化学反応に最適な酵素を予測するための
機械学習を用いた EC 番号予測モデルの開発

武藤　克弥[1]，岩﨑　源司[2]，浅野　泰久[2]*，奥原　浩之[3]*

[1]富山県立大学工学研究科電子・情報工学専攻

[2]富山県立大学工学部生物工学科

[3]富山県立大学工学部情報システム工学科

（〒 939-0398　富山県射水市黒河 5180）

要旨

4 桁からなる EC 番号には酵素名とその酵素が触媒する化学反応も記載されている．本研究では，有機合成に用いる化学反応に対して最適な酵素候補を EC 番号として予測するモデルの作成を行った．そして，Kyoto Encyclopedia of Genes and Genomes (KEGG）および BRENDA など文献に記載されている酵素反応データを用い，正解 EC 番号の予測に関するモデルの評価を行った．今回は，基質 2 種類，生成物 2 種類からなる EC 3 に属する酵素反応に対し，EC 番号の subclass（2 桁目）および sub-subclass （3 桁目）を予測する Random Forests (RF)予測モデルを開発した．初めに，KEGG より EC 番号と反応式の文字データを取得し，数値に変換した．数値化の際には，各反応式で，基質が生成物に変化する際の 208 種類の記述子(物理・化学特性値)の変化量を計算し，208 次元の反応式の特徴ベクトルを作成した．次に，SMOTE を適用し，特徴ベクトルのデータ数を 962 から 3100 にオーバーサンプリングした．さらに，モデル作成の前処理として記述子選択を行い，RF に対して forward selection を適用し，23 種類の記述子が選択された．また，パラメータ調整では決定木の最大深さ 15，決定木数 800 となった．これらのデータ・パラメータ調整で作成したモデルの予測結果として，KEGG のテストデータに対し，F1 スコア平均 0.99 が得られた．また，BRENDA などの文献反応 12 種に対しても，現状十分な予測精度が得られた．

キーワード：EC 番号，有機合成，機械学習，特徴選択

---

*連絡先　E-mail: asano@pu-toyama.ac.jp, okuhara@pu-toyama.ac.jp

緒言

有機合成化学分野では，化学反応の設計や予測に酵素を生体触媒として利用する機会が著しく増加している．化学触媒と比較して，生体触媒はしばしば触媒効率が高く、かつ環境に優しい条件で利用できるため，化学反応を効率的に進めることができる．そのため，目的の生成物を生み出す反応に対して，最適な酵素を予測することが重要になっている. 酵素には 4 桁の Enzyme Commission number (EC 番号) [1] が割り振られており，どの反応を触媒するか，どの結合や基質に作用するかによって酵素が分類される．本研究では機械学習を用い，特定の化学反応に対して最適な酵素候補を EC 番号として予測するモデルの開発を行った．機械学習による EC 番号予測は，主に代謝経路の解析等において，未知酵素に EC 番号の割り当てを行い，酵素の性質を特定する目的で行われてきた．一方で本研究では，有機合成において，基質に対して用いるべき酵素を絞り込むために EC 番号を予測する．機械学習による EC 番号予測モデルが実用化できれば，将来的に酵素候補を予測された EC 番号内の酵素に絞り込むことができ，実験による探索時間やコストの短縮が期待できる．

EC 番号予測の方法として，基質から生成物に変化する際の物理化学的特性値や化学構造の変化に着目したもの [2,3]，タンパク質配列を用いたものなどがある [4]．これらの手法は酵素への EC 番号割り当てを主軸としているため，本研究では，有機合成の目的で EC 番号予測手法を用いる．具体的には，従来と同様に物理化学的特性値の変化に着目した手法と，Kyoto Encyclopedia of Genes and Genomes (KEGG) [5]の酵素反応データを用いて, EC 番号予測モデルを構築する．また，モデル構築に用いなかった KEGG のデータ(テストデータ)に対して EC 番号を予測する．そして，新たに BRENDA [6]やその他文献の酵素反応データに対しても EC 番号予測を行う．生体由来の基質に関する酵素反応を多数収録している KEGG に対し，BRENDA は合成基質に作用する酵素反応も多く扱っており [7]，より有機合成に対する予測に適していると考えられる．

本研究の手順として，初めに機械学習における学習を行うためのデータを収集した．KEGG には，EC 番号に登録されている酵素名の他に，その酵素が作用して発生する化学反応式が登録されているため，これらを用いた．次に，取得したデータを加工し，化学反応式を機械学習器に入力可能な数値データに変

換した．そして，各反応式において，基質から生成物に変化する際の物性値・化学特性値の変化 (特性値変化量)に着目し，数値化を行った．ここで，特性値とは分子量や部分電荷，LogP などの物性値・化学特性値を指す．複数の特性値の変化量を計算することで，反応式は多数の特性値変化量を要素に持った多次元ベクトルとなる．その後，ベクトル化した反応式に対して EC 番号を紐づけしたデータ (反応データ)を作成し，機械学習を行うことで，EC 番号予測モデルを作成した．最後に，正解の EC 番号が分かっている，KEGG のテストデータをモデルに入力し，正しい EC 番号が予測できるかどうかの予測精度の評価を行った．そして，BRENDA や文献に記載されている酵素反応データを入力し，新たに用いた酵素反応に対する EC 番号の予測精度を確認した．

