逆合成方向の思考をもとにした手法である．化学反応予測は，出発物質を決め，目的の生成物を得るための反応は起こるのか，副産物は何が生成されるのかといった正合成方向の思考を元にしている．化学合成経路の設計を考えるソフトウェアは古くから開発されてきているが，化学反応予測のためのソフトウェアは，様々な因子が複雑に絡み合うために，開発数が少ないとされてきた．しかし，近年では機械学習の発展によって，その実態が変わりつつある．研究例として，化学反応時の電子移動に関する，極性反応とラジカル反応について予測したものがある [8]．ここでは，1110 個の極性反応，103 個のラジカル反応からデータベースを作成し，機械学習の分類を行っている．分子の反応部位や構造情報，原子の複数の性質などを，量子科学計算によって数値化および特徴ベクトルとして表現し，10 分割交差検証によって，分類精度を評価している．

このように，化学反応や化合物における，特定の特徴の数値化など，コンピュータ上で扱いやすく，機械学習に組み込みやすい形式を開発にすることによって，精度の良い反応予測を可能にしている．

§ 2.2　酵素と EC 番号

酵素は生体内で必要な化学反応を触媒するタンパク質で，生物が生きていくためには必要不可欠なものである．酵素には基質特異性という，特定の反応物 (基質) のみに触媒反応を示す特性を持っている．これは，Emil.H.Fischer が唱えた「鍵と鍵穴説」と呼ばれる，基質を鍵，酵素を鍵穴と見立てて考えられている．基質が酵素に結合することで反応が始ま

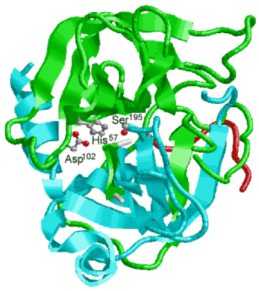


図 2.2: α-キモトリプシンの立体構造 [9]

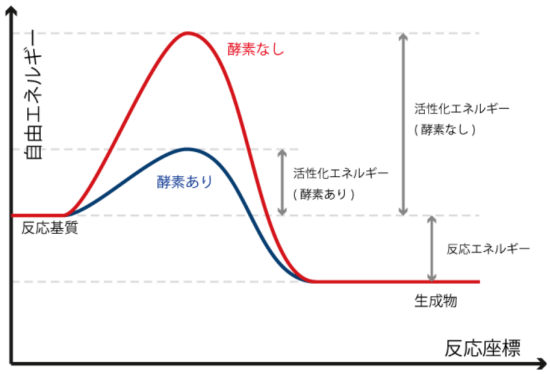


図 2.3: 反応の進行と必要なエネルギー [10]

り，基質が生成物へと変化すると結合が外れる．このとき，酵素自体は変化することなく元の状態に戻るため，触媒として繰り返し利用できる．酵素の構造イメージを図 2.2 に示す．

より多くの基質と結びついて作用する用途の広い触媒とするために，基質特異性を広げるタンパク質工学と呼ばれる分野がある．ここでは，アミノ酸配列の一部を置き換えることで酵素の性質を改変したり，ランダムに変異させた変異体ライブラリを作成し，スクリーニングによって所望の触媒機能をもったものを選択するといったことが行われている．

酵素を用いることのメリットとして以下のことが挙げられる．

反応速度の増加

酵素は生体触媒として，基質の化学反応をより早く，安定して進めることができる．化学反応において, 反応が進むにつれてエネルギーが増加し，遷移状態をピークに減少していく．この反応開始から遷移状態になるために，必要なエネルギーを活性化エネルギーと呼び，大きいほど反応が進みにくくなる．しかし，酵素を用いることで必要な活性化エネルギーが低下し，反応を速く進めることができる．酵素を用いた場合と用いなかった場合のエネルギー遷移の様子を図 2.3 に示す．

グリーンケミカル

通常，化学触媒は高温や高圧といった条件下で使用することが適している場合が多い．一方，生体触媒は常温，常圧で使用することができ，これらの条件下での反応であれば，高温・高圧にするためのエネルギー削減につながる．

高選択性

そして，選択性が高いとは，化合物の特定の構造を持つ部位のみを選んで，化学変化させることが可能であることを指し，合成ステップの省略などにつながる．

上記の理由から，生体触媒を使って，医薬品を生成する事例が増えている．例として新型コロナウイルスの治療薬として，治験が進められているモルヌピラビル (MK-4482) の合成がある [2]．ここでは従来 10 ステップで行っていたものを，生体触媒を取り入れることによって 3 ステップまで短縮している．図 2.4 にモルヌピラビルの従来と提案された合成方法の比較を示す．合成ステップを短縮することは，使用する試薬などのコストが減り，結果として環境にも優しい．

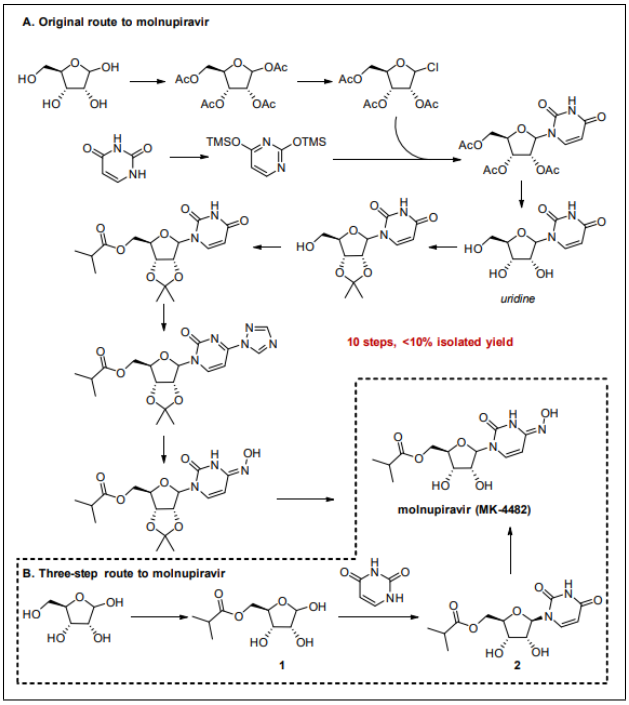


図 2.4: モルヌピラビルの合成

酵素番号 (Enzyme Commission numbers：EC 番号)

酵素は EC 番号という，4 組の数字の組み合わせからなる番号で管理されており，酵素の性質ごとに分類されている．EC ○. ○. ○. ○というように番号が振られ，1 番目の数字はどの反応を触媒するかによって，1(酸化還元酵素),2(転移酵素),3(加水分解酵素),4(離脱酵素),5(異性化酵素),6(合成酵素),7(輸送酵素) の 7 つに分類されている [11]．2 番目の数字では，どの結合に作用するか，3 番目の数字ではどの基質 (化合物) に反応するかや，必要とする補酵素情報というように分類され，4 番目の数字で 1 から 3 番目までの組み合わせ番号 (EC ○. ○. ○) に属する酵素の番号 (登録順) を表している [12]．以下にデータベース Enzyme Nomenclature [13] での EC 1.1.1.1 の酵素を参照した場合の例を図 2.5 に示す．ここでは,「EC 3.1 to EC 3.3」の「separate」のリンク先に行くと EC3(加水分解酵素) のうち，エステル結合に作用する EC 3.1，グリコシド結合に作用する EC 3.2，エーテル結合に作用する EC 3.3 の階層を見ることができる．さらに，EC 3.1 の下層に注目すると，カルボン酸エステルに作用する EC 3.1.1，チオエステルに作用する EC 3.1.2，リン酸モノエステルに作用する EC 3.1.3 という分類がなされている．

§ 2.3　化学・酵素データベース

化学・生物分野において用いられているデータベースについて，いくつか説明する．

Kyoto Encyclopedia of Genes and Genomes(KEGG) [14]

遺伝子・タンパク質情報，タンパク質相互作用を可視化した 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,

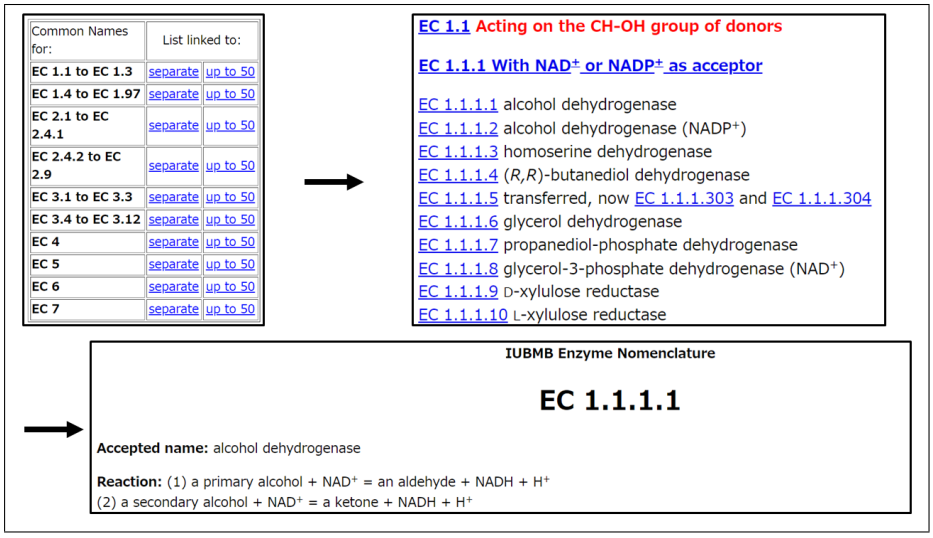


図 2.5: Enzyme Nomenclature で EC.1.1.1.1 を参照する場合の例

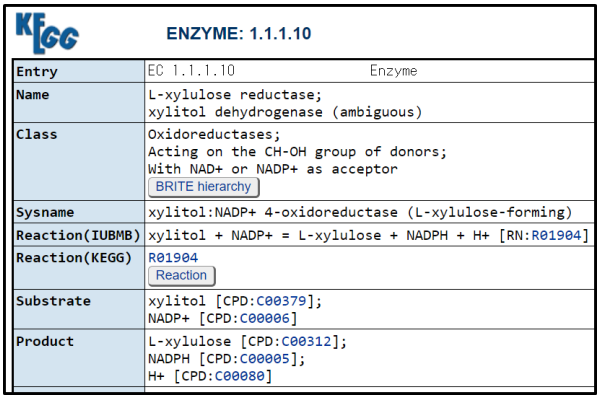


図 2.6: KEGG ENZYME データベース

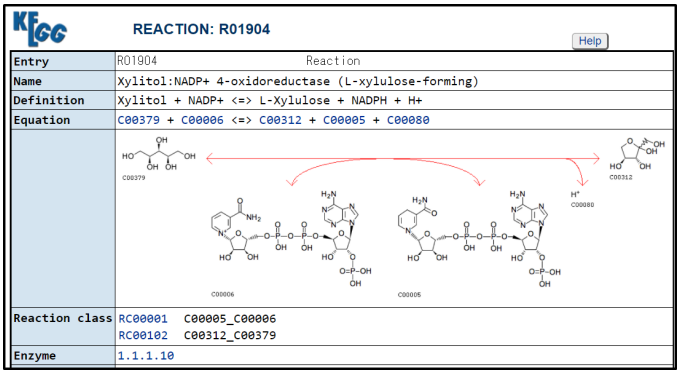


図 2.7: KEGG REACTION データベース

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である．KEGG ENZYME では各酵素の情報を該当する EC 番号から検索して得ることができ，酵素の別名，その酵素を用いた生体内の反応式，基質・生成物情報，遺伝情報，文献情報等について書かれている．KEGG REACTION には酵素を用いて起こる化学反応についての情報を記している．それぞれの反応は R から始まる 5 桁の数字で管理されており，反応に用いられる酵素と EC 番号，化合物名・C 番号・構造式でそれぞれ表した反応式等が書かれている．KEGG COMPOUND では C から始まる 5 桁の C 番号で化合物を管理しており，主に KEGG PATHWAY 中や KEGG REACTION 中に現れる化合物を扱っている．C 番号，名前，分子式で検索することができ，そのリンク先には，別名，分子式，分子量，構造式，登場する R 番号，PATHWAY MAP の MAP 番号，EC 番号のリンク先，他のデータベースへのリンク先などが掲載されている．サイト内のリンクのつながりによって，EC 番号から R 番号，R 番号から C 番号とたどることができる．図 2.6 および図 2.7 に KEGG データベースの例を示す．

