ICIC Express Letters
Part B: Applications
Volume **14**, Number **7**, July **2023**

ICIC International ©2023 ISSN 2185-2766
pp. 663–672

CHEMOINFORMATICS USING DIMENSIONALITY REDUCTION 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
Graduate School of Engineering
[2]Department of Information Systems Engineering
[3]Department of Biotechnology
Faculty of Engineering
Toyama Prefectural University
5180 Kurokawa, Imizu, Toyama 939-0398, Japan
u255018@st.pu-toyama.ac.jp; { iwasaki; asano }@pu-toyama.ac.jp
*Corresponding author: okuhara@pu-toyama.ac.jp

Received November 2022; accepted January 2023

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.*

Keywords: EC number, Chemoinformatics, Dimensionality reduction, Clustering

1. **Introduction.** The global outbreak of COVID-19 is currently increasing the need for new drug development. Because of it, the chemoinformatics has attracted much attention. Chemoinformatics analyzes properties and structures of chemical compounds or uses machine learning method to classify, design, and predict chemical reactions.

In the field of organic synthesis, enzymes are increasingly used as biocatalysts in the design and prediction of chemical reactions. Compared to chemical catalysts, biocatalysts are known to be environmentally friendly and allow chemical reactions to proceed more efficiently. Therefore, it is becoming one of the important factors to predict the most suitable enzyme for a particular reaction to produce desired products. Enzymes are assigned an EC number (Enzyme Commission number) consisting of four pairs of numbers and are classified according to which reaction they catalyze and which bond or substrate they act on [1, 2]. While organic chemists spend time searching for enzymes by consulting databases or working with enzyme experts, predicting the optimal EC number for a given reaction allows them to move quickly to the next experimental step, such as selecting which enzyme product of that number to use.

There are many studies about EC number prediction, using protein sequences [3, 4], structural properties of compounds [5, 6], physical and chemical properties [7], etc. While enzymes are classified according to their properties up to the third digit of the EC number, the fourth digit distinguishes enzymes with the same properties by name. Therefore, as it is difficult to predict the fourth digit, many studies predict up to the third digit. This study focuses on the prediction of the fourth digit of the EC number. In contrast to

DOI: 10.24507/icicelb.14.07.663

previous studies that predict the fourth digit using protein sequences [3, 4], the proposed method not only predicts using the amount of change in physical and chemical properties from reactants to products, but also visualizes the similarity between chemical reactions that are assigned EC numbers.

Typical chemical reactions data using enzymes belonging to EC numbers are obtained from KEGG (Kyoto Encyclopedia of Genes and Genomes) [8]. The EC number of the chemical reaction with the highest similarity is predicted as the best enzyme by comparing the changes in the characteristic values of the target chemical reaction with those of the EC number chemical reaction.

This paper is organized as follows. In this section, the background and purpose of this study were explained. In Section 2, enzymes and EC numbers are described. The next section introduces the structural representation of compounds in chemoinformatics and the clustering method used in this study. In Section 4, the proposed method for EC number prediction is explained. Next, the procedure of experiments, result and discussion are described. Finally, a "Conclusion" ends this paper.

2. **Enzymes and EC Numbers.** Enzymes are proteins that catalyze chemical reactions necessary in living organisms and are indispensable for their survival. They are classified according to their properties by EC numbers, which consist of four pairs of numbers, X.X.X.X. The first number is classified into seven categories according to which reaction it catalyzes: 1 (Oxidoreductases), 2 (Transferases), 3 (Hydrolases), 4 (Lyases), 5 (Isomerases), 6 (Ligases), and 7 (Translocases) [1, 2]. The second number indicates which bond the enzyme acts on, the third number indicates which substrate (compound) it reacts with and the coenzyme information it requires, and the fourth number indicates the name of the enzyme (in the order of registration) belonging to the first to third combination numbers (EC X.X.X). Figure 1 shows an image of the EC number classification. For example, carboxylesterases belong to the enzymes called hydrolase and catalyze a hydrolysis reaction of compound with a carboxyl bond (ID C02391) that are classified as ester bond shown in the bottom of Figure 1.