実験方法

実験概要　本研究では，EC 3 の subclass (2 桁目) および sub-subclass (3 桁目)の番号の予測に焦点を当てる．使用する subclass および sub-subclass の組み合わせは，20 通り存在する．208 種類の特性値を基に作成した反応式の特徴ベクトルをランダムフォレスト(RF) [8] に学習し，EC 番号予測モデルを作成する．モデル作成の前処理として，データの整形とオーバーサンプリングを行い，208 種類から必要な記述子のみを選ぶ特徴選択，および RF のパラメータ調整を実行する．作成したモデルの予測精度の評価は 2 種類の方法で行う．第 1 番目に，学習に用いなかった KEGG の反応データ(テストデータ)を予測モデルに入力し，20 種類の EC 番号に対するクラス分類を行う．第 2 番目として，BRENDA 等から取得した，EC 3.1.1, EC 3.7.1 および EC 3.5.3 の酵素反応データを入力し，これらを予測された EC 番号に分類する．この 2 つの方法で，入力データを正解の EC 番号に正しく分類できるかの分類精度を評価する (Fig. 1).

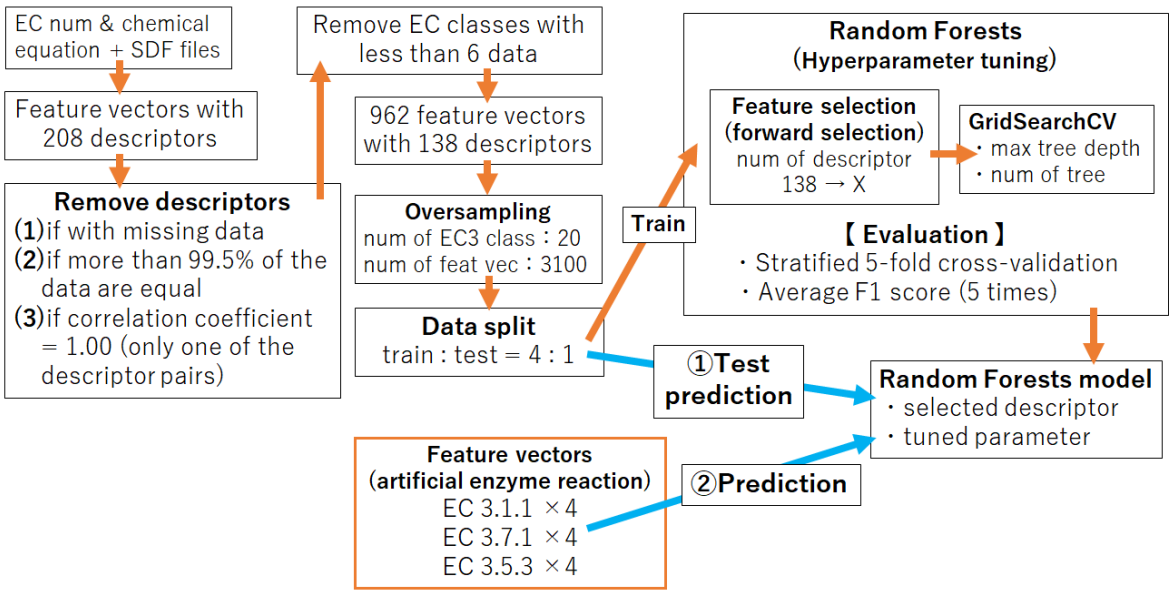

Fig. 1. The process of EC number prediction.

学習データ作成 まず，KEGG から EC 番号と反応式情報を取得した．反応式中の各化合物には KEGG の化合物 ID が割り当てられている．この ID は PubChem Substance ID に対応しているため，今回は PubChem [9]から化合物の構造情報を示した MOL ファイルを取得した．次に Python の RDKit ライブラリ [10] を用いて，MOL ファイル形式の化学構造を Simplified Molecular Input Line Entry System (SMILES) の文字形式に変換した．これにより，EC 番号と SMILES 形式で書かれた反応式を取得した (Table 1)．行ラベルは EC 番号，列ラベルはそれぞれ反応式の左辺第 1 項，第 2 項および右辺第 1 項、第 2 項を表す．この際，SMILES が空白や欠損文字，または「H+」や「HX」の表記となっている化合物は物性値を計算できないため，それらの表記を含む反応式は除去した．また，今回用いる EC 番号のクラスを EC3 (加水分解酵素)とし，反応式が 6 種類以上登録されているクラスを採用した．また，反応式の左辺および右辺にある化合物がそれぞれ 2 つずつのものに限定し，合計 962 種類の反応式を用いた．

Table 1.　EC number and chemical equation written in SMILES.

ENZYME	left1	left2	right1	right2
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
3.1.1.6	*OC(C)=O	[H]O[H]	*O	CC(=O)O
3.1.1.1	*OC(*)=O	[H]O[H]	*O	*C(=O)[O-]
3.1.1.7 3.1.1.8	CC(=O)OCC[N+](C)(C)C	[H]O[H]	C[N+](C)(C)CCO	CC(=O)O
3.1.1.8	*C(=O)OCC[N+](C)(C)C	[H]O[H]	C[N+](C)(C)CCO	*C(=O)[O-]
...	...	...	...	...

