

Chemoinformatics Using Dimensionality Reduction and Clustering for Enzyme Commission Number Prediction in Organic Synthesis

Katsuya Mutoh¹, Genji Iwasaki²,
Koji Okuhara², Yasuhisa Asano²

¹Graduate School of Engineering, ²Faculty of Engineering
Toyama Prefectural University, Japan

ICICIC 2022 Online, 11:40-13:40 Friday, September 16, 2022.

- 0. Contents
- 1. Introduction
- 2. EC Numbers
- 3. Chemoinformatics and Information Technology
- 4. Proposed Method
 - 4.1.1.
 - 4.1.2.
 - 4.2.
 - 4.3.
- 5. Experimental Results and Discussion
 - 5.1.
 - 5.2.1.
 - 5.2.2.
 - 5.2.3.
 - 5.2.4.
- 6. Conclusion

0. Contents

1. Introduction

2. EC Numbers

3. Chemoinformatics and Information Technology

4. Proposed Method

4.1.1.

4.1.2.

4.2.

4.3.

5. Experimental Results and Discussion

5.1.

5.2.1.

5.2.2.

5.2.3.

5.2.4.

6. Conclusion

- 1 : Introduction
- 2 : EC numbers
- 3 : Structural Representation of Compounds
- 4 : Proposed Method
- 5 : Experimental Results and Discussion
- 6 : Concluding Remarks

1. Introduction

3/15

The global outbreak of COVID-19 increased the need for new drugs development. It is becoming important factors to use enzymes in organic synthesis for producing desired products more efficiently.

Use of enzymes

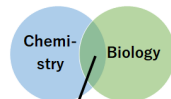
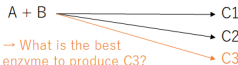
Enzymes (biocatalysts)

- React efficiently
- Eco-friendly

Increasingly
used



• Reaction prediction considering enzymes



Difficult to predict only with knowledge of organic synthesis.

Process of Enzyme selection

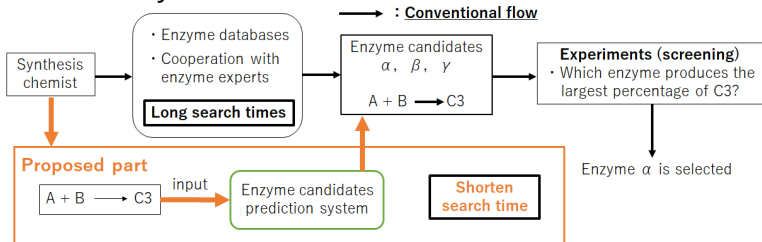
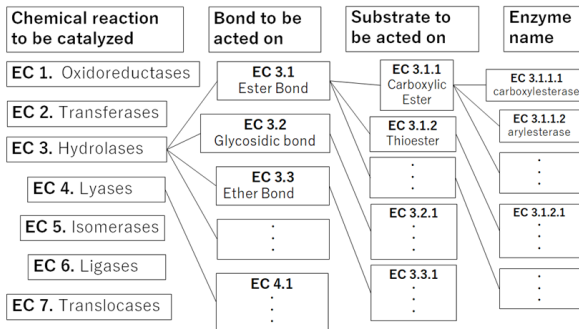


Fig. 1: Overview of enzyme use

2. EC Numbers (Enzyme Commission Numbers)

4/15

EC number consists of four pairs of numbers, EC X.X.X.X. Enzymes are classified by these pairs of numbers according to their properties.



Typical chemical equation using
EC 3.1.1.1 (**observed in nature**)

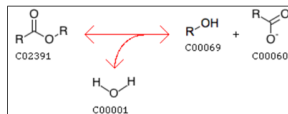


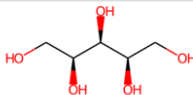
Fig. 2: EC number classification

3. Structural Representation of Compounds

5/15

Various representations are used to handle chemical structures in computer.

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



Structural
formula object

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

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

Simplified molecular
input line entry system
(Smiles)

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

The value of chemical
characteristics
(molecular weight)

```
152.14600000000002
```

Fig. 3: Structural representation by RDKit¹

¹ "The RDKit Documentation", <https://www.rdkit.org/docs/GettingStartedInPython.html#list-of-available-descriptors>

4.1 Proposed Method: Prediction of EC numbers using changes in characteristic values

6/15

When comparing the structural changes of the target chemical equation and the EC chemical equation, the EC number of the most similar EC chemical equation is predicted as the optimal enzyme candidates.