SciFinder^n [15]

Chemical Abstracts Service(CAS) が提供する，データベース．主に，「Substances(化学物

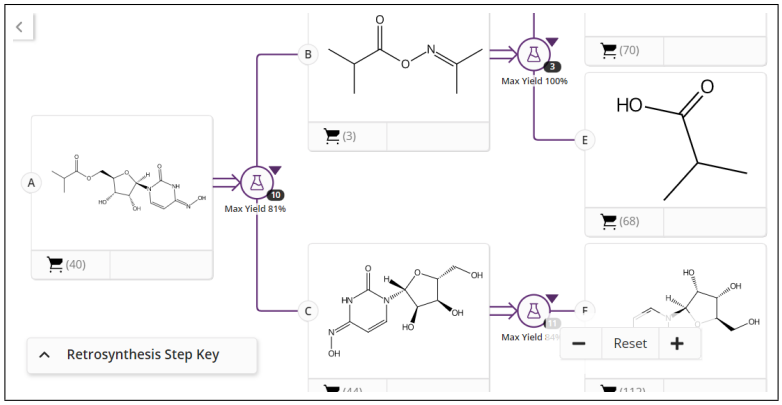


図 2.8: SciFinder^n による逆合成設計予測

質情報)」,「Reactions(反応情報)」,「References(文献情報)」,「Suppliers(カタログ情報)」,「Biosequences(配列情報)」の項目から検索することができる.

「Substances」では化学物質の名前，CAS 登録番号，分子式やスペクトル，物性値などで検索できる [16].「Reactions(反応情報)」では化学物質名，構造式などから検索され，その化合物が反応式・生成物として用いている反応式を調べることができる．また，生成物の収率，反応に用いる試薬や溶媒，文献情報などを条件に入れてフィルター検索もできる.「References」ではキーワード，著者名，文献番号，雑誌情報，機関名などで検索される.「Suppliers(カタログ情報)」では，検索した化合物を取り扱う企業などのカタログ情報を取り扱っている．検索結果には取扱業者名と純度情報，化合物の購入サイトへのリンクと取り扱い分量等が表示される.「Biosequences」では DNA,RNA, タンパク質の配列情報や類似する配列などで検索される.

SciFinder^n では，構造式をユーザ自信が描画・編集して検索することが可能である．化合物の構造が一致するもの，または構造の類似度に基づいて検索できる他，関連する反応式，文献情報，提供元の情報も参照できる．さらに，作成した構造式を生成物として，逆合成ルートを設計・予測する「Retrosynthesis Planner」というツールが存在する．ここでは合成ステップ数やコストなどを設定し，複数パターンの合成プランが設計される．各合成ルートは既知の反応または，予測された反応で構成され，最大の収率が表示される．図 2.8 にモルヌピラビルをターゲットとして，逆合成ルートの設計・予測をした様子を示す.

PubChem [17]

化合物名，分子式，化合物の 2D(もしくは 3D) 形式の構造イメージ，化学・物理特性，生物学的活性情報，毒性情報，文献情報等のデータを収録している．データ提供者からアップロードされた，約 2.8 億種の化学物質情報や約 140 万種の生物学的実験データ，標準化された約 1.1 億種の化学構造情報，また，約 10 万種の遺伝子データなどから構成される [18]. さらに，PubChem Compound，PubChem Substance，PumChem BioAssay の 3 つのデータベースがある.

PubChem Substance では，研究者がアップロードしたデータを管理している．複数の提供者から重複するデータがアップロードされることがあるため，標準化によって，同様の情報を集約し，PubChem Compound に格納される [19]．また，PumChem BioAssay では，

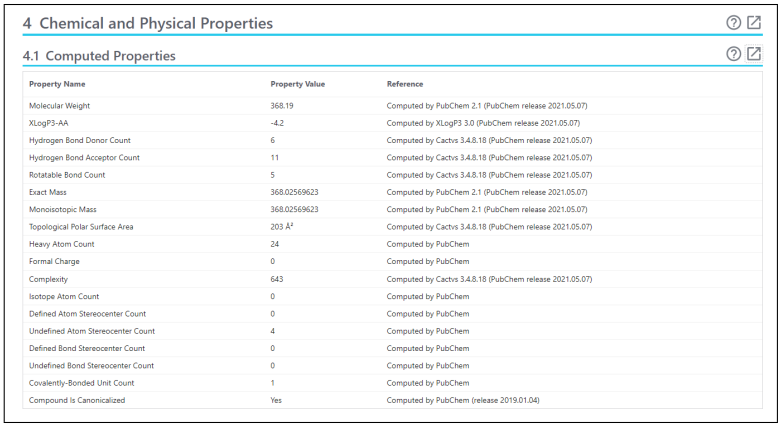


図 2.9: PubChem の例

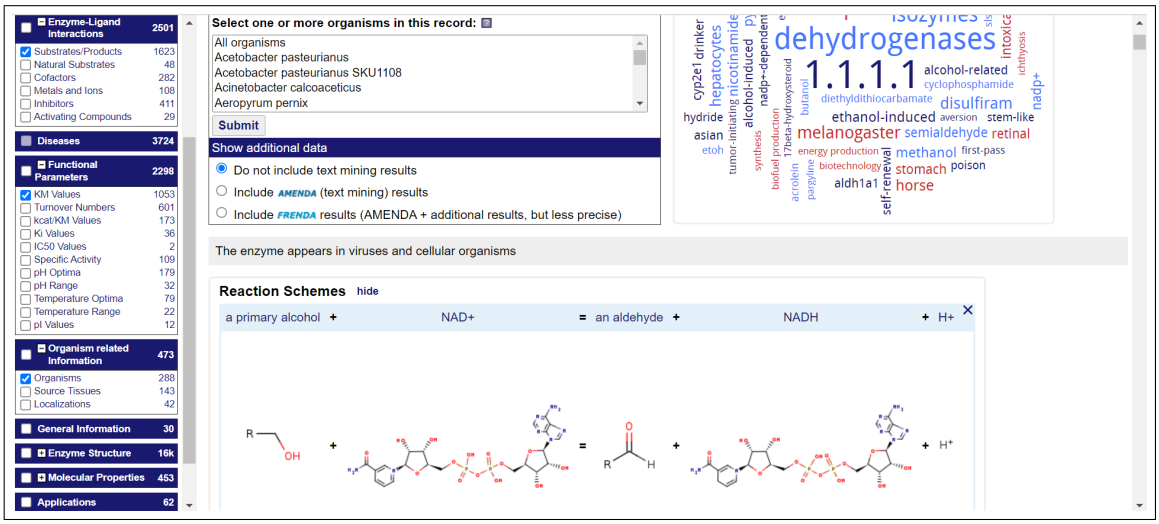


図 2.10: BRENDA 上の EC1.1.1.1 に関する情報

データ提供者の実験環境によってばらつきが生じる生物活性データ等を，実験に用いられた化合物と，実験結果ごとに紐づけを行うことで管理している．それぞれのデータベース中のデータには，SID(SubstanceID)，CID(CompoundID), AID(AssayID) が割り振られている．特に SID は KEGG のほとんどの C 番号と対応している．図 2.9 に PubChem のデータベースの例を示す．

BRENDA [20]

酵素に関するデータを，文献の情報をもとに網羅したデータベース．酵素名，生物種，CAS 登録番号，EC 番号，特性値などで検索することができる．例として，EC 番号で検索した様子を図 2.10 に示す．検索した EC 番号のページに行くと，その酵素に関係している単語のワードマップや用いられている反応式が書かれている．図 2.10 の画面左にある画面から目的とする詳細情報を表示できる．例えば，Substrates/Products では，検索した EC 番号の酵素を使った反応の基質・生成物のペアが記されている．Organisms では，酵素を作る由来となった生物種のリストが表示されている．また，「Functional Parameters」ボックス

内の KM Values では，酵素の由来となった生物種・基質ごとの Km 値 (基質と酵素の親和性を表す指標) を見ることができる．

第3章

ケモインフォマティクスと情報技術

§ 3.1　化学データベースからの情報抽出

Web サイト等から収集した大量の情報の中から，自然言語処理を用いて有用な情報を抽出するテキストマイニングにおいては，スクレイピングが用いられることがある．スクレイピングとは，Web サイトから文章をプログラミングによって自動取得する方法であり，効率的にデータを収集できる．一方でデータベースを管理している Web サイト等においては，独自のアプリケーション・プログラミング・インターフェース (Application Programming Interface: API) を備えている場合があり，指定された形式でプログラムを記述すれば，データベース上の情報を自動的に取得することができる．

化学データベースにも公式の API が公開されているものがいくつか存在する．KEGG では KEGG API [21]，PubChem では POWER USER GATEWAY(PUG) [18] と呼ばれる API が公開されており，本節ではこの 2 つの API について説明する．

KEGG API の構成

KEGG API のフォーマットは以下のようになる [21]．<operation>の部分に上記の 7 つのいずれかを指定する，例えば,「list」を指定した場合．以下のフォーマットに従う．<dbentries>で目的のデータがある KEGG データベース名を指定する．例えば,「pathway」を指定することで，完成するリンク先へ行くと，各 Pathway のマップ番号と，Pathway 名の対応リストを取得できる．図 3.1 に番号と Pathway 名の対応表を示す．このように,「http://rest.kegg.jp/」以下の部分で指定された識別子を設定することで，データが保存されている URL に移動することができ，各プログラム言語で実装されている，リンク先の中身を取得するコードによって，必要なデータを取得することができる．

PUG

Common Gateway Interface(CGI) を経由して，PubChem のデータをプログラミングによって，取得する機能を提供するシステム [22]．データのやり取りは URL ではなく XML を用いる．XML によるリクエストを CGI へ送り，リクエストの内容が実行された後，結果が

```
http://rest.kegg.jp/<operation>/<argument>[/<argument2[/<argument3> ...]]
<operation> = info | list | find | get | conv | link | ddi
```

図 3.1: KEGG API の URL 構成 1

```
http://rest.kegg.jp/list/<dbentries>

<dbentries> = Entries of the following <database>
<database> = pathway | brite | module | ko | genome | <org> | vg | vp | ag |
    compound | glycan | reaction | rclass | enzyme | network | variant |
    disease | drug | dgroup | <medicus>
```