次に，SMILES を RDKit の構造式オブジェクトに変換し，各化合物に対して特性値を計算した． RDKit には特性値を計算する 208 種類の記述子と呼ばれるものが用意されており，1 つの特性値に対し，1 つの記述子が対応している．各化合物に対し，記述子を用いて特性値を計算し，基質の和と生成物の和の差を取った，特性値変化量を導出する．このとき，1 つの反応式は 208 種類の特性値変化量を持ち，208 次元のベクトルで表される．本研究では，この特徴ベクトルを，機械学習に入力する特徴量とする．今回学習に使用した反応式は全て，左辺が基質と H_2O，右辺が生成物 2 種類である加水分解反応としているため，右辺第 1 項と第 2 項の特性値の和と左辺第 1 項と第 2 項の和の差を取った，特性値変化量となる．これを全ての反応式で行い，各反応式が 208 次元の特性値変化量からなる特徴ベクトルを持つデータを作成した．一部抜粋した出力結果 (Table 2)において，行ラベルは EC 番号，列ラベルは各記述子の特性値変化量を表し，特定 EC 番号に属する反応式に対する 208 次元の特徴ベクトルデータとなっている．

Table 2. EC number and Feature vector for each chemical equation.

	MaxEStateIndex	MinEStateIndex	MinAbsEStateIndex	qed	MolWt	HeavyAtomMolWt	ExactMolWt
3.5.1.	1.415	-1.6875	0.8125	-0.053711	61.040001	58.015999	61.016376
3.6.1.	-7.82413	3.657444	-4.738124	-0.206899	79.978996	78.971001	79.966331
3.6.1.	-8.56906	4.22743	-4.681925	-0.097512	0.0	0.0	-0.0
3.6.1.	-8.56906	4.068902	-3.057568	-0.272087	0.0	0.0	-0.0
3.5.4.	0.017361	0.014793	0.036602	-0.063832	0.0	0.0	0.0
...	...	...	...	...	...	...	...
3.1.1.	-6.828515	-0.668523	-1.021632	-0.087708	0.0	0.0	0.0
3.2.1.	-8.720595	1.384253	-0.518534	-0.069449	0.0	0.0	-0.0
3.2.1.	-8.708548	1.381274	-0.495988	-0.069449	0.0	0.0	-0.0

次に，特徴ベクトルデータの整形を行った．反応式の特性値変化量に欠損値が含まれる 13 種の記述子を削除した．そして，各クラスの特性値変化量が 99.5%以上等しい値となる記述子を 55 種削除した．さらに相関係数が 1 となる記述子ペアが 2 種存在したため，それぞれ片方の記述子を除外した．結果，138 次元の特徴ベクトルが得られた．

また，本研究では反応式がどの EC 番号に属するかのクラス分類の精度を検証するため，使用する EC 番号のクラス 20 種と各クラスのデータ数の内訳を示した (Table 3(a))．EC 3 の反応式のうち，データ数が 6 以上の EC 番号のみを採用している．初めに，データ数が最小の 6 から最大の 155 と不均衡なデータ分布のため，Python の imbalanced-learn ライブラリにある SMOTE を用いてオーバーサンプリングを行った [11]．全てのクラスのデータ数を 155 になるように調整し，結果的に 3100 個のデータが得られた (Table 3(b))．最終的に，各クラスで学習データとテストデータの割合が 4:1 となるよう均等に分割し，合計 2480 個のデータを機械学習器に学習させた．

Table 3. The number of feature vectors of used EC number.
(a) original data and (b) data after using SMOTE.

(a) Before applying SMOTE

EC Class	Num of Equation	EC Class	Num of Equation	EC Class	Num of Equation
3.1.1	122	3.2.2	24	3.5.5	12
3.1.2	59	3.3.2	6	3.5.99	11
3.1.3	152	3.4.13	6	3.6.1	94
3.1.4	29	3.4.19	7	3.7.1	35
3.1.6	14	3.5.1	155	3.8.1	16
3.1.7	8	3.5.3	25	3.13.1	9
3.2.1	131	3.5.4	47	Total	962

(b) After applying SMOTE

EC Class	Num of Equation	EC Class	Num of Equation	EC Class	Num of Equation
3.1.1	155	3.2.2	155	3.5.5	155
3.1.2	155	3.3.2	155	3.5.99	155
3.1.3	155	3.4.13	155	3.6.1	155
3.1.4	155	3.4.19	155	3.7.1	155
3.1.6	155	3.5.1	155	3.8.1	155
3.1.7	155	3.5.3	155	3.13.1	155
3.2.1	155	3.5.4	155	Total	3100

Random Forests(RF)と記述子選択による予測モデルの作成・評価　機械学習器として RF を用い，記述子の閾値に応じて特徴ベクトルを分類するモデルを作成した．データ分類する際に用いる特徴量の順番と閾値についてはジニ不純度を用いて決定した．初めに，forward selection による特徴選択を行った．ライブラリとして Python mlxtend.feature_selection の SequentialFeatureSelector [12] 用い，RF 分類において学習データの予測精度が増加しなくなるまで，記述子を 1 つずつ追加した．次に，RF のパラメータ(決定木の最大深さ，決定木数)

