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

Katsuya Mutoh[1], Genji Iwasaki[3], Koji Okuhara[2] and Yasuhisa Asano[3]

[1]Department of Electrical and Computer Engineering, [2]Department of Information Systems Engineering,

[3]Biotechnology Research Center and Department of Biotechnology, Toyama Prefectural University, 5180 Kurokawa, Imizu, Toyama 939-0398, Japan

Abstract

The outbreak of COVID-19 has increased the demand for new drug development. That has led to a growing interest in chemoinformatics, which is valuable information technology to predict chemical reactions. The use of enzymes as catalysts is gaining importance in terms of the environment and reaction efficiency. In order to predict the best enzyme to obtain the desired product, the target chemical equation is compared with typical chemical equations of enzymes classified by Enzyme Commission Number (EC number) using clustering. The EC number of the chemical equation that is evaluated to have the highest similarity is predicted.

1. Introduction

In recent years, enzymes are increasingly used as biocatalysts in the design and prediction of chemical reactions for green chemistry and efficiency. Therefore, it is becoming important to predict the most suitable enzyme for a chemical reaction by machine learning. The first step of this study, we make the machine learner learn the changes in physical and chemical property values from reactants to products in the chemical equations described in Enzyme Commission numbers (EC numbers). Next, the best enzymes to use for a target chemical equation is predicted as an EC number by learned machine learner. Finally, we consider the prediction accuracy of the learner.

Keywords: EC Number, Chemoinformatics, Feature Selection, Clustering

2. Chemoinformatics and Information Technology

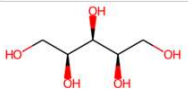
Structural Representations of Compounds

Various representations are used to handle chemical structures in computer as chemoinformatics. This study uses a method quantifying physical/chemical characteristics.

Using RDKit [1], we calculate many type of characteristic values (called descriptors) of compounds for machine learning.

RDKit reads chemical structure information files obtained from databases and draws structural formula object. After that, Smiles or Characteristic values are calculated.

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



Structural formula object

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

Text representation (Smiles)

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

```
from rdkit.Chem import Descriptors
Descriptors.MolWt(Xylit
152.14600000000002
```

Calculate characteristics value (molecular weight)

3. Proposed Method

3.1 Method Outline

This study compares the structural changes of a target chemical equation and EC number chemical equations. Finally, the optimal EC number is predicted for the target. we assume that, if the target chemical structural change from reactant to product is similar to an EC chemical equation structural change, the same EC enzymes can be used to target chemical reaction. This is based on the concept of molecular similarity used in chemistry [2].

Therefore, when an EC chemical equation structural change is similar for target, this EC number is predicted as the optimal enzyme candidates.

Structural change = Amount of changes in characteristic values from reactants to products

Assumptions : Structural changes of the target and EC chemical equation are similar.
→ Target product can be obtained by using the EC enzymes(concept of molecular similarity).

Retrosynthesis analysis identified reactants

Target: Reactant A + Reactant B → Target Product

Characteristic value changes (descriptor): Mol Weight **X**, Charge **Y**, Hydrophobicity **Z**, ⋯ } Feature Vector $V1 = (X, Y, Z, \cdots)$

EC 3.X.X.X ⋯ Typical EC Number Chemical Equations: Reactant 1 + Reactant 2 → Product 1 $V2 = (X', Y', Z')$

Proposed Part

if (V1≒V2)

Predicted EC 3.X.X.X

This number enzymes are candidates

Experiments after this study

Select one of the best from EC 3.X.X.X enzymes

3.2 Features Representing Structural Change

Previous studies [3]: Classification of EC chemical equation (focusing on fingerprint changes)

$$\text{EC X.X.X.X}: \quad RCT1 + RCT2 \rightarrow PDT1 + PDT2$$
$$\rightarrow RFP = FP_{PDT1+PDT2} - FP_{RCT1+RCT2}$$

FP : Fingerprints of each compound
RFP : Reaction difference fingerprints

One kind of fingerprint have a limit in representing the structural changes.(Many fingerprints have been developed)

Proposed method : Amount of change in characteristic values

Characteristic values of n descriptors for each reaction equation: cv_j

$$cv_j = (PD_1 + PD_2) - (RT_1 + RT_2) \qquad (j = 1, 2, ..., n)$$

(RT_i (PD_i): Characteristic value of reactant i (product j))

Feature vectors for each chemical equation: $\boldsymbol{DF}_i (i = 1, 2, ..., m)$