図 3.2: KEGG API の URL 構成 2

表 3.1: リンク先の対応表

path:map00010	Glycolysis / Gluconeogenesis
path:map00020	Citrate cycle (TCA cycle)
path:map00030	Pentose phosphate pathway
path:map00040	Pentose and glucuronate interconversions
path:map00051	Fructose and mannose metabolism
path:map00052	Galactose metabolism
path:map00053	Ascorbate and aldarate metabolism
path:map00061	Fatty acid biosynthesis
path:map00062	Fatty acid elongation
path:map00071	Fatty acid degradation
path:map00073	Cutin, suberine and wax biosynthesis
path:map00100	Steroid biosynthesis
path:map00120	Primary bile acid biosynthesis
path:map00121	Secondary bile acid biosynthesis
path:map00130	Ubiquinone and other terpenoid-quinone biosynthesis
path:map00140	Steroid hormone biosynthesis
path:map00190	Oxidative phosphorylation
path:map00195	Photosynthesis
path:map00196	Photosynthesis - antenna proteins
path:map00220	Arginine biosynthesis
path:map00230	Purine metabolism
path:map00232	Caffeine metabolism
path:map00240	Pyrimidine metabolism
path:map00250	Alanine, aspartate and glutamate metabolism
path:map00253	Tetracycline biosynthesis
path:map00254	Aflatoxin biosynthesis
path:map00260	Glycine, serine and threonine metabolism
path:map00261	Monobactam biosynthesis
path:map00270	Cysteine and methionine metabolism
path:map00280	Valine, leucine and isoleucine degradation

XML で返信される仕組みとなっている．例として，CID1 と CID99 の化合物の構造を SDF ファイル形式の gzip 圧縮でダウンロードする場合，図 3.4 のような XML 構造のリクエスト応答となる．PubChem ではアクセス簡略化のため，PUG-SOAP と PUG-REST というシステムが実装されている．本研究では PUG-REST を用いるため，PUG-REST について説明する．

PUG-REST

PUG や PUG-SOAP で用いられている XML 形式の記述を必要とせず，簡単な記述方でデータを取得することができる API．PUG-REST のリクエストは以下のような URL で表記される [18]．<input specification>はさらに<domain>/<namespace>/<identifiers>で構成されており，何のデータを取ってくるのかを定める．<domain>では，substance，compound，assay などの対象とするデータベースを指定する．また，<namespace>では CID(cid) や化合物名 (name)，分子式 (formura) 等を指定し，<identifiers>では，CID の番号，化合物名・分子式の文字列といった，<namespace>に対する具体的な名前を指定する．<operation specification>では<input specification>で指定したデータ保管場所にアクセスした際，どのような操作を所望しているのかを記述する．例えば，<input specification>で CID 番号の情報を記述している状態で，synonyms を指定するとその CID 番号の化合物名に対する同義語のリストが返される．同様のケースで，<compound property>で property/XXX,YYY,$\cdots$,ZZZ/ を指定すると，その化合物の物性値や化学的特性値を複数取得することができる．<output

```
https://pubchem.ncbi.nlm.nih.gov/rest/pug/<input specification>/
<operation specification>/[<output specification>][?<operation_options>]

<input specification> = <domain>/<namespace>/<identifiers>
<operation specification> = record | <compound property> | synonyms | sids |
    cids | aids | assaysummary | classification | <xrefs> | description |
    conformers
<output specification> = XML | ASNT | ASNB | JSON | JSONP [ ?callback=<
    callback name> ] | SDF | CSV | PNG | TXT
```

図 3.3: PUG-REST のリクエスト

図 3.4: PUG における XML 応答の例 [18]

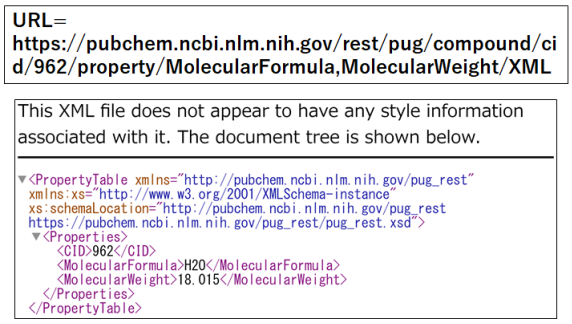


図 3.5: 水 (H_2O) の情報を取得するリクエスト URL とデータ取得結果

specification>の部分では取得したいデータをどのような形式で出力するかを指定する．基本的には，<input specification>/<operation specification>/<output specification>の部分を指定すれば良く，例として，水 (CID968) の分子式 (MolecularFormula) と分子量 (MolecularWeight) を XML で取得した場合を図 3.5 に示す.

§ 3.2　化合物の構造表現法と EC 番号予測手法

化合物同士の構造比較について述べる前に，ケモインフォマティクスで一般的に使われている化合物の構造表現について説明する．

MOL ファイル

化合物の構造情報を記したテキスト形式のファイル.「.mol」の拡張子で保存されることが多い．ファイル内には結合している原子と各原子の 3 次元座標リストやどの原子同士が結びついているかのリストが記述されている．通常の構造式と mol ファイルを比較したものを図 3.6 に示す

SDF ファイル

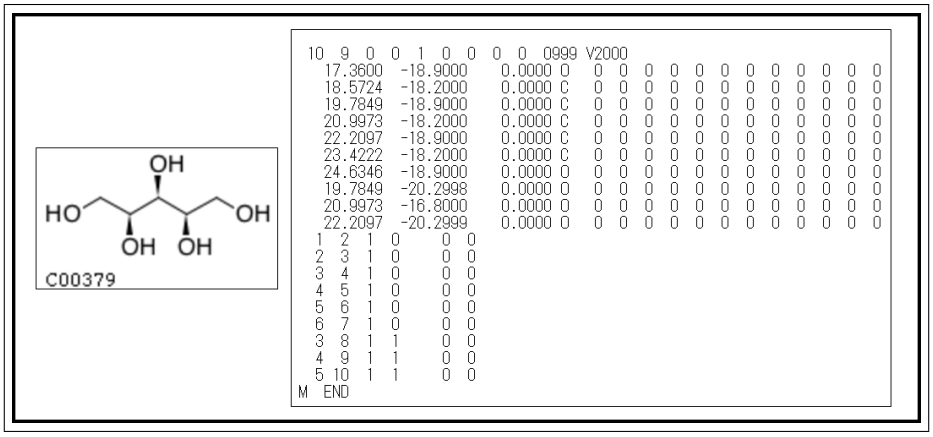


図 3.6: KEGG COMPOUND で取得できる構造式と MOL ファイルの例

MOL ファイルを複数統合した拡張子「.sdf」のファイル．2 つ以上の分子の MOL ファイルをデータベースから同時に入手する際は，この形式となることが多い．

SMILES

化合物の構造を文字列で表したもの．以下の規則に従って文字列に変換していく [23].

1. 原子は元素記号で表し，2 文字で区別がつきにくい原子 (Nb と NB 等) は [ ] で囲む
2. 水素原子は省略する
3. 隣接する原子は隣に記す
4. 二重結合は =，三重結合は#で表し，単結合・芳香族結合は省略する (芳香族原子は小文字の c などで表記する)
5. イオンなどで結合がない部分は「.」で分ける
6. 構造が分岐する箇所は ( ) で表記する
7. 環構造は切断して切断箇所を記すとともに (C1 など)，鎖錠構造で表す．

以上の規則に基づいて作成されたものを generic SMILES と呼ぶが，以下の規則を加えたものを isomeric SMILES と呼ぶ

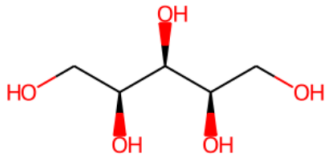
1. 同位体 (例えば炭素) がある場合 [13C] という表記にする
2. 立体異性体を区別するための絶対配置を「@」または「@@」で表現する
3. 二重結合などで生じる幾何異性を「/」と「¥」で表す

フィンガープリント

化合物の構造や特徴をビット列で表現したもの．構造のどの部分に注目するか，または性質などで種類がある．例えば，MACCS Keys のフィンガープリント [24] では，166 種の特徴的な構造を化合物が持っているかどうかを 0 と 1 で表現している．フィンガープリントは主に化合物同士の類似性比較で用いられる．

化合物の数値化 (特徴ベクトル)

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




```
Chem.MolToSmiles(Xylitol)
```

```
'OC[C@H](O)[C@@H](O)[C@H](O)CO'
```

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

```
152.14600000000002
```

図 3.7: rdkit を用いた化合物の情報

```
import pubchempy
pubchempy.get_properties([ 'MolecularFormula','MolecularWeight',
                           'IsomericSmiles' ], 'glucose',
                          'name', as_dataframe=True)
```

CID	MolecularFormula	MolecularWeight	IsomericSMILES
5793	C6H12O6	180.16	C([C@@H]1[C@H]([C@@H]([C@H](C(O1)O)O)O)O)O

図 3.8: PubChempy でグルコースの情報を取得した結果

機械学習で様々な予測を行うためには，化合物を数値化して表現する必要がある．その方法として，前述のフィンガープリントではビット列で化合物を数値化しているが，物性値を特徴として用いられることも多い．一般的に複数の物性値が用いられ，多次元の特徴ベクトルとして化合物の特徴を表現する．これらの構造情報や物性値で化合物の特徴を表したものは記述子と呼ばれている．

RDKit [25] を用いた化合物のデータ化

RDKit は Python 提供されている，化合物の構造を扱うライブラリである．SDF ファイルや MOL ファイルを読み込んで構造式の画像を出力したり，SMILES やフィンガープリントに変換することができる．RDKit では読み込んだ構造式から，化合物の記述子を計算することができるため，化合物同士の類似性を評価したり，機械学習に発展させることができる．例として，化合物の分子量を意味する MolWt を知りたい場合，MOL ファイルから読み込んだ化合物のインスタンスを生成し，RDKit の Descriptors クラスにある MolWt メソッドに生成したインスタンスを渡すことで，MolWt が計算され出力される．図 3.7 に，rdkit を用いて化合物の構造式と SMILES を出力した様子，および化合物の MolWt を計算した結果を示す．

PubChemPy [26]

PUG REST を用いて PubChem のデータを取得するための Python ライブラリ．化合物名や CID を引数にして，対象化合物の物性値や SMILES を取得することができる．例として，グルコースの分子式，分子量，IsomericSMILES を取得した結果を図 3.8 に示す．

以上の構造表現法を用いて，酵素予測の研究が多く行われている．ここでは主に，反応式を基質と生成物のペアとみなし，そのペアに対して最適な酵素を，EC 番号として予測する研究が行われており，手法として 2 種類存在する．1 つはアミノ酸配列の類似性を用いる手法である．酵素はタンパク質であるため，アミノ酸配列で表現される．アミノ酸配列の類似性に基づいて，該当する EC 番号を予測する．

もう 1 つの手法として，基質と生成物の構造に着目したものがある．構造をフィンガープリントなどで表したり，構造として特徴的な部分の化学変化に注目したりなどがある．前者