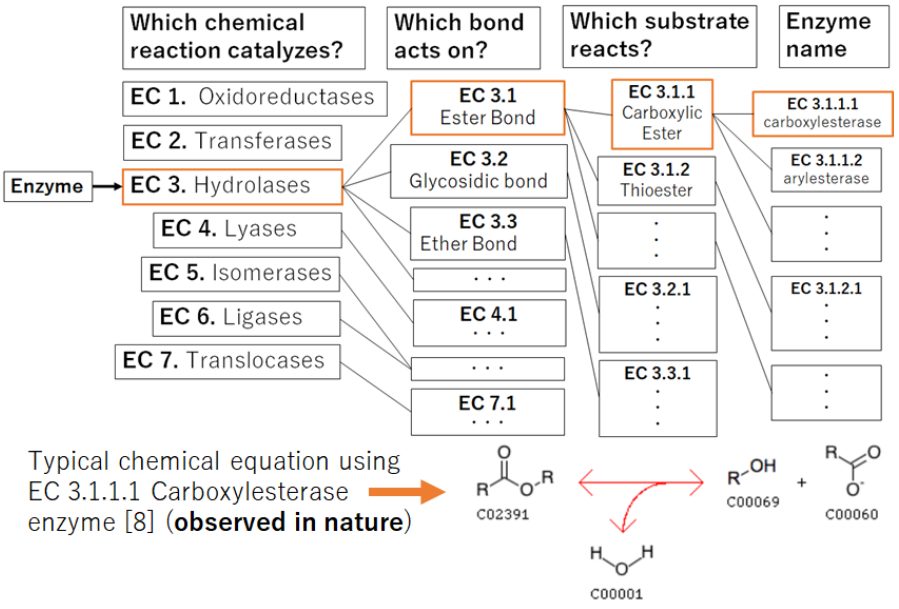
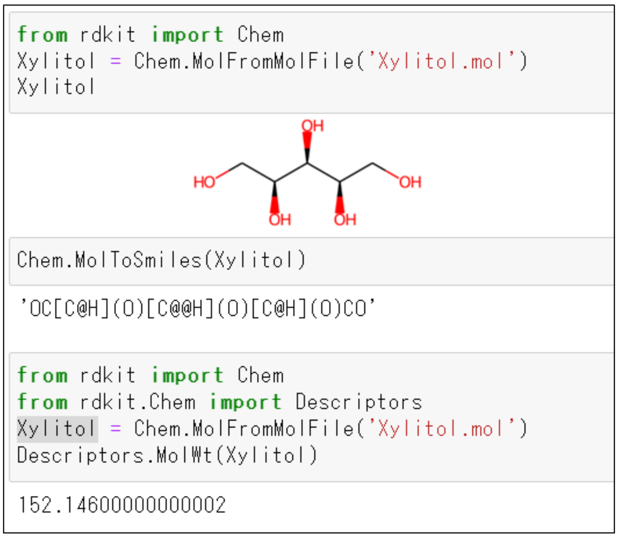


FIGURE 1. Classification of enzymes by EC number (based on [8])

3. Chemoinformatics and Information Technology.

3.1. **Structural representation of compounds.** Chemoinformatics is the study of predicting and classifying chemical reactions by representing the structure of compounds in a format that is easy to handle on a computer. One way to express the characteristics of a compound on a computer is to quantify its physical properties and chemical characteristics. These property values are called descriptors. A compound can be represented as a multidimensional feature vector with many property values.

The Python library, RDKit [9], handles this representation method. It can read chemical structure information files obtained from databases and create objects to draw structural formula. By calculating characteristic values of compounds from the structural formula object, it is possible to evaluate the similarity between compounds and make predictions by machine learning. Figure 2 shows how the structural formula of a compound is drawn using RDKit, and the results of calculating the Molecular Weight (MolWt).



```
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
```

FIGURE 2. Compound information using RDKit

3.2. **Clustering method.** This section describes a clustering method used in dimensionality reduction.