に対して，グリッドサーチによるパラメータ調整を実行した．なお，特徴選択とパラメータ調整時の評価指標として F1 スコアを用い，層化 5 分割交差検証による平均値で評価した．その後，調整パラメータで RF モデルを作成し，テストデータを入力することで，分類精度の評価を行った．

文献反応に対する EC 番号予測精度の評価　3 つの EC クラスに対して，化合物合成[13]や BRENDA に記載されている文献から，基質を用いた酵素反応を取得し，各 4 個，計 12 個の反応式を取得した (Fig. 2, 3, 4)．これらの反応式は，どの EC 番号に属する酵素反応であるかが既に判明している．今回は，SMOTE 実行前のデータ数が多い EC 3.1.1，並びにデータ数が中規模の EC 3.7.1 および EC 3.5.3 の反応式をそれぞれ 4 個ずつ選択した．予測モデルに入力した際に正しい EC 番号に予測されるかを，予測確率を用いて評価した．EC 3.1.1 の 4 つの反応式データのうち，2 つは SciFinder^{n}[15]や PubChem で入手した MOL ファイルを RDKit で SMILE に変換し，特徴ベクトルを計算することで取得した (Fig. 2(a)(b))．残り 2 つのデータ，および EC 3.7.1, EC3.5.3 のデータは BRENDA から MOL ファイルを取得したのち，同様の方法で作成した．