の手法では，基質と生成物をそれぞれ分子の部分構造 (フラグメント) に着目したフィンガープリントで表している．基質のフィンガープリントから生成物のフィンガープリントを引いた反応差分フィンガープリントを定義している [27]．そして，EC 番号の基質と生成物の反応差分フィンガープリントとの類似性をユークリッド距離で求め，最も類似した反応差分フィンガープリントの EC 番号を割り当てるという手法を用いている．KEGG REACTION の R00005 に登録されている反応式 $C01010 + C00001 <=> 2C00011 + 2C00014$ に対して，各分子の分子フィンガープリントを MFP として，反応差分フィンガープリント RFP は以下のように定義されている．

$$RFP_{R00005} = MFP_{C01010+C00001} - MFP_{2C00011+2C00014} \tag{3.1}$$

後者の手法 [28] では，RDM パターンと呼ばれる，基質と生成物の各構造に対して，反応中心原子 (R atom)，その近傍の原子で異なっている領域 (D atom) と一致している領域 (M atom) を定義している．EC 番号の基質と生成物の RDM パターンと，入力した反応の RDM パターンの類似性を比較することで EC 番号を予測している．

§ 3.3　クラスタリング手法

本研究では 2 つのクラスタリング手法を用いるが，それにあたってクラスタリングについて説明する．クラスタリングは観測されたデータのみを扱う教師なし学習の一つで，特定の基準に従い類似しているデータどうしでクラスタを形成し，分類する手法である．データが 1 つのクラスタのみに属するクラスタリングをハードクラスタリングと呼ばれており，種類によっては，複数のクラスタに属することを許容するソフトクラスタリングも存在する．クラスタリングは主に階層的クラスタリングと非階層的クラスタリングに分けられる．階層的クラスタリングではさらに，分割型のものと凝集型のものに分けられる．分割型ではデータを全て 1 つのクラスタとみなしたのち，細かいクラスタに分割していく手法である．凝集型では，データそれぞれを 1 つのクラスタとみなし，特定の基準にしたがって複数データが属するクラスタを形成する．複数データを持つクラスタ同士も連結され，新たなクラスタを形成し，指定したクラスタ数になるまで繰り返される．

階層型では凝集型が主に用いられ，以下では，凝集型におけるクラスタを形成していく基準について述べる．

クラスタ C_1，C_2 に属するデータの集合をそれぞれ $\mathbf{x_1}$,$\mathbf{x_2}$，$\mathbf{x_1}$ と $\mathbf{x_2}$ の距離を $d(\mathbf{x_1}, \mathbf{x_2})$ としたときのクラスタ間の距離を $d(C_1, C_2)$ とする [29].

最短距離法

2 つのクラスタ内のデータどうしで，最も距離が近い組を基準として，新しきクラスターを作成する．計算量は少ないが，外れ値に弱いとされている．

$$d(C_1, C_2) = \min_{\mathbf{x_1} \in C_1, \mathbf{x_2} \in C_2} \{d(\mathbf{x_1}, \mathbf{x_2})\} \tag{3.2}$$

最長距離法

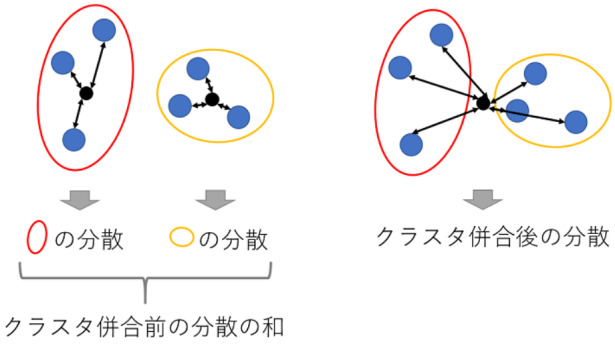


図 3.9: ウォード法のイメージ [30]

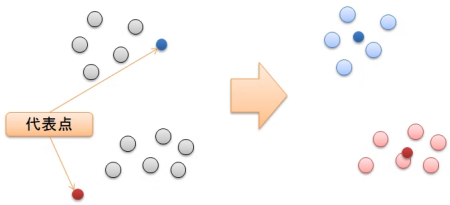


図 3.10: k-means 法のイメージ [31]

最短距離法に対して，最も距離が遠い組を基準としたもの，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．

$$d(C_1, C_2) = \max_{\mathbf{x_1} \in C_1, \mathbf{x_2} \in C_2} \{d(\mathbf{x_1}, \mathbf{x_2})\} \tag{3.3}$$

群平均法

2 つのクラスタ内の要素同士の距離を合計し，各クラスタサイズで割った平均を基準としたもの．外れ値の影響が少なく．クラスタが帯状に並ぶ鎖効果が起こりにくいとされている．

$$d(C_1, C_2) = \frac{1}{|C_1||C_2|} \sum_{x_1 \in C_1} \sum_{x_2 \in C_2} d(x_1, x_2) \tag{3.4}$$

ウォード法

あらかじめ 2 つのクラスタを結合し，結合したクラスタ内の重心に対する，データの分散 $E(C_1 \cup C_2)$ に対して，結合前の各クラスタ内のデータの分散 $E(C_i)$ を引いた差が，最小となるクラスタのペアを結合する方法．計算量は多くなるものの，分類感度が良いとされ，階層的クラスタリングで最も用いられている．ウォード方のイメージを図 3.9 に示す．クラスタ C_1 の重心を $\boldsymbol{c_i}$ として，以下のように表される．

$$\boldsymbol{c_i} = \sum_{\boldsymbol{x} \in C_i} \frac{\boldsymbol{x}}{|C_i|} \tag{3.5}$$

$$E(C_i) = \sum_{\boldsymbol{x} \in C_i} d(\boldsymbol{x}, \boldsymbol{c_i})^2 \tag{3.6}$$

$$d(C_1, C_2) = E(C_1 \cup C_2) - E(C_1) - E(C_2) \tag{3.7}$$

非階層的クラスタリングでは，あらかじめクラスタリング数を決めておき，各手法で定められている基準にしたがって，データを分類する．非階層的クラスタリングの手法をいくつか以下に示す．

k-means 法

データに対して，ランダムにクラスタを割り振り，重心に基づいてクラスタを再構成していく手法．以下の手順に沿ってクラスタリングを行う．

1. 最初に指定した k 個のクラスタリングに，データ点をランダムに割り振る

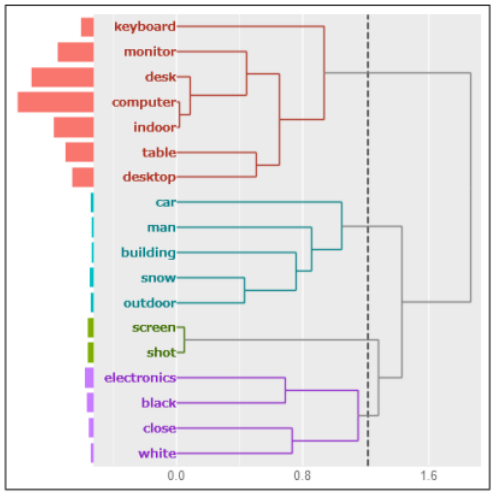


図 3.11: 階層的クラスタリングによる行動識別

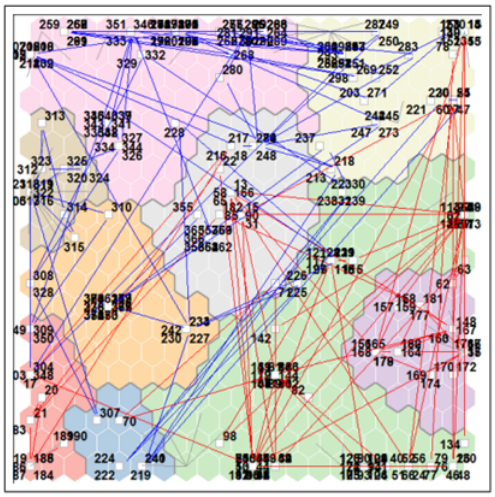


図 3.12: SOM を用いた行動時系列分析

2. 各クラスタ内の各データに対して重心を計算し，データ点が最短距離にある重心のクラスタに属するように，データ点へのクラスタを振り直す．
3. 振り直しで全てのデータ点のクラスタが固定されるまで，上記の手順を繰り返す．

自己組織化マップ (Self-Organizing Map: SOM) [32]

多次元データを低次元にマッピングし，可視化するクラスタリング手法．以下そのアルゴリズムを示す [33]．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 $m_i (i = 1, 2, ..., k)$ で構成された，2 次元平面上に写像する．このとき各ユニットの重心を $\boldsymbol{r}_i = (r_{i1}, r_{i2})$ とし，これを m_i の位置ベクトルとする．さらに，各ユニットは，重みベクトル $\boldsymbol{\xi}_i = (\xi_{i1}, \xi_{i2}, ..., \xi_{ip}), (i = 1, 2, ..., k)$ を持っているとする．ここで，$\boldsymbol{x}_j$，m_i をそれぞれ，入力層，出力層と呼び，次の手順によって出力層を更新する．

1. $j = 1$ から n までの順に，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．

$$\begin{cases} \boldsymbol{\xi}_i \leftarrow \boldsymbol{\xi}_i + h(t)\{\boldsymbol{x}_j - \boldsymbol{\xi}_i\} & i \in N_c \\ \boldsymbol{\xi}_i \leftarrow \boldsymbol{\xi}_i & i \notin N_c \end{cases} \tag{3.8}$$

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．

$$h(t) = \alpha(t) \exp\left[\frac{-||\boldsymbol{r}_c - \boldsymbol{r}_i||}{2\sigma^2(t)}\right] \tag{3.9}$$

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

クラスタリングを用いた研究例として，ヒトの行動パターンを解析し，行動識別を行ったものがある [34]．まず，画像認識 API を用いて，視界に映っている物体を認識し，その物体名をテキストデータに出力している．次に，テキストマイニングデータのクラスタリングを行うソフトウェア KH Corder [35] を用いて，物体名の同時出現頻度に関して，階層的クラスタリングを行っている．それによって，クラスタ内に含まれる物体名から行動全体のイベント性を分析している．また，SOM によるクラスタリングも行われている．観測されたデータから順番に SOM の 2 次元マップ上にプロットしていき，プロット点を線で結んでいくことで，行動の時系列を作り，複数の測定における行動の類似性を分析している．階層的クラスタリングと SOM を用いた行動分析の様子を図 3.11 および図 3.12 に示す．