N types of changes in characteristic values when reactants are converted to products = **Indicator of structural change of compound**

Assumptions: The structural change of the target and EC chemical equation is similar.
→ There is a possibility that the target product can be obtained by using the enzyme with that EC number (concept of molecular similarity).

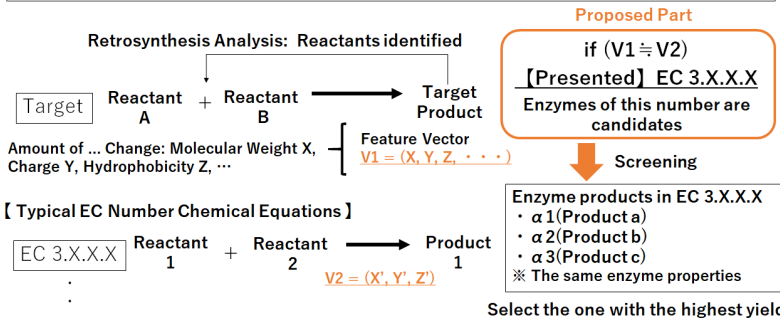


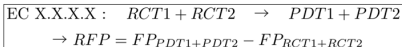
Fig. 4: The image of prediction flow

4.1 Proposed Method: Features Representing Structural Change

7/15

Compare similarity by the amount of change in characteristic values when changing from reactants to products.

Previous studies² : Classification of EC chemical equation (focusing on fingerprint changes)



FP : Fingerprints of each compound

RFP : Reaction difference fingerprints

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

→ Describe them **using many physical/chemical property values changes**
(RDKit: 208 descriptors to output each property value)

	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}
\vdots	\vdots	\vdots	\ddots	\vdots
DF_m	cv_{m1}	cv_{m2}	...	cv_{mn}

Proposed method : Amount of change in characteristic values

Characteristic values of n descriptors for each reaction equation: cv_j

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

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

Feature vectors for each chemical equation: $DF_i (i = 1, 2, \dots, m)$

$$DF_i = (cv_{i1}, cv_{i2}, \dots, cv_{ij}, \dots, cv_{in})$$

Present the EC number of the DF_i most similar to the Target

Tbl. 1: Feature vectors



4.2 Proposed Method: Dimensionality Reduction Using Clustering

8/15

The Creation of Composite Descriptors

Combine highly correlated descriptors by complete linkage method.

● Clustering is performed as long as descriptor pairs with a correlation coefficient of 0.9 or higher exist ↓

- ① Sequentially merge pairs with the highest correlation coefficient.
- ② Merge clusters based on the complete linkage method.
- ③ Standardize and average the characteristic values of descriptors in a cluster.

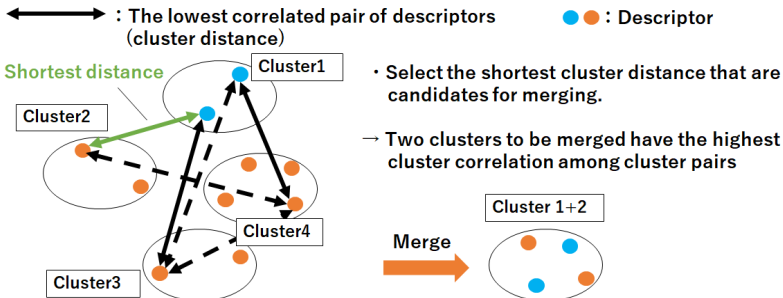
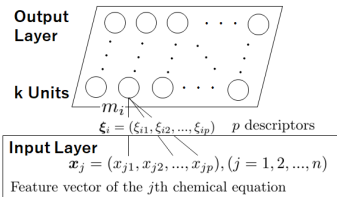


Fig. 5: The complete linkage method

4.3 Proposed Method: Clustering of chemical equations using SOM

9/15

Using self-organizing map, feature vectors after dimensionality reduction are mapped on a two-dimensional plane and clustered.



Unit: $m_i (i = 1, 2, \dots, k)$

Random initialization

Center of gravity of unit i : $r_i = (r_{i1}, r_{i2})$

Weight vector of unit i : $\xi_i = (\xi_{i1}, \xi_{i2}, \dots, \xi_{ip}), (i = 1, 2, \dots, k)$

Euclidean distance between x_j and ξ_i : $\|x_j - \xi_i\|$

Step(1): Find the winner unit m_c such that $\|x_j - \xi_i\|$ is minimized.