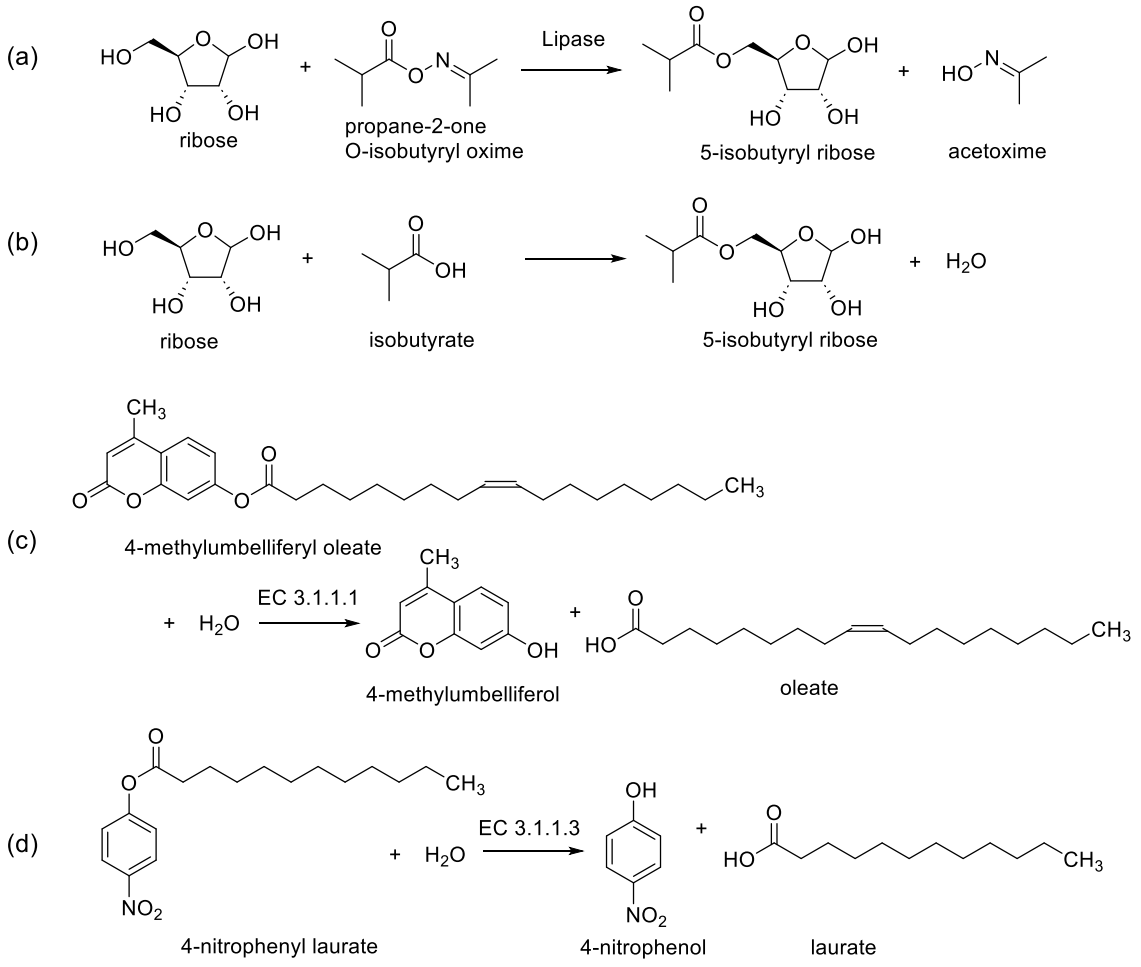

Fig. 2. Literature reactions (EC 3.1.1). (a) is the first reaction in the enzyme screening to synthesize mornupiravir [13]. Since the enzyme Novozym 435 (*Candida antarctica* lipase B) belonging to EC 3.1.1.3 [14] gave the highest yield, EC 3.1.1.3 was considered to be the correct class. Although, in the literature, acetoxime is not described in the chemical equation, it was added to the right side of the equation assuming that it is also produced as a product. In addition, tert-amyl alcohol was included in the left-hand side of the chemical equation because it was used in the synthesis of (a), and the amount of changes in characteristic value were calculated as the difference between the sum of terms 1 to 3 on the left-hand side and the sum of terms 1 to 2 on the right-hand side. (b) is the enzyme reaction we assumed to occur. In (a) and (b), the reverse reaction was input to be predicted. (c) and (d) are one of the literature reactions of EC 3.1.1.3 and EC3.1.1.1 (EC3.1.1.13) described in BRENDA, respectively.

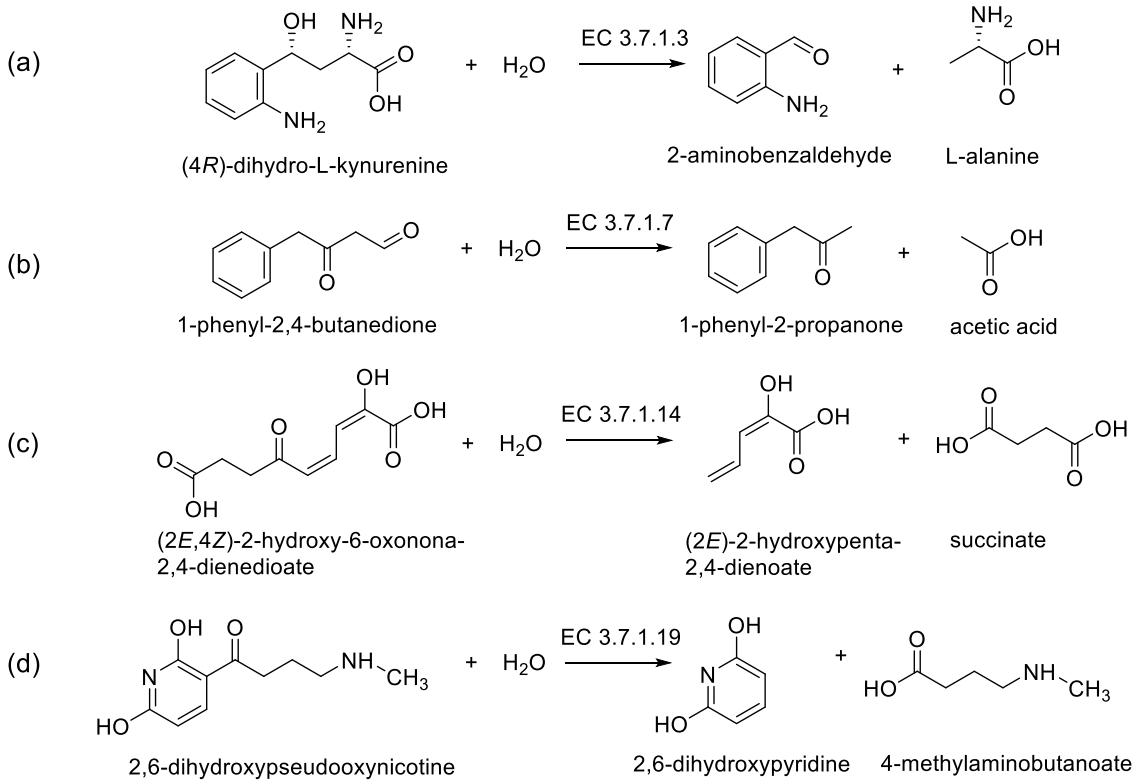

Fig. 3. Literature reactions (EC 3.7.1). (a), (b), (c) and (d) are one of the literature reactions of EC 3.7.1.3, EC 3.7.1.7, EC 3.7.1.14 and EC 3.7.1.19 listed in BRENDA, respectively.