第4章

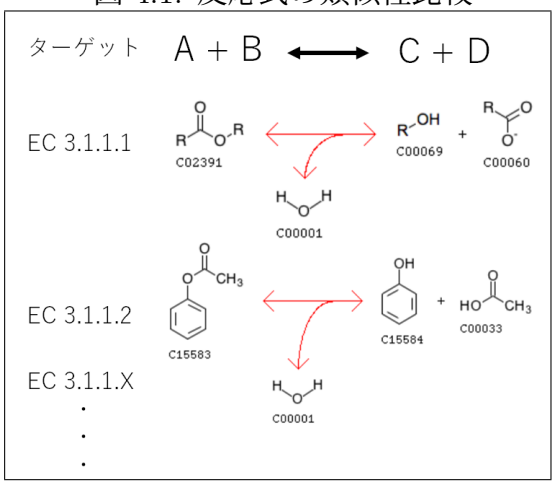
提案手法

§ 4.1　EC 番号の予測

医薬品などの新規化合物を開発する分野において，それに必要な有機合成を効率的かつ，なるべく環境に負荷を与えない形で行えるほうが望ましい．その点，生体触媒の酵素を用いると，反応物の特定の部位だけの選択的合成，反応の効率化など，グリーンケミストリーの優れた反応となるため，酵素を用いる機会が増加している．それに伴い特定の反応を行うために最適な酵素を選択することも重要となってきている．一方で，基質特異性などの酵素の性質は，生物分野に関わる内容であるため，有機合成の知識のみでは解決が難しい．酵素研究の専門家と協力して，または，酵素データベースなどを参照して最適な酵素候補の目途をつけ，その後のスクリーニングなどで，1 つの酵素に絞っていく．ここで，目的とする反応情報を与えた際に，酵素候補を予測するシステムがあれば，酵素候補の探索にかかる時間を著しく短縮することができ，次のスクリーニングの段階まで研究をスムーズに進めることができる．酵素は EC 番号で分類されており，EC 番号には生物の体内で起こる酵素を用いた代表的な反応が反応式として記載されている．EC 番号の酵素を用いた様々な生物由来の酵素製品が開発されているが，EC 番号を予測することでスクリーニング候補として，その EC 番号内の酵素に絞り込むことができる．

そこで，本研究では，Pyhton と R 言語を用いて，ターゲットとなる反応式を与えた際，生体触媒として最適な酵素の EC 番号を予測する．予測の方法として，対象とする反応式 (ターゲット反応式) と EC 番号の代表的な反応式 (EC 反応式) の類似性を比較する．図 4.1 に比較のイメージを示す．ターゲット反応式で目的とする生成物は，逆合成的な思考でどの反応物を用いれば，得られるのか分かっている．ここでは，その反応を効率良く行うための酵素候補を絞り込むことが重要となる．ターゲット反応式における反応物から生成物への変化が，EC 反応式における反応物から生成物の構造変化の特徴に類似しているならば，EC 反応式で用いられている酵素をターゲットで使用することで，反応の効率が上がり，高い収率でターゲット生成物が得られる可能性がある．これは化学の分野で用いられている類似性の概念 [36] に関係している．反応による構造変化を，反応式の類似性の評価指標とした理由として，ターゲット反応式の化合物と，各 EC 反応式の化合物どうしの比較では反応式の類似性を正確に評価できないためである．例えば，ターゲット反応式の反応物が，EC3.1.1.1 の反応物に最も類似していると評価されたとしても，生成物は EC3.1.1.2 の生成物に最も類似されていると，評価される可能性がある．また，反応物や生成物は複数ある場合が多いため，より反応式の類似性を評価することが難しくなる．そのような理由から反応による構造変化を類似性の比較として用いる．

図 4.1: 反応式の類似性比較



反応物から生成物への構造変化を捉える特徴として，記述子を用いた物性値・化学特性値の変化量 (特性値変化量) を用いる．従来研究では，反応物から生成物に変化する際の，構造記述子の変化が，反応変化の特徴として用いられている [27]．ここでは，反応物の生成物の部分構造に注目したフィンガープリントを求め，その差分を比較することで，類似する EC 反応式の酵素を予測している．しかし，フィンガープリントには様々な種類があり，それぞれ化合物のどのような特徴を説明しているのかが異なっている．つまり，1 つのフィンガープリントでは反応変化の特徴を全てとらえるのは難しい．一方で，物理・化学的な特性値を表す記述子も，化合の構造を表現する指標として考えられる．RDKit では 208 種類の特性値に関する記述子が実装されており，読み込んだ分子構造式から簡単に特性値を計算できる．そのため，RDKit 記述子を多数用いて，多次元の特性値変化量を要素に持つ特徴ベクトルを求めることで，ターゲット反応式と EC 反応式の反応時の構造変化を表現する．

差分フィンガープリントと同様に特性値変化量を以下のように定義する．各反応の反応物と生成物の個数をそれぞれ 2 個としたとき，反応物 i の特性値を RT_i，生成物 i の特性値を PD_i とする．このとき，各 n 種の記述子に対する特性値変化量 $cv_j(j = 1, 2, ..., n)$ および，m 個の反応式の特徴ベクトル $\boldsymbol{DF}_i(i = 1, 2, ..., m)$ を以下のように表す．

$$cv_j = (PD_1 + PD_2) - (RT_1 + RT_2) \tag{4.1}$$

$$\boldsymbol{DF}_i = (cv_{i1}, cv_{i2}, ...cv_{ij}, ..., cv_{in}) \tag{4.2}$$

これらをもとに，表 4.1 のような $i \times j$ の各反応式の特性値表を作成する．行ラベルはターゲット反応式 T と EC 番号，列ラベルは記述子名となる．

特徴ベクトル比較のために必要となる化合物などのデータは，KEGG と PubChem から収集する．これらのデータベースを用いる理由として 2 つ挙げられる．1 つ目は，API でデータを取得するフォーマットが整っていることである．API によって必要となるデータを簡単に取得できることは，プログラミングで自動収集するシステムの，開発のしやすさにつながり，効率的なデータ収集を行える．2 つ目はリンクによってデータベースどうしの行き来がしやすい点にある．異なるデータベースへの参照リンクが多いほど，多種多様な

表 4.1: 各反応式に対する記述子ごとの特性値

	記述子 1	記述子 2	$\cdots$	記述子 n
$\boldsymbol{DF}_1$	cv_{11}	cv_{12}	$\cdots$	cv_{1n}
$\boldsymbol{DF}_2$	cv_{21}	cv_{22}	$\cdots$	cv_{2n}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
$\boldsymbol{DF}_m$	cv_{m1}	cv_{m2}	$\cdots$	cv_{mn}

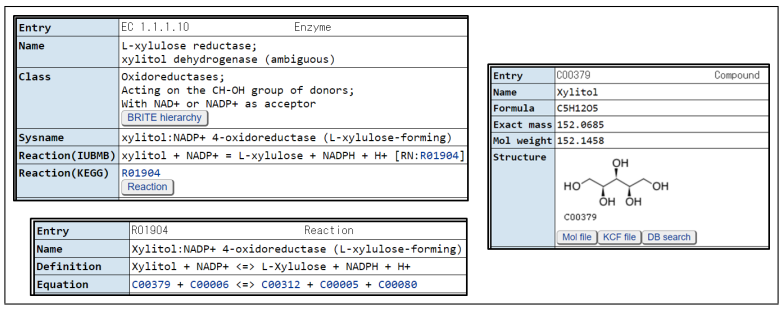


図 4.2: EC 番号，R 番号，C 番号の参照 [14](一部抜粋)

データを収集をしたり，1 つのデータベース内では見られないデータ間の関係を得ることができる．必要となるデータを API で取得し，集めたデータ関係を分析する，または，新たなデータ関係を見出すデータベースを構築することも可能となる．

KEGG では図 4.2 のように，KEGG ENZYME，KEGG REACTION，KEGG COMPOUND 間で，リンクによって EC 番号から R 番号，R 番号から C 番号とたどることができる．この関係をもとに EC 番号と代表的な反応式な反応式を構成する各化合物の ID を取得する [40]．

PubChem Compound では化合物の特性情報など KEGG にはない情報が記載されており，CID で管理されている．さらに，CID は PubChem Substance において SID とともに併記されていることが多く，SID は KEGG COMPOUND の化合物情報にリンクとして表記されている．これによって，R 番号の C 番号で書かれた反応式からそれぞれの化合物の詳細情報を得ることができる．今回は PubChem から化合物の SMILES 情報を取得し，C 番号と SID・CID の対応によって SMILES 形式の反応式と，EC 番号の対応表を作成する．図 4.3 に C 番号と SID の対応関係を表す． SMILES 形式の化合物を RDKit で読み込むことで，化合物の構造オブジェクトに変換できる．それにより，構造式をコンピュータ上で表現するとともに，RDKit の記述子を用いて特性値を計算し，化合物を数値で表現する．各反応式において，生成物と反応物の差分を取り，作成した SMILES 反応式と EC 番号の対応表より，EC 番号と各反応式の特性値変化量の特徴ベクトルを取得する．

§ 4.2　凝集型クラスタリングによる次元削減

反応変化の特徴として，特性値変化量からなる多次元の特徴ベクトルを用いるが，次元のサイズが大きすぎることが，問題となるケースがある．一般的には多重共線性や次元の

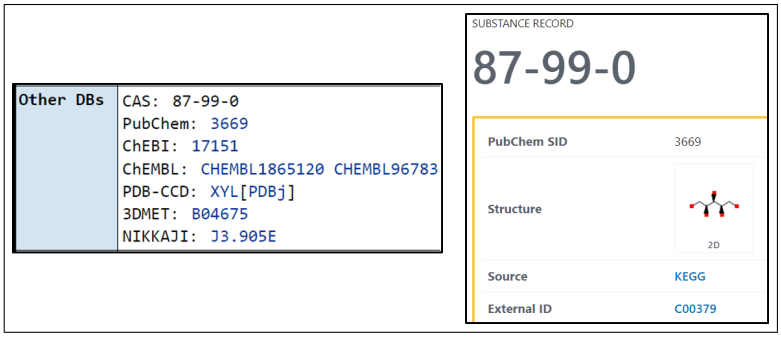


図 4.3: KEGG の C 番号と PubChemSID の対応 [17](一部抜粋)

表 4.2: 相関係数の逆数を要素に持つ距離行列

	記述子 0	記述子 2	$\cdots$	記述子 n
記述子 1	0	$1/s_{12}$	$\cdots$	$1/s_{1n}$
記述子 2	$1/s_{21}$	0	$\cdots$	$1/s_{2n}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
記述子 n	$1/s_{n1}$	$1/s_{n2}$	$\cdots$	0

呪いに絡んでくる．多重共線性とは，説明変数間に高い相関があるときに起きる現象で，汎化性能や分類精度の低下の原因とされている．次元の呪いは，用いる変数が多い場合に起こり，過学習の原因とされる問題である．今回のケースでは相関の高い記述子のペアが存在すると，同じような記述子が存在することになり，構造変化の特徴の一部として，他の記述子よりも重みづけが大きくなると考えられる．そのため，多重共線性の問題を解決しつつ，次元の削減も同時に行う．

多重共線性を解決するためには，相関の高いペアの変数に対して，どちらか片方を取り除く方法が取られることが多い．しかし，誤って重要な変数を除去していまう可能性や 3 個以上の変数間の高い相関には対処できない等の問題がある．そのため，相関に基づき，記述子間で凝集性クラスタリングを行うことで多重共線性をなくす方法を用いる [37]．ここでは，最長距離法をクラスタ間の距離としてクラスタリングを行う．プログラムは Python の sklearn に実装されている凝集性クラスタリングである，AgglomerativeClustering ライブラリ [38] を用いる．記述子 u，v 間の相関係数を s_{uv} としたとき，以下のように表される．

$$s_{uv} = \frac{\sum_{i=1}^{m}(cv_{iu} - c\bar{v}_u)(cv_{iv} - c\bar{v}_v)}{\sqrt{\sum_{i=1}^{m}(cv_{iu} - c\bar{v}_u)}\sqrt{\sum_{i=1}^{m}(cv_{iv} - c\bar{v}_v)}} (\text{ただし，} c\bar{v}_u\text{，} c\bar{v}_u \text{ は記述子 } u,v \text{ の特性値平均}) \tag{4.3}$$

s_{uv} に対して，逆数を取った，$1/s_{uv}$ を記述子間の距離とし，Python で表 4.2 のような距離行列を作成してクラスタリングする．ここでは，相関係数が 1 となる要素を 0 としている．AgglomerativeClustering では，入力データとして通常の特徴ベクトルだけでなく，距離行列を用いることができ，記述子間でマージするときの閾値を指定することができる．今回は相関係数 $s_{ij} \geqq 0.9$ すなわち，$1/s_{ij} \leqq 1/0.9 \approx 1.11$ で記述子をマージする．クラスター間でのクラスタリングにおいて，最遠距離法を用いたとき，クラスター間距離 $d(C_1, C_2)$ は，

```
somm <- som(d, xnodes, ynodes, topol="hexa", rlen=c(rlen1,rlen2))
```