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

Present the EC number of the $\boldsymbol{DF}_i$ most similar to the Target

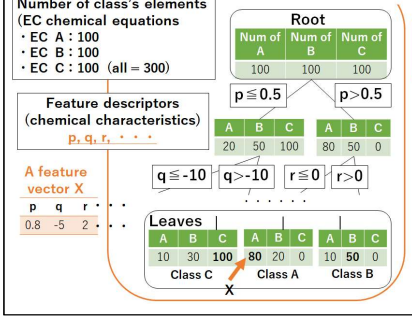
	Desc 1	Desc 2	⋯	Desc n
Target	cv_{T1}	cv_{T2}	⋯	cv_{Tn}
DF_1	cv_{11}	cv_{12}	⋯	cv_{1n}
DF_2	cv_{21}	cv_{22}	⋯	cv_{2n}
⋮	⋮	⋮	⋱	⋮
DF_m	cv_{m1}	cv_{m2}	⋯	cv_{mn}

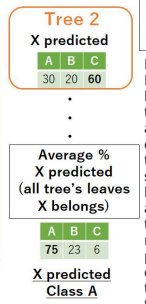
Many type of physical/chemical property value changes describe structural changes

(RDKit has 208 descriptors)

Some studies mainly use reactions that occur in nature as described in KEGG [4]. This study is significant in terms of making predictions for experimental reactions that are not described.

3.3 Predicting Method (Random Forests [5])





Tuning Parameters
- **Num of Descriptors**
- **Tree depth (root → leaf)**
- **Num of Tree**

Random Forests uses many trees. Each EC equation's feature vector go to a leaf from root classified by descriptor thresholds. Where a descriptor is used and how its threshold is determined, depends on the information gain about the bias of the class distribution when splitting data from one node to two lower nodes. The descriptor location and threshold are determined so that the information gain of each node is maximized. The class prediction probability of the feature vector is obtained by the average class ratios of the leaf nodes that it belongs to in each tree.

4. Results and Discussions

4.1 Process of Proposed Method and Preliminary Results

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

20 classes
962 equations

Data Splitting (train : test=4 : 1 in 1 class)

Train data (769) with 208 descriptors

Remove non-calculable descriptors

Random Forests Learning (Parameter Tuning)

【Parameter】
- Num of descriptors → selected 27
- Tree depth → 15
- Number of Trees → 100

【Evaluation indicator】
- Stratified 5-Fold Cross-validation
- F1 average score

Test data (193) with reduced descriptor

①Test Prediction

Created classification model (Tuned Parameter & reduced descriptor)

②Prediction

Target equation
EC 3.1.1 ×4
EC 3.5.3 ×4
EC 3.7.1 ×4

→ : Learning
→ : Prediction

Only selected descriptors for highest classification accuracy on training data

4.2. Test data Prediction Result

	precision	recall	f1-score	num of equations
3.1.1.	0.96	0.96	0.96	25
3.1.2.	0.92	1.00	0.96	12
3.1.3.	0.91	0.94	0.92	31
3.1.4.	0.86	1.00	0.92	6
3.1.6.	1.00	1.00	1.00	3
3.1.7.	0.00	0.00	0.00	2
3.13.1.	1.00	0.50	0.67	2
3.2.1.	0.96	0.96	0.96	26
3.2.2.	0.83	1.00	0.91	5
3.3.2.	1.00	1.00	1.00	1
3.4.13.	0.00	0.00	0.00	1
3.4.19.	1.00	1.00	1.00	1
3.5.1.	0.94	0.97	0.95	31
3.5.3.	0.83	1.00	0.91	5
3.5.4.	0.89	0.89	0.89	9
3.5.5.	1.00	1.00	1.00	2
3.5.99.	1.00	0.50	0.67	2
3.6.1.	0.86	0.95	0.90	19
3.7.1.	1.00	0.71	0.83	7
3.8.1.	1.00	0.67	0.80	3
accuracy			0.92	193
average	0.85	0.80	0.81	193

While the classes having many test data were classified with high accuracy, some of the classes with fewer data were all predicted as different classes.