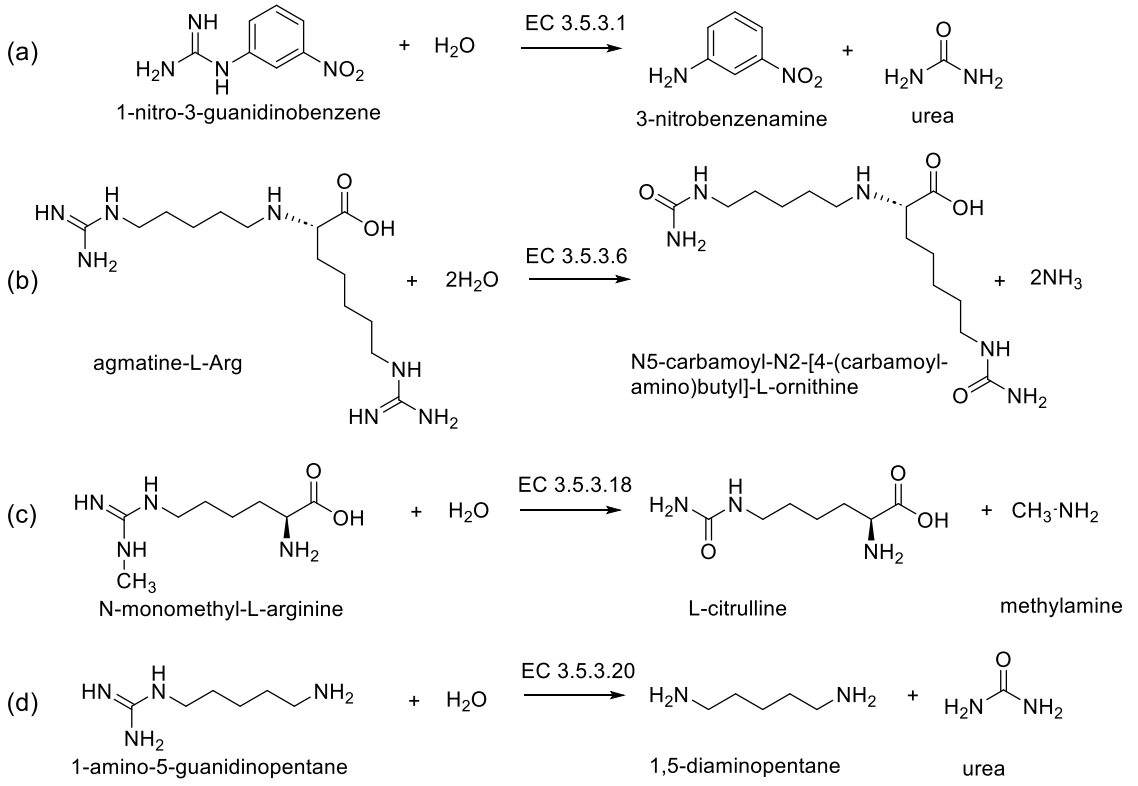

Fig. 4. Literature reactions (EC 3.5.3). (a), (b), (c) and (d) are one of the literature reactions of EC 3.5.3.1, EC 3.5.3.6, EC 3.5.3.18 and EC 3.5.3.20 listed in BRENDA respectively. The coefficients of the reaction (b) are not considered in the calculation of the amount of change in characteristic value.

実験結果

KEGG のテストデータに対するモデルの予測精度　RF を用いた記述子選択とパラメータ調整により，23 個の記述子が選択され，決定木の最大深さ 15，決定木数 800 となった．KEGG のテストデータをモデルに入力した際の，各クラスの適合率(precision)，再現率(recall)，F1 スコア，および各平均値を出力した (Table 4)．F1 スコアに関して，SMOTE 適用前 (Table 4(a)) はばらつきがあり，値が極端に低い EC 番号が見られたが，SMOTE 適用後 (Table 4(b)) には全体的に高い値となり，予測精度の改善が見られた．

Table 4. Prediction accuracy of created models for KEGG test data.

(a) Before applying SMOTE

EC Num	Num of Equation	Precision	Recall	F1 Score	EC Num	Num of Equation	Precision	Recall	F1 Score
3.1.1	25	0.96	0.96	0.96	3.4.19	1	1.00	1.00	1.00
3.1.2	12	0.92	1.00	0.96	3.5.1	31	0.94	0.97	0.95
3.1.3	31	0.91	1.00	0.96	3.5.3	5	0.83	1.00	0.91
3.1.4	6	0.86	1.00	0.92	3.5.4	9	0.89	0.89	0.89
3.1.6	3	1.00	1.00	1.00	3.5.5	2	1.00	1.00	1.00
3.1.7	2	0.00	0.00	0.00	3.5.99	2	1.00	0.50	0.67
3.2.1	26	0.96	0.96	0.96	3.6.1	19	0.86	0.95	0.90
3.2.2	5	0.83	1.00	0.91	3.7.1	7	1.00	0.71	0.83
3.3.2	1	1.00	1.00	1.00	3.8.1	3	1.00	0.67	0.80
3.4.13	1	0.00	0.00	0.00	3.13.1	2	1.00	0.50	0.67
					Total	193			
					Average		0.85	0.80	0.81
					Accuracy				0.92

(b) After applying SMOTE

EC Num	Num of Equation	Precision	Recall	F1 Score	EC Num	Num of Equation	Precision	Recall	F1 Score
3.1.1	31	1.00	0.94	0.97	3.4.19	31	1.00	1.00	1.00
3.1.2	31	1.00	1.00	1.00	3.5.1	31	1.00	0.97	0.98
3.1.3	31	0.97	1.00	0.98	3.5.3	31	0.97	0.97	0.97
3.1.4	31	1.00	0.97	0.98	3.5.4	31	1.00	0.97	0.98
3.1.6	31	1.00	1.00	1.00	3.5.5	31	0.89	1.00	0.94
3.1.7	31	1.00	1.00	1.00	3.5.99	31	1.00	1.00	1.00
3.2.1	31	1.00	1.00	1.00	3.6.1	31	1.00	0.97	0.98
3.2.2	31	0.97	1.00	0.98	3.7.1	31	1.00	1.00	1.00
3.3.2	31	1.00	1.00	1.00	3.8.1	31	1.00	1.00	1.00
3.4.13	31	1.00	1.00	1.00	3.13.1	31	1.00	1.00	1.00
					Total	620			
					Average		0.99	0.99	0.99
					Accuracy				0.99