図 4.4: SOM のソースコード

式 3.3 より以下のようになる．

$$d(C_1, C_2) = \max_{\mathbf{u}\in C_1, \mathbf{v}\in C_2} \frac{1}{s_{uv}} \tag{4.4}$$

これらを用いて次の手順で記述子間のクラスタリングを行う．

1. $1/s_{ij} \leqq 1.11$ を満たす，記述子のペアにおいて，互いの距離が最短となるものをマージする．
2. $d(C_1, C_2)$ が最小となるクラスタ C_1，C_2 をマージする．記述子全体で 1 つとするクラスタが完成するまで，これを繰り返す．
3. クラスタリングを終了後，クラスタ番号とそのクラスタに所属する記述子の対応表を取得する．
4. 元の入力データに対して，あるクラスタ番号に属する記述子の特性値列を全て抜き出し，標準化を行い平均化したのち，新たな合成記述子として使用する．

表 4.1 において，同クラスタの記述子同士をまとめ，合成記述子 clusterX(X はクラスタ番号) と置き換えることで，次元削減を行う．

§ 4.3 SOM による反応式クラスタリング

次元削減した特徴ベクトルを用いて，反応式間の類似度を比較する手法として，SOM によるクラスタリングを行う．SOM を用いることの利点として，2 つ挙げられる．1 つ目は，低次元空間への可視化が可能になる点である．高次元の特徴ベクトルの場合，反応式どうしの位置関係が把握しにくいが，2,3 次元まで圧縮することで，その関係を把握することが可能となる．2 つ目として，クラスタリングによる類似比較が挙げられる．類似度を比較する手法として，コサイン類似度や相関係数等が用いられるが，複数の反応式間の類似度を調べたいときには，直感的な理解が難しい場合がある．その際に，クラスタリングを用いることによって，全ての反応式の類似性を把握することができる．これらのことから，ターゲット反応式の近くに分布する類似性の高い EC 反応式を複数同時に確認できる他，他の反応式どうしの類似性も見ることができるようになる．用いる SOM のプログラムとして，KH Corder で出力される SOM の R 言語ファイルを参考に作成された，R 言語使用のソースコードを用いる [34]．入力するデータは次元削減及び標準化を行った，記述子 n 個に対する各ターゲット・EC 反応式の特性値変化量となる．SOM のプログラム中には R 言語のパッケージとして実装されている som を使用している [39]．プロット点のラベルはターゲット (T) と反応式の EC 番号を扱う．SOM の学習を行うソースコードは図 4.4 のようになる．

d は入力データ，xnodes, ynodes はユニット (ノード) 数を決める変数で，(ノード数)= $xnodes \times ynodes$ となる．topol はユニットの形状を表し，rect が四角形，hexa は六角形を表

す．rlen は学習回数を表している．大まかな順位付けを行う段階 (rlen1) と収束段階 (rlen2) の 2 段階に分かれており，KH Corder のデフォルト (KH Coder3 リファレンス・マニュアルに記載) では rlen1= 1,000，rlen2= 500,000 となっている．今回は，(ノード数)= 20×20，topol=hexa，rlen1= 1,000，rlen2= 200,000 とする．変数 somm に実行結果が格納され，その後クラスタ結果の描画を行う．クラスタ数は 9 に指定してそれぞれ色分けを行う．なお，今回は時系列データとなっていないため，プロット点どうしを結ぶ線は除外する．

第5章

実験結果並びに考察

§ 5.1　数値実験の概要

本研究の実験の流れについて説明する．まず，化合物の特徴ベクトルを求めるため，KEGG と PubChem から各反応式の情報を取得する．次に，反応式内の反応物・生成物の SMILES を出力する．さらに，RDKit にある 208 種の記述子を用いて，化合物の物理・化学特性値を計算し，特性値変化量を求めることで，ターゲット反応式と EC 反応式を，208 次元の特徴ベクトルで表現する．さらに，不適切な値を含む記述子を除外し，凝集型クラスタリングによって相関の高い記述子同士をまとめて，新たな合成記述子を作成することで，次元削減を行う．最後に，SOM によって反応式をクラスタリングし，ターゲット反応式に対して適切な酵素を予測する．

具体的なデータ整理，前処理，および分析条件について以下で説明していく．

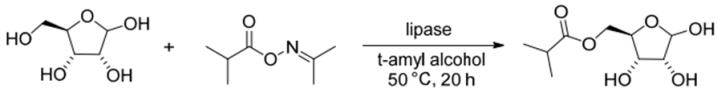
ターゲット反応式と提案手法の評価方法

今回は，モルヌピラビルを生成する過程における，1 ステップ目の合成の反応式に焦点を当てる [2]．図 5.1 に対象とする反応式と酵素スクリーニングの結果を示す．この反応はリボース (左辺第 1 項) の第一級アルコール部分を選択的にエステル化する反応である．ここでは，8 つの酵素製品に対する，生成物のアッセイ収率を調べるための，スクリーニングを行っている．最終的に Novozym435 の酵素製品が一番優れた結果となったが，これは BRENDA によると EC3.1.1.3 に分類される酵素とされている．特性値変化量が，酵素番号予測を行うために，十分な特徴を備えている場合，この反応式をターゲットとして他の EC 反応式とともにクラスタリング行えば，EC3.1.1.3 の反応式ターゲットの付近に位置すると考えられる．

一方で，この反応式における，2 つの反応物を用いた反応は，通常の方法では起こりえないと考えられる．そこで，図 5.2 のように左辺第 2 項の化合物をイソ酪酸に置き換えた，反応式をターゲットする．一般的な加水分解反応の視点から見れば，リボースに対してイソ酪酸を用いる場合の方が，反応として起こりうると推測される．この合成反応を確認した後に，収率を上げるために等価体として，置き換え前の化合物を用いたと考えられる．よって，図 5.2 の化合物に置き換えた反応式をターゲットとして EC 反応式と比較し，最適な酵素を予測する．

比較対象となる EC 反応式

今回の予測は比較する EC 反応式をあらかじめ絞ったうえで行う．ターゲットの反応はエステル加水分解の逆反応となるエステル化反応のため，用いる酵素として，EC3.1.1 の加



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

図 5.1: ターゲットと酵素スクリーニング

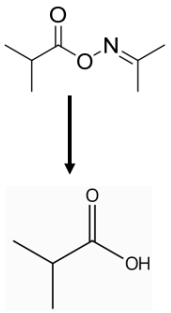


図 5.2: 反応物の置き換え

水分解酵素が適当であると考えられる．これは，加水分解が一般的には可逆的な反応であり，加水分解酵素でエステル化も可能であるためである．したがって，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を行う．一方で，全ての EC 反応式に対して比較することも重要であると考えられる．しかし，EC3.1.1 以外の番号で類似していると認識される可能性を考慮し，今回は，EC3.1.1 と認識された場合を仮定し，そのうえで提案した記述子変化量とその記述子選択によって，適切な酵素を予測できるか検証する．

データの対応表取得と整理

まず，KEGG の ID や反応式を取得するソースコードを用いて [40]，EC3.1.1 に属する EC 番号，対応する R 番号，C 番号で構成された R 番号の反応式を取得し，それぞれの対応表を作成した．同様に PubChem からは C 番号と CID，SID の対応表を取得した．次に C 番号と CID または SID を参照し，PubChemPy によって，各反応式の反応物と生成物の SMILES を取得することを試みた．しかし，C 番号に対する CID がまだ登録されていない化合物や，SID を引数にして，PubChemPy から SMILES を取得できないなどの問題が発生した．そこで，PubChem の SID で検索したリンク内で SDF ファイルを入手し，それを RDKit で読み込み SMILES に変換した．また，ターゲットの SMILES は SciFinder^n で入手した MOL ファイルを RDKit で変換することで取得した．それらの SMILES をまとめて，ターゲットおよび EC 番号に対する，各反応物・生成物の SMILES 対応表を作成した．以下，表 5.1，表 5.2, および表 5.3 にそれぞれの対応表を示す．

対応表の前処理 1

得られた SMILES 対応表には KEGGC COMPOUND に登録されていない (番号が新しい) 化合物，あるいは登録されているが，構造式が記載されていない化合物が存在する．そのため，対応表内の SMILES の項に空白となる部分が発生するため，その項を含む反応式は除外した．また，ターゲットは反応物 2 個，生成物 1 個の組み合わせであるため，EC 反応式はターゲット同様の組み合わせにする．これは，提案手法の特性値変化量の計算に用いられる化合物の数が増加または減少することで，構造変化とは別の要因による変化が影響すると考えられるためである．つまり，ターゲットの物理・化学特性値の純粋な変化と比較して，化合物の多寡による特性値の変化も追加されると推測されるためである．以上の理由から，ほとんどの反応式に含まれている H_2O を除外した場合の，反応物が 2 個，生成物が 1 個の組み合わせとなる EC3.1.1 反応式 113 種のみを採用した．

表 5.1: EC 番号と KCID

	ENZYME	left1	left2	right1	right2	right3
0	3.1.1.22	C04546	C00001	C01089	None	None
1	3.1.1.20	C01572	C00001	C01424	None	None
2	3.1.1.40	C02868	C00001	C01839	None	None
3	3.1.1.33	C02655	C00001	C00031	C00033	None
4	3.1.1.6	C01883	C00001	C00069	C00033	None
...	...	...	...	...	...	...
173	3.1.1.111	C18125	C00001	C22237	C00162	None
174	3.1.1.115	C22218	C00001	C22219	None	None
175	3.1.1.117	C22373	C00001	C22374	C00069	None
176	3.1.1.118	C01194	C00001	C22400	C00162	None
177	3.1.1.118	C00416	C00001	C03974	C00162	None

表 5.2: 各化合物 ID 対応表

	cid	pubchem_SID	pubchem_CID
1	C00001	3303	962
2	C00002	3304	5957
3	C00003	3305	5893
4	C00004	3306	439153
5	C00005	3307	5884
...	...	...	...
18594	C22269	405226444	6365572
18595	C22272	405226445	11788398
18596	C22273	405226446	11411510
18597	C22274	405226447	135567131
18598	C22275	405226448	44468216