Step(2): Update weights of m_c to be closer to x_j

$$\xi_i = \xi_i + \alpha(t)(x_j - \xi_i)$$

$\rightarrow \alpha(t)$: Learning rate coefficient

(decreases with the number of learning times t)

Step(3): Update m_i in the neighborhood of m_c .

(so that m_i is also somewhat closer to x_j)

$$\xi_i = \xi_i + h(t)(x_j - \xi_i)$$

$\rightarrow h(t)$: Neighborhood function

(The further away from the winner unit, the weaker its influence.)

$$h(t) = \alpha(t) \exp \left[\frac{-\|r_c - r_i\|}{2\sigma^2(t)} \right] \quad (\sigma^2(t): \text{Adjustment function})$$

Step(4): Repeat Step(1) ~ (3) from x_{j+1} to x_n .

Step(5): Repeat (1) ~ (4) until the threshold learning count is reached.

\rightarrow Mapping the feature vectors of the chemical equation on each finalized winner unit.

Step(6): Color-coded clustering of each unit

- Merge criteria between units: Euclidean distance between ξ_i
- Merge method between clusters: Ward's method

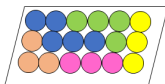


Fig. 6: SOM clustering

5.1 Numerical Experiment

10/15

Targets : Chemical reaction using EC 3.1.1.3 enzymes in the experiment

- ① SOM is applied to each target and EC 3.1.1 class feature vector.
- ② Predict the EC number of the EC chemical equation closest to the target as the best enzyme.

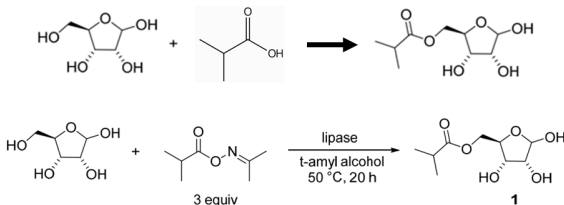
**Target1
Feature Vector**

(EC 3.1.1.3 labeled)

**Target2³
Feature Vector**

VS

**EC 3.1.1.X
Feature Vector**



**Typical EC 3.1.1.X
chemical equation**

× 112



X=? (from 65 types)

(Is X=3 predicted?)

Fig. 7: Comparison of each target and EC 3.1.1.X feature vector

5.2 Experimental Results (Dimensionality Reduction)

11/15

- Target1 and EC 3.1.1.X : 16 composite descriptors (clusters) and 73 dimensional feature vectors were created.
- Target2 and EC 3.1.1.X : 15 composite descriptors (clusters) and 76 dimensional feature vectors.

	cluster0	cluster1	cluster2	cluster3	cluster4	cluster5	cluster6	cluster7	cluster8	cluster9	...	Kappa3	EState_VSA8
target1	-0.811847	0.083828	4.450476	0.487355	0.249143	-0.157514	0.192650	-0.269549	0.670170	-0.277871	...	-105.319683	0.0
33	-0.811847	0.083828	-0.101680	0.194684	0.249143	-0.157514	0.210124	-0.042677	0.616525	-0.277871	...	-103.916364	0.0
6	-0.811847	0.083828	-0.101680	0.058794	0.249143	-0.157514	0.142875	-0.042677	-0.312148	1.964942	...	-1.005737	5.106527
1	1.141110	0.083828	-0.101680	0.061725	0.249143	1.630786	0.195868	-0.122933	-0.448792	1.964942	...	-1.278912	5.106527
7_8	-0.811847	0.083828	-0.101680	0.192381	0.249143	-0.157514	0.209894	-0.042677	0.643709	-0.277871	...	-103.686009	0.0
...
111	1.141110	0.083828	-0.101680	0.137586	0.249143	-0.157514	0.205520	-0.042677	-0.922301	-0.277871	...	-105.038909	4.736863
118	1.141110	0.083828	-0.101680	0.135367	0.249143	-0.157514	0.214994	-0.042677	-1.328664	-0.277871	...	-105.080776	0.0
2	-0.811847	0.083828	-0.101680	0.205582	-3.984159	-0.157514	0.206484	-0.042677	0.808984	1.964942	...	-104.229693	0.0
5.1	1.141110	0.083828	-0.101680	0.163351	0.249143	1.630786	0.218918	-0.122933	-0.693718	-0.277871	...	-106.172572	-4.523747
28.1	1.141110	0.083828	-0.101680	0.586779	0.249143	-1.945814	0.397498	0.852081	-0.713532	-2.520683	...	-104.746194	0.0

113 rows × 73 columns

- Period:** distinguish the chemical equations which have the same EC number.
- Underscore:** Identify the chemical equations registered in multiple EC numbers.
- E:** EC numbers other than EC 3.1.1.X.