文献の反応データに対するモデルの予測精度　BRENDA と文献から取得した酵素反応がどの EC 番号のクラスに属する(予測される)のか，予測確率の高いものから順に 1st, 2nd, 3rd で示した (Table 5)．EC 3.5.3 の反応式はいずれも正解となる EC 番号が 1 番目に予測され，EC 3.1.1 の反応式で 2 つ，EC 3.7.1 の反応式では 1 つだけ，異なる EC 番号が 1 番目に予測された．また，正解

EC 番号が 1 番目に予測された Target について，その予測確率(Probability)に偏りが見られた．

Table 5.　EC number prediction probability for literature reactions (Target).

(a) Target Equation (registered as EC 3.1.1)

	1st	2nd	3rd
Target1	3.2.1.	3.1.1.	3.8.1.
Probability	0.349167	0.095	0.08375
Target2	3.2.1.	3.1.1.	3.7.1.
Probability	0.230313	0.179375	0.1525
Target3	3.1.1.	3.7.1.	3.2.1.
Probability	0.561849	0.094382	0.067454
Target4	3.1.1.	3.7.1.	3.5.1.
Probability	0.922124	0.052175	0.007679

(b) Target Equation (registered as EC 3.7.1)

	1st	2nd	3rd
Target1	3.13.1.	3.7.1.	3.5.99.
Probability	0.251406	0.127435	0.124286
Target2	3.7.1.	3.1.2.	3.5.1.
Probability	0.293299	0.2725	0.119097
Target3	3.7.1.	3.1.1.	3.1.2.
Probability	0.992454	0.002546	0.0025
Target4	3.7.1.	3.1.2.	3.5.1.
Probability	0.991954	0.0025	0.001625

(c) Target Equation (registered as EC 3.5.3)

	1st	2nd	3rd
Target1	3.5.3.	3.13.1.	3.7.1.
Probability	0.952604	0.009063	0.007774
Target2	3.5.3.	3.5.99.	3.5.4.
Probability	0.7375	0.07625	0.07
Target3	3.5.3.	3.5.4.	3.5.99.
Probability	0.90625	0.05	0.01625
Target4	3.5.3.	3.5.99.	3.1.1.
Probability	0.99875	0.00125	0

考察

BRENDA や文献に記載されている酵素反応において，9 種の Target で正解 EC 番号が 1 番目に予測され，3 種の Target で 2 番目に予測される結果となった．提案した EC 番号予測モデルで，従来用いられてきた KEGG に加え，新たに BRENDA や文献の酵素反応に対し，20 種用いた EC 番号の中から正解 EC 番号が少なくとも 2 番目以上に予測された点で，現状十分な予測精度が得られた．一方で，正解 EC 番号を 2 番目に予測した EC 3.1.1 の Target 2 種と EC 3.7.1 の Target 1 種に関して，1 番目に予測された EC 番号と正解 EC 番号の間に酵素の性質や反応の構造に共通点があると考えられる．特徴量選択手法や機械学習手法の検討，EC 番号内の酵素反応データの分析などを行い，予測モデルに反映することが今後の課題となる．また，予測確率に偏りが見られる点についても，偏りを減らし，全ての Target で正解 EC 番号をより高確率で予測できるモデル作成が必要である．選択された記述子 23 種について，それぞれが表す物理・化学的指標の類似性を調べるとともに，各記述子の重要度を算出し，それに応じて重みづけを行い，モデルに再学習する方法も予測精度を高める策として有効と考えられる．また今回は，化学反応式中の化合物にかかる係数の反映や，基質や生成物が 3 種類以上の反応式を用いた学習を行っていない．モデル作成に用いた基質 2 種類および生成物 2 種類の反応式と同時に用い

ると，1 つの EC 番号に属する反応式の特性値変化量のばらつきが大きくなり，予測精度に影響を及ぼす可能性があるためである．一方で，これらを用いることで，汎化性能の向上も期待できるため，ばらつきを少なくできる方法を検討することが重要である．

本研究では，forward selection と SMOTE を用いて，EC 3class の subclass（2 桁目）および sub-subclass (3 桁目）の酵素反応を予測する RF モデルを作成した．そして，KEGG に加えて，新たに BRENDA やその他文献の酵素反応データに対して EC 番号予測を行い，現状十分な予測精度が得られた．今後の展望として，EC 番号の class (1 桁目)および sub-sub-subclass (4 桁目）に対しても予測できるモデルを作成することが挙げられる．class (1 桁目）に関しては，EC 3 の加水分解酵素 EC 1 (酸化還元酵素)や EC 4(転移酵素)など，より汎用的に予測ができるモデルの作成が必要となる．また，EC 番号の sub-sub-subclass (4 桁目）は酵素の性質や作用する結合によって class (1 桁目)から sub-subclass (3 桁目）まで振り分けた酵素を番号で区別しているのみである．したがって，特徴が類似している反応が多くなるため，予測が難しい．これらに共通して必要なのは，適切に予測していくための，データ加工法の修正やより良い記述子選択・機械学習手法の検討，RDKit 以外の新たな記述子の追加などが考えられる．そして，BRENDA や文献から取得可能な酵素反応のデータ数を増やし，今回行った新たな予測の範囲の拡大と精度の信頼性向上を目指すことが必要である．また，将来的には新規の化学反応に対して，最適な酵素候補を提示できるモデルを開発していくことが求められる．