表 5.3: EC 番号と SMILES の対応表

	ENZYME	left1	left2	right1	right2	right3
0	3.1.1.22	C[C@@H](O)CC(=O)O[C@H](C)CC(=O)O	[H]O[H]	C[C@@H](O)CC(=O)O	N	N
1	3.1.1.20	O=C(O)c1cc(O)c(O)c(OC(=O)c2cc(O)c(O)c(O)c2)c1	[H]O[H]	O=C(O)c1cc(O)c(O)c(O)c1	N	N
2	3.1.1.40	Cc1cc(OC(=O)c2c(C)cc(O)cc2O)cc(O)c1C(=O)O	[H]O[H]	Cc1cc(O)cc(O)c1C(=O)O	N	N
3	3.1.1.33	CC(=O)OC[C@H]1O[C@@H](O)[C@H](O)[C@@H](O)[C@@H]1O	[H]O[H]	OC[C@H]1OC(O)[C@H](O)[C@@H](O)[C@@H]1O	CC(=O)O	N
4	3.1.1.6	*OC(C)=O	[H]O[H]	*O	CC(=O)O	N
...	...	...	...	...	...	...
173	3.1.1.111	*C(=O)OCC(O)COP(=O)(O)OC[C@H](N)C(=O)O	[H]O[H]	N[C@@H](COP(=O)(O)OCC(O)CO)C(=O)O	*C(=O)O	N
174	3.1.1.115	O=C1OC[C@](O)(CO)[C@H]1O	[H]O[H]	O=C(O)[C@H](O)C(O)(CO)CO	N	N
175	3.1.1.117	N	[H]O[H]	N	*O	N
176	3.1.1.118	*C(=O)OC[C@H](COP(=O)(O)O[C@H]1[C@H](O)[C@@H](...	[H]O[H]	N	*C(=O)O	N
177	3.1.1.118	*C(=O)OCC(COP(=O)(O)O)OC(*)=O	[H]O[H]	*C(=O)O[C@H](CO)COP(=O)(O)O	*C(=O)O	N

物理・化学特性値および記述子変化量の計算

ターゲット+113 種の SMILES 対応表を元に，RDKit の rdkit.chem.descriptor から 208 種の記述子名を取得し，反応式 1，反応式 2，生成物それぞれの場合で特性値を計算した．その後，特性値変化量を求め，図 5.4 のような 208 次元ベクトルを持つ反応式の表を作成した．

対応表の前処理 2

図 5.4 の表から，nan 値や発散している要素を持つ記述子 12 種を除外した．また，全ての反応式において等しい特性値を持つ記述子を除外し，最終的に 128 種類の記述子で次元削減を行う．

§ 5.2　実験結果と考察

特徴ベクトルの次元削減

相関係数の逆数である距離行列を入力とした，最遠距離法の凝集型クラスタリングによる次元削減を行った．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**??**に，クラスタ番号と，そのクラスタにマージされた記述子の対応表を示す．12 番のクラスタに所属する記述子「MaxEStateIndex」と「MaxAbsEStateIndex」は相関係数が 1 であるが，ともにマージされていることが分かる．また，記述子名が類似している記述子が，同じク

表 5.4: 各反応式の特性値変化量

	MaxEStateIndex	MinEStateIndex	MaxAbsEStateIndex	MinAbsEStateIndex	qed	MolWt	HeavyAtomMolWt	ExactMolWt	NumValenceElectrons
Target	-8.378152	0.949632	-8.378152	-0.144028	-0.330982	-18.015	-15.999	-18.010565	-8
3.1.1.33	-7.632875	0.794822	-7.632875	-1.064815	-0.343138	-18.015	-15.999	-18.010565	-8
3.1.1.6	-6.597222	0.946759	-6.597222	-0.949074	-0.409219	-18.015	-15.999	-18.010565	-8
3.1.1.1	-6.486111	0.972222	-6.486111	-0.675926	-0.331106	-17.007	-15.999	-17.003288	-8
3.1.1.7 3.1.1.8	-7.085822	0.351574	-7.085822	-0.914074	-0.484689	-18.015	-15.999	-18.010565	-8
...	...	...	...	...	...	...	...	...	...
3.1.1.106	-8.896201	0.794521	-8.896201	-0.839784	-0.462056	-18.015	-15.999	-18.010565	-8
3.1.1.113	-6.747917	0.372685	-6.747917	-0.872685	-0.398840	-18.015	-15.999	-18.010565	-8
3.1.1.112	-7.033650	0.317731	-7.033650	-0.979769	-0.421762	-18.015	-15.999	-18.010565	-8
3.1.1.111	-8.902683	0.535378	-8.902683	-0.657129	-0.360318	-18.015	-15.999	-18.010565	-8
3.1.1.118	-8.839073	0.535378	-8.839073	-0.575822	-0.317022	-18.015	-15.999	-18.010565	-8

表 5.5: 次元削減のためのクラスタリング結果

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	MaxESta	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	Fraction	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									

ラスタに属している傾向があることが分かる．表 5.6 にはクラスタリングでマージされた記述子，および次元削減の結果を示す．合成された記述子は，クラスタ番号 X を後ろにつけた「clusterX」で表示されている．また，ターゲットを T とし，EC 番号の下 1 桁のみ表示している，ピリオド以下の番号は，EC 番号の代表の反応式が，複数ある場合の区別に用いられている．アンダーバーで区切られているものは，その反応式が複数の EC 番号間で重複している場合の区別となっている．

SOM による反応式のクラスタリング結果

SOM のプログラムによって，特徴変化量に基づいて，反応式をクラスタリングした結果は図 5.3 のようになった．E は EC3.1.1 以外の反応式を表す．色分けされた各クラスタ領域において，青色のクラスタが離れているが，これは，SOM が本来は 2 次元平面を円柱状に丸め，さらに円柱を曲げて切り口をつないだトーラス型の形状をしており，2 次元平面にした際に分裂したものと考えられる．ターゲット (T) と同じクラスタに属し，かつ付近に位置する EC 反応式として，EC3.1.1.80，EC3.1.1.93 という結果となった．これらのターゲット反応式と比較すると図 5.5 のようになった．EC 3.1.1.93 の右辺 (反応物) である C00191 は，リボースと同じ糖類に分類されるグルコースの構造が含まれている．また，ターゲット反応式を，モルヌピラビルの 1 ステップ目の反応 (イソ酪酸置き換え前) として，同様の条件で SOM によるクラスタリングを行った．その結果を図 5.4 に示す．ターゲットと同じクラスタに属するのは，EC 3.1.1.75(2 つ)，EC 3.1.1.76，EC 3.1.1.101 となり，これらの反応式

表 5.6: 次元削減後の特徴ベクトル

	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

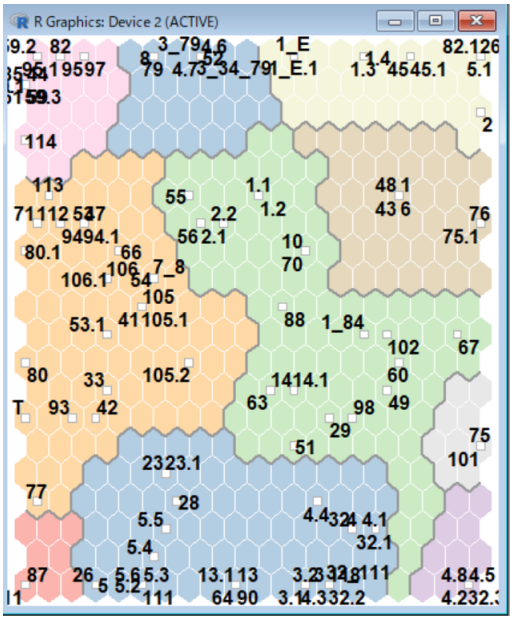


図 5.3: SOM による反応式のクラスタリング (イソ酪酸)

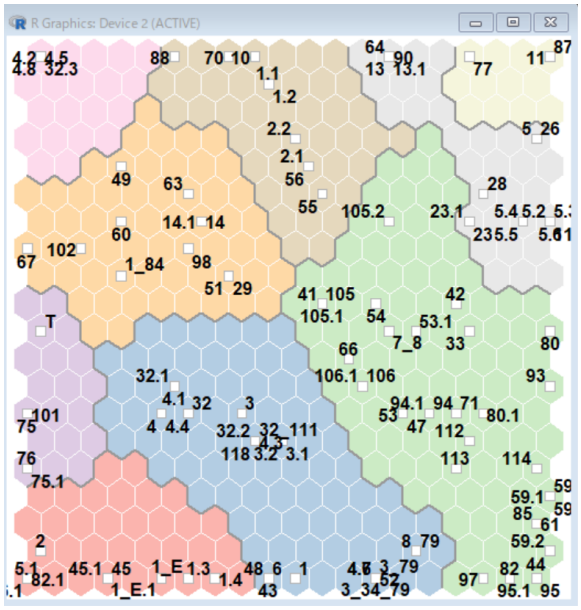


図 5.4: SOM による反応式のクラスタリング (イソ酪酸置き換え前)

中には重合体が必ず含まれる形となった．いずれの結果においても EC 3.1.1.3 はターゲットと異なるクラスタに属する結果となった．

考察

今回目的とする EC 3.1.1.3 が，SOM のマップ上でターゲットの付近に位置しなかった原因として，4 つのことが挙げられる．

1 つ目は反応式中の係数を反映していなかったことが挙げられる．用いられる全ての化合物の比率を平等にした場合でも，特性値変化量は構造変化の特徴として機能すると考えられる．しかし，反応式中の 1 つの化合物が用いられる分子数だけ特性値を上乗せすることで，より化学的に構造変化を捉えられると考えられる．

2 つ目は，ターゲットの反応において，提案した特徴変化量では，捉えきれていない要因が多く影響している点である．モルヌピラビルの論文のサポート資料 [41] では，tert-アミルアルコールを溶媒として用いており，50 ℃で 20 時間振とうを行うことでターゲットの生成物を生成している．また，用いた反応物の分量なども異なっている．一方で，EC 反応式は生物の体内等で起こる反応であり，基本的には有機溶媒等を用いない．実験に用いた試

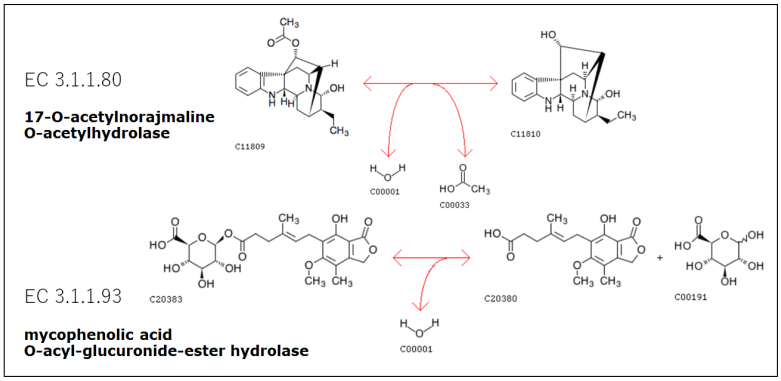


図 5.5: ターゲットの近くに位置している反応式

薬や，溶媒，実験環境，配合比率などの様々な要因によって，天然に起こる反応との差異を発生させていると考えられる．特性値変化量だけでなく，反応以外の要因も考慮した特徴作成が必要である．