4.3. Target Prediction Result

EC 3.1.1 target pred

	1st	2nd	3rd
target1	3.1.1.	3.2.1.	3.7.1.
Pred %	0.59	0.13	0.08
target2	3.2.1.	3.3.2.	3.2.2.
Pred %	0.45	0.15	0.14
target3	3.1.1.	3.5.1.	3.7.1.
Pred %	0.62	0.16	0.07
target4	3.1.1.	3.7.1.	3.1.2.
Pred %	0.97	0.02	0.01

EC 3.7.1 target prediction

	1st	2nd	3rd
target1	3.7.1.	3.1.2.	3.5.1.
Pred %	0.31	0.17	0.152
target2	3.2.2.	3.13.1.	3.5.4.
Pred %	0.3495	0.210526	0.060526
target3	3.1.1.	3.7.1.	3.6.1.
Pred %	0.29	0.22	0.11
target4	3.1.1.	3.5.1.	3.7.1.
Pred %	0.28	0.24	0.13

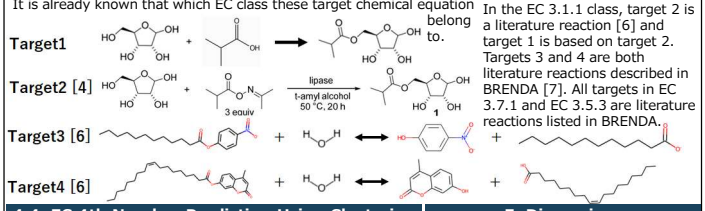
EC 3.1.1 targets except target 2 were predicted EC 3.1.1 class with the highest probability. EC 3.5.3 targets were predicted the same class with better probability than EC 3.1.1. However, in the three EC 3.7.1 target, different EC numbers had the highest predicted probabilities.

EC 3.5.3 target prediction

	1st	2nd	3rd
target1	3.5.3.	3.1.6.	3.5.99.
Pred %	0.96	0.02	0.01
target2	3.5.3.	3.5.4.	3.5.99.
Pred %	0.95	0.03	0.02
target3	3.5.3.	3.5.4.	3.5.99.
Pred %	0.89	0.06	0.03
target4	3.5.3.	3.2.1.	3.1.1.
Pred %	0.99	0.01	0.0

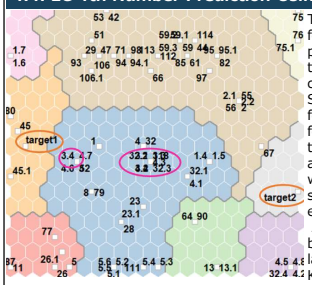
Target equation labeled EC 3.1.1 class

It is already known that which EC class these target chemical equation belong to.



In the EC 3.1.1 class, target 2 is a literature reaction [6] and target 1 is based on target 2. Targets 3 and 4 are both literature reactions described in BRENDA [7]. All targets in EC 3.7.1 and EC 3.5.3 are literature reactions listed in BRENDA.

4.4. EC 4th Number Prediction Using Clustering



The EC 3.1.1's forth number was predicted for testing using the clustering method SOM. It compares feature vectors a few hundred thousand times and places those with high similarity near each other.

A SOM program based on R language file of KH coder [8] was used.

For 2 targets of EC 3.1.1.3, five EC 3.1.1.3 equations belonged to other cluster.

While the EC first to three numbers classify enzymes according to enzyme properties, the fourth number classifies by their name. We will develop a method that can successfully predict up to the fourth number.

5. Discussions

In the third number of EC 3 class prediction, there were some classes with high accuracy and others with low accuracy. It is necessary to develop a feature selection method to obtain a certain level of prediction accuracy for all classes or add new features. In addition, this study did not take into account coefficients of each term in the chemical equation. The ratio of compounds need to be incorporate in the future.

6. Conclusion

This study proposed a method to predict the second and third number of the EC 3 class target chemical equation using random forests.

Future work focuses on developing a prediction method of other EC classes (EC 1,2,4,5,6 and 7) and a method based on a classification rules that introduces a clustering method for EC numbers.

References

[1] "The RDKit Documentation", https://www.rdkit.org/docs/GettingStartedInPython.html#list-of-available-descriptors
[2] Johnson, M.A., Maggiora, G.M. (1990). *Wiley.*
[3] Hu Q-N, Zhu H, Li X, Zhang M, Deng Z, Yang X, et al, (2012). *PLoS ONE.* 7, 12. doi:10.1371/journal.pone.0052901.
[4] "KEGG: Kyoto Encyclopedia of Genes and Genomes", https://www.genome.jp/kegg/kegg_ja.html
[5] Breiman, L. (2001). *Machine Learning.* 45, 5-32. doi:10.1023/A:1010933404324.
[6] Benkovics, T., McIntosh, J.A., Silverman, S.M., et al, (2020). *ChemRxiv. doi:10.26434/chemrxiv.13472373.v1*
[7] "BRENDA The Comprehensive Enzyme Information System", https://www.brenda-enzymes.org/index.php
[8] "KH Coder", http://khcoder.net/en/