謝辞

本研究の実施にあたり，ケモインフォマティクスの分野に関して多くのご助言をいただいた荒木通啓先生（国立研究開発法人医薬基盤・健康・栄養研究所 AI 健康・医薬研究センター 副センター長），渡邉直暉様（国立研究開発法人医薬基盤・健康・栄養研究所）に深く御礼申し上げます．

本研究は、くすりのシリコンバレーTOYAMA 創造コンソーシアムの助成および科学研究費(A) 22H00361（浅野泰久）により実施されました．この場を借りて御礼申し上げます．

文献

1) Enzyme Nomenclature: https://iubmb.qmul.ac.uk/enzyme/ (2023/3/12).
2) Latino, D. A. R. S., and Aires-de-Sousa, J.: *Journal of Chemical Information and Modeling*., 49, 1839-1846 (2009).
3) Hu, Q-N., Zhu, H., Li, X., Zhang, M., Deng, Z., Yang, X., and Deng, Z.: *PLoS ONE*., 7, e52901 (2012).
4) Ryu, J. Y., Kim, H. U., and Lee, S. Y.: *Proceedings of the National Academy of Sciences*., 116, 13996-14001 (2019).
5) KEGG: Kyoto Encyclopedia of Genes and Genomes: https://www.genome.jp/kegg/kegg_ja.html (2023/3/12).
6) BRENDA Enzyme Database: https://www.brenda-enzymes.org/index.php (2023/3/1).
7) 荒木　通啓：生物工学会誌, 92, 304 (2014).
8) Breiman, L.: *Machine Learning*., 45, 5-32 (2001).
9) PubChem: https://pubchem.ncbi.nlm.nih.gov/ (2023/3/12).
10) The RDKit Documentation: https://www.rdkit.org/docs/index.html (2023/3/1).
11) Chawla, N. V., Bowyer, K. W., Hall, L. O., Kegelmeyer, W. P.: *Journal of Artificial Intelligence Research*., 16, 321-357 (2002).
12) Mlxtend.feature selection: http://rasbt.github.io/mlxtend/api_subpackages/mlxtend.feature_selection/ (2023/3/12).
13) Benkovics, T., McIntosh, J. A., Silverman, S. M., Kong, J., Maligres, P., Itoh, T., Yang, H., Huffman, M. A., Verma, D., Pan, W., Ho, H-I., Vroom, J., Knight, A., Hurtak, J., Morris, W., Strotman, N. A., Murphy, G., Maloney, K. M., and Fier, P. S.: *ChemRxiv*., (2020).
14) Ravelo, M., Gallardo, M. E., Ladero, M., Garcia-Ochoa, F.: *Catalysts*., 12, 1531 (2022).
15) CAS SciFinder^{n}: https://scifinder-n.cas.org/ (2023/3/26).

Development of EC number prediction model using machine learning to predict the optimal enzyme for a chemical reaction

Katsuya Mutoh[1], Genji Iwasaki[2], Yasuhisa Asano[2]*, and Koji Okuhara[3]*
([1]Department of Electrical and Computer Engineering, Graduate School of Engineering, [2]Department of Biotechnology Faculty of Engineering, [3]Department of Information Systems Engineering Faculty of Engineering, Toyama Prefectural University, 5180 Kurokawa, Imizu, Toyama, 939-0398)

Abstract

The four-digit EC number contains the enzyme names and the chemical equations which the enzyme acts on. In this study, we created a model that predicts an EC number of optimal enzyme candidates for a chemical reaction. After that, the model was evaluated whether if it predicts the correct EC number, using the enzyme reaction data listed in the Kyoto Encyclopedia of Genes and Genomes (KEGG), BRENDA and other literature. We developed a Random Forests (RF) prediction model to predict the subclass and sub-subclass (second and third digits) of the chemical equation belonging to EC 3, which consists of two substrates and two products. First, character data of EC number and chemical equation were obtained from KEGG and converted into values. For quantification, the amounts of changes in 208 descriptors (physical and chemical property values) when substrates change into products were calculated for each chemical equation and 208-dimensional feature vectors of the chemical equations were created. Next, SMOTE was applied to oversample 962 feature vectors to 3100 vectors. Then, descriptor selection was performed for model creation. Forward selection was applied to the RF and 23 descriptors were selected. Parameter tuning resulted in a maximum decision tree depth of 15 and the number of trees of 800. The predictive results of the created model yielded an average F1 score of 0.99 for the test data of KEGG. In addition, the prediction accuracy was currently enough for 12 literature reactions listed in BRENDA.

Keywords: EC number, Organic Synthesis, Machine Learning, Feature Selection