3 つ目は，相関係数に基づくクラスタリングで，同一クラスタに存在する記述子を合成した際に，構造変化に重要な特徴の影響力を弱めてしまった可能性が挙げられる．今回は，相関の高い記述子ペアに対し，片方を除外することで重要な記述子を誤って削除するのを避けるため，クラスタ内の記述子どうしで標準化，および平均化を行った．しかし，やはり重要な記述子と重要でない記述子が混ざったクラスタが存在し，平均化によって，重要な記述子の説明力を薄めてしまった可能性がある．改善策としては，相関係数に応じて形成されたクラスタ内で，記述子の重要度に基づいて，重みづけをする手法を提案できれば，適切な記述子選択と次元削減ができると考えられる．

4 つ目として，次元削減後に用いられた記述子は 80 種とまだまだ多く，不必要な記述子によって構造変化を上手く説明できなかったことが考えられる．今回は主に多重共線性の対策としての手法を提案したが，少数かつ構造変化を十分に説明できるような，記述子の組み合わせを提案する手法を検討していきたい．

第6章

おわりに

近年，新型コロナウイルスになどの影響で，新薬開発の需要が高まり，化学反応の設計や予測を行う研究が発展を続けている．一方で，反応の効率化と環境面から，酵素の生体触媒を用いて合成が行われる機会が増えており，目的の反応に対して最適な酵素を予測することが重要視されている．しかし，基質特異性などの酵素の性質は生物分野にかかわるため，有機合成の知識のみでは解決が難しく，酵素研究の専門家と協力する，または，酵素データベースを参照するなどして最適な酵素候補は探索されていた．

目的とする反応に対して，酵素候補を予測するシステムがあれば，次のステップである1つの酵素に絞るスクリーニングまでスムーズに進めることができる．また，酵素はEC番号と，生体内で自身が使用されて起こる代表的な反応の反応式で管理されている．

これらのことから，本研究では，ターゲットとなる反応式を与えた際に，EC番号の代表的な反応式と比較し，最も類似する反応式のEC番号を最適な酵素として予測する．予測の方法として，化合物の物理・化学的な特性値を計算し，反応物から生成物の差分を取った特性値変化量をもとに，反応式どうしの類似性を比較することを提案した．

KEGGやPubChemなどで必要とするデータを取得し，RDKitを用いて各反応に対して208種の特性値変化量を計算し，特徴ベクトルを作成した．その後，凝集型クラスタリングによって特徴ベクトルの次元を80次元まで削減し，SOMによって反応式のクラスタリングを行った．

2パターンでの検証を行い，1つ目では，ターゲットの反応式に対して，EC3.1.1.80，EC3.1.1.93の反応式が，2つ目では，EC 3.1.1.75，EC 3.1.1.76，EC 3.1.1.101が類似していると判定された．本来予測されるはずのEC 3.1.1.3を予測することはできなかったが，ターゲット反応式・EC反応式，または各反応式間で共通する特徴を確認することができた．

今後の課題として，化学反応時の構造変化を特徴としてより詳細に捉えるため，重要な記述子を残しつつ，さらに次元を削減していく手法を開発することが挙げられる．また，EC番号の反応式のクラス分類に着目し，最も精度よく分類できる記述子の組み合わせを特定したのち，ターゲットの反応式の酵素予測において検証していくことが考えられる．

謝辞

本研究を遂行するにあたり，多大なご指導と終始懇切丁寧なご鞭撻を賜った富山県立大学工学部電子・情報工学科情報基盤工学講座の奥原浩之教授，António Oliveira Nzinga René 講師に深甚な謝意を表します．そして，有機化学・酵素化学に関して貴重なご意見をいただいた工学部生物工学科酵素化学工学講座の浅野泰久教授，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補佐の岩崎源司博士 (薬学) に感謝申し上げます．最後になりましたが，多大な協力をしていただいた研究室の同輩諸氏に感謝致します．

2022 年 2 月

武藤 克弥

参考文献

[1] “ケモインフォマティクス市場、2021 年から 2026 年の間に CAGR13 %で成長見込み”, https://prtimes.jp/main/html/rd/p/000002048.000071640.html, 閲覧日 2022.1.6.

[2] Tamas Benkovics, John A. McIntosh, Steven M. Silverman, Jongrock Kong, Peter Maligres, Tetsuji Itoh, Hao Yang, Mark A. Huffman, Deeptak Verma, Weilan Pan, Hsing-I Ho, Jonathan Vroom, Anders Knight, Jessica Hurtak, William Morris, Neil A. Strotman, Grant Murphy, Kevin M. Maloney, and Patrick S. Fier1,, “Evolving to an Ideal Synthesis of Molnupiravir, an Investigational Treatment for COVID‐19”, *ChemRxiv*, 2020.

[3] 北川勲，磯部稔, ”天然物化学・生物有機化学 I”. 朝倉書店, 2008. 3-4 ページ

[4] 西村淳, 樋口弘行, 大和武彦, ”有機合成化学入門 -基礎を理解して実践に備える” 丸善株式会社, 2010. 1 ページ

[5] “日本化学会・ケモインフォマティクス部会”, https://cicsj.csj.jp/, 閲覧日 2022.1.23.

[6] 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.

[7] 佐藤寛子, “化学情報学 -化学反応の系図と反応予測” 国立情報学研究所, 2003. 21-23 ページ

[8] 藤波 美起登, 清野 淳司, “量子化学計算情報を記述子とした機械学習に基づく反応予測手法の開発”, *Journal of Computer Chemistry, Japan*, Vol. 15, No. 3, pp. 63-65, 2016.

[9] “酵素の化学”, http://www.sc.fukuoka-u.ac.jp/~bc1/Biochem/biochem5.htm, 閲覧日 2022.1.31.

[10] “酵素基質とは”, https://bizcomjapan.co.jp/iris-biotech/knowledge/substrate/, 閲覧日 2022.1.31.

[11] “新設された酵素分類 EC7 の和名提案について”, https://www.jbsoc.or.jp/notice/ec_translocase.html, 閲覧日 2022.1.15.

[12] 白兼孝雄, “酵素の分類と命名法”, JAS 情報, 2017

[13] “Enzyme Nomenclature”, https://iubmb.qmul.ac.uk/enzyme/, 閲覧日 2022.1.15.

[14] ”KEGG: Kyoto Encyclopedia of Genes and Genomes”, https://www.genome.jp/kegg/kegg_ja.html, 閲覧日 2022.1.17.

[15] "CAS SciFinder^n" https://scifinder-n.cas.org/ 閲覧日 2022.1.23.

[16] "CAS SciFinder^n 検索ガイド" https://www.jaici.or.jp/scifinder-n/ref/sfn.pdf 閲覧日 2022.2.3.

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

[18] "About PubChem", https://pubchemdocs.ncbi.nlm.nih.gov/about, 閲覧日 2022.2.6

[19] Sunghwan Kim, Paul A. Thiessen, Evan E. Bolton, Jie Chen, Gang Fu, Asta Gindulyte, Lianyi Han, Jane He, Siqian He, Benjamin A. Shoemaker, Jiyao Wang, Bo Yu, Jian Zhang, Stephen H. Bryant, "PubChem Substance and Compound databases", *Nucleic Acids Research*, Vol. 44, No. 1, pp. 1202-1213, 2016.

[20] "BRENDA The Comprehensive Enzyme Information System", https://www.brenda-enzymes.org/index.php, 閲覧日 2022.2.1

[21] "KEGG API", https://www.kegg.jp/kegg/rest/keggapi.html, 閲覧日 2022.2.1

[22] Sunghwan Kim, Paul A. Thiessen, Evan E. Bolton, Stephen H. Bryant, "PUG-SOAP and PUG-REST: web services for programmatic access to chemical information in PubChem", *Nucleic Acids Research*, Vol. 43, No. 1, pp. 605-611, 2015.

[23] "SMILES 記法は化学構造の線形表記法" https://future-chem.com/smiles-smarts/, 閲覧日 2022.1.27

[24] Joseph L. Durant, Burton A. Leland, Douglas R. Henry, and James G. Nourse, "Reoptimization of MDL Keys for Use in Drug Discovery", *American Chemical Society*, Vol. 7, No. 12, 2012.

[25] "The RDKit Documentation", https://www.rdkit.org/docs/GettingStartedInPython.html#list-of-available-descriptors, 閲覧日 2022.2.6

[26] "PubChemPy documentation", https://pubchempy.readthedocs.io/en/latest/#, 閲覧日 2022.2.6

[27] Qian-Nam Hu, Hui Zhu, Xiaobing Li, Manman Zhang, Zhe Deng, Xiaoyan Yang, and Zixin Deng, "Assignment of EC Numbers to Enzymatic Reactions with Reaction Difference Fingerprints", *J. Chem. Inf. Comput. Sci.*, Vol. 42, No. 6, 2002.

[28] Yoshihiro Yamanishi, Masahiro Hattori, Masaaki Kotera, Susumu Goto, Minoru Kanehisa, "E-zyme: predicting potential EC numbers from the chemical transformation pattern of substrate-product pairs", *Bioinformatics.*, Vol. 25, pp. 179-186, 2009.

[29] "クラスタリング (クラスター分析)", https://www.kamishima.net/jp/clustering/#bib_cutting, 閲覧日 2022.2.3

[30] “クラスタリングとは — 概要・手順・活用事例を紹介”, https://ledge.ai/clustering/, 閲覧日 2022.2.3

[31] “クラスタリング手法の列挙 (一部)”, https://qiita.com/sotoattanito/items/b885ef2dd3fe11cb817d, 閲覧日 2022.2.8

[32] Teuvo KOHONEN, ”Self-organized formation of topologically correct feature map”, *Biological Cybernetics*, Vol. 43, pp. 59–69, 1982.

[33] 亀岡瑶, 宗像昌平, 八木圭太, 山本儀郎, “自己組織化マップによる顧客の分類とその可視化”, 計算機統計学, Vol. 29, No. 2, pp. 181-188, 2016.

[34] 福嶋 瑞希,”環境認識ライフログからの行動パターン解析による類似性・イベント検出”, 富山県立大学学位論文 2018.

[35] ”KH Coder” "http://khcoder.net/" 閲覧日 2022.1.30

[36] Mark A. Johnson, Gerald M. Maggiora, “Concepts and Applications of Molecular Similarity”, *Wiley*, New York, 1990.

[37] “[Python コード付き] 相関係数で変数選択したり変数のクラスタリングをしたりしてみましょう”, https://datachemeng.com/variable_selection_and_clustering_based_on_r/, 閲覧日 2022.1.29

[38] “sklearn.cluster.AgglomerativeClustering ”, https://scikit-learn.org/stable/modules/generated/sklearn.cluster.AgglomerativeClustering.html, 閲覧日 2022.2.3

[39] “Package ‘ som ’ ”, https://cran.r-project.org/web/packages/som/som.pdf, 閲覧日 2022.2.3

[40] ”KEGG API を用いてデータ取得”, https://rstudio-pubs-static.s3.amazonaws.com/472676_97a2c135b5704dc1b52f7759b73466e8.html#kegg-compound, 閲覧日 2022.12.28.

[41] “Supporting Information for: Evolving to an Ideal Synthesis of Molnupiravir, an Investigational Treatment for COVID‐19”, https://europepmc.org/api/fulltextRepo?pprId=PPR257265&type=FILE&fileName=EMS109513-supplement-Supporting_Information.pdf&mimeType=application/pdf, 閲覧日 2022.2.6