Agglomerative clustering (complete linkage method)

Each data is considered as a cluster, and the distance between the data in one cluster and the data in another cluster is calculated using Euclidean distance, etc. The pair with the farthest distance is considered to be the cluster distance. The two clusters with the smallest cluster distance are merged one after another until the specified number of clusters is reached. If the data belonging to clusters C_1 and C_2 are $\mathbf{x_1}$ and $\mathbf{x_2}$, respectively, the distance between $\mathbf{x_1}$ and $\mathbf{x_2}$ is $d(\mathbf{x_1}, \mathbf{x_2})$ and the distance between the clusters is $d(C_1, C_2)$ [10], the distance between clusters in the complete linkage method is as follows:

$$d(C_1, C_2) = \max_{\mathbf{x_1} \in C_1, \mathbf{x_2} \in C_2} \{d(\mathbf{x_1}, \mathbf{x_2})\} \tag{1}$$

4. Proposed Method.

4.1. **Prediction of EC numbers using changes in characteristic values.** In this study, the structural changes of the target chemical equation and the representative chemical equation of the EC number (EC chemical equation) are compared, and the EC number of the most similar EC chemical equation is predicted as the optimal enzyme candidates. Here, a representative chemical equation refers to a reaction change from a substrate to a product that occurs in nature, such as in living organisms, and is registered as one or

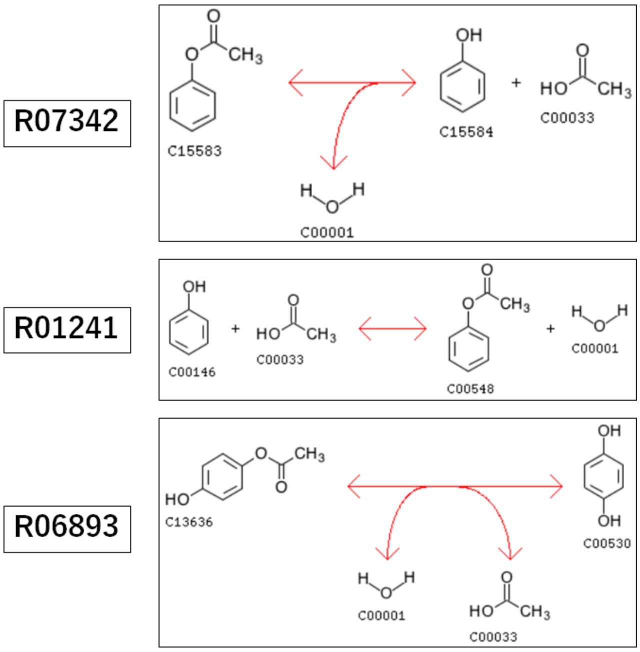


FIGURE 3. Typical chemical equation for EC 3.1.1.2 (based on [8])