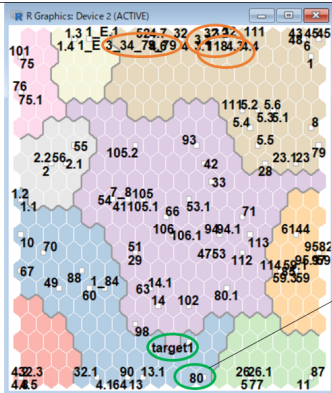
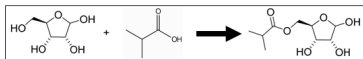
Fig. 8: The result of dimensionality reduction (Target1 and EC 3.1.1.X)

5.2 Experimental Results (SOM Clustering)

12/15

A SOM program based on R language file of KH coder⁴ was used.

target 1



- Number of units: 400 (20 x 20)
- Number of training
 - (1) Rough ordering phase: 1,000
 - (2) Convergence phase: 200,000

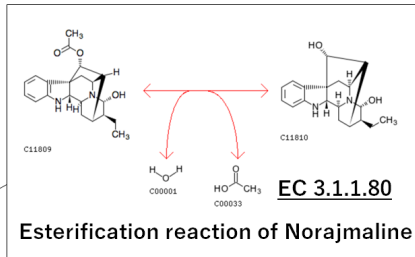


Fig. 9: Clustering results using SOM

⁴ "KH Coder", <http://khcoder.net/en/>

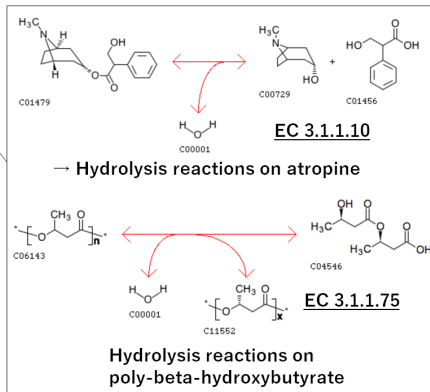
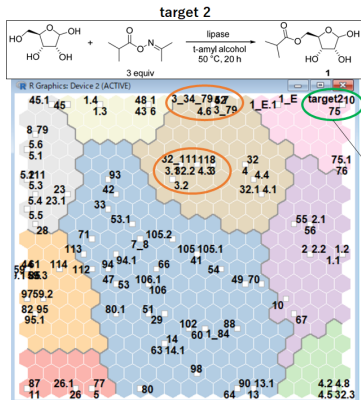


Fig. 10: EC 3.1.1.X chemical equations located near the target

5.2 Discussion

14/15

(1) The five EC 3.1.1.3 chemical equations used were all far from the target and belonging to different clusters.

[Reason 1] The most important descriptors were not weighted when combining highly correlated descriptors.

→ Consider a method to select only a small number of important descriptors.

[Reason 2] Several factors other than the amount of change in characteristic values have affected the EC number prediction.

Example (Target 2 reaction)

- Use solvent (tert-amyl alcohol)
- Shaking at 50°C for 20 hours



EC chemical equation


- Occurs in nature
- Chemical solvents is not used basically.

→ Create features that take into account factors such as the reagents, the solvent, the experimental environment, and the combination ratio used in the experiment.

(2) Investigating reason the predicted EC numbers are located near the target seems to lead better enzyme predictions.

These EC numbers are likely to be little-known enzymes.

The number of references in BRENDA⁵

EC 3.1.1.80 = 2, EC 3.1.1.10 = 12, EC 3.1.1.75 = 81  EC 3.1.1.3 = 366

→ Examine whether these EC enzymes are better than EC 3.1.1.3 through experiments.

Conclusion

This study proposed a method for predicting EC numbers to **reduce the time required to search for enzyme candidates**.

- A feature vector was created using amount of changes in characteristic values for each chemical equation.
- The clustering of feature vectors was performed using SOM.

Enzymes of EC 3.1.1.3 were not selected as the most suitable enzymes. Further examination of the predicted EC numbers seemingly result in a discovery of better enzymes than those of EC 3.1.1.3.

Future Work

- Development of a method for dimensionality reduction by **detecting important descriptors that should be kept** to describe the structural changes of chemical reactions in more detail.
- Focus on a method to **select the appropriate combination of descriptors that can most accurately classify EC classes** to predict the optimal EC numbers for the targets.