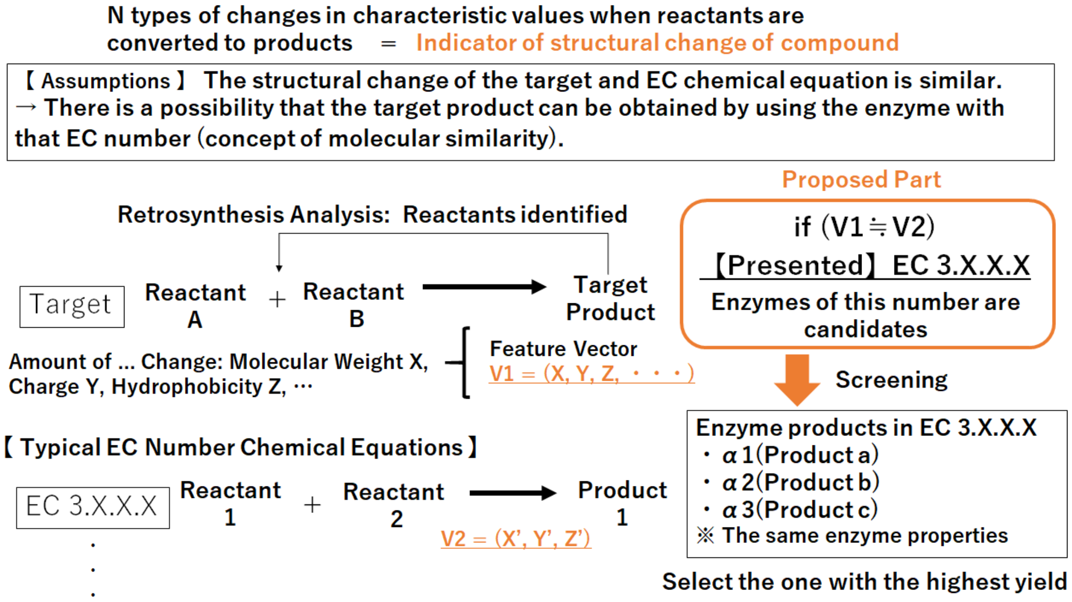

N types of changes in characteristic values when reactants are converted to products = Indicator of structural change of compound
【Assumptions】 The structural change of the target and EC chemical equation is similar.
→ There is a possibility that the target product can be obtained by using the enzyme with that EC number (concept of molecular similarity).
Proposed Part
Retrosynthesis Analysis: Reactants identified
Target
Reactant A
+
Reactant B
Target Product
Amount of ... Change: Molecular Weight X, Charge Y, Hydrophobicity Z, …
Feature Vector
V1 = (X, Y, Z, ・・・)
if (V1≒V2)
【Presented】 EC 3.X.X.X
Enzymes of this number are candidates
Screening
【Typical EC Number Chemical Equations】
EC 3.X.X.X
Reactant 1
+
Reactant 2
Product 1
V2 = (X', Y', Z')
Enzyme products in EC 3.X.X.X
・α1(Product a)
・α2(Product b)
・α3(Product c)
※ The same enzyme properties
Select the one with the highest yield

FIGURE 4. Comparison of similarity of chemical equation

more for each EC number. As an example, in EC 3.1.1.2 of KEGG ENZYME [8], 3 typical chemical equations are registered with R numbers as shown in Figure 3.

In the comparison of structural changes, the image is as shown in Figure 4. In the target chemical equation, it is known which reactant is used to obtain the desired product by retrosynthesis analysis. If the structural change from reactant to product of the target chemical equation is similar to that of an EC chemical equation, we assume that using the enzyme of the chemical equation in the target will increase the efficiency of the reaction to obtain the target product in high yield. This is based on the concept of molecular similarity used in chemistry [12].

The following characteristic value changes are used as an indicator of structural change. When the number of reactants and products in a chemical equation is two each, the characteristic value of reactant i is RT_i and that of product i is PD_i. In this case, the change in characteristic value cv_j $(j = 1, 2, \ldots, n)$ for n descriptors of physical and chemical property values is defined as follows:

$$cv_j = (PD_1 + PD_2) - (RT_1 + RT_2) \tag{2}$$

For each chemical equation, the amount of change in the characteristic value of n descriptors is derived. The n-dimensional feature vector with these elements is used as the feature of the structural change of the chemical equation. The feature vector $\boldsymbol{DF}_i$ $(i = 1, 2, \ldots, m)$ for m chemical equations are expressed as follows:

$$\boldsymbol{DF}_i = (cv_{i1}, cv_{i2}, \ldots, cv_{ij}, \ldots, cv_{in}) \tag{3}$$

The similarity of the feature vectors between the target and each EC chemical equation is evaluated, and the EC number of the chemical equation with the highest evaluation is predicted as the best enzyme candidate. SOM (Self-Organizing Map) [11] is used to evaluate the similarity.

4.2. **Dimensionality reduction by agglomerative clustering.** A complete linkage method of agglomerative clustering is used for the 208 descriptors to adjust the number of descriptors and reduce the dimensionality of the feature vectors [14]. The agglomerative clustering implemented in sklearn of Python [15] is used as the clusterring program. In this study, the inverse of the correlation coefficient between descriptors is used as the distance between descriptors in each cluster. When the correlation coefficient between descriptors u and v is s_{uv}, $1/s_{uv}$ is the distance between descriptors. A distance matrix such as Table 1 is created by Pyhton for clustering. Here, elements with a correlation coefficient of 1 are set to 0.

Table 1. Distance matrix

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

In quantitative structure-activity relationship analysis, which predicts the properties of structurally similar compounds by modeling the structure of the compounds with physical and chemical properties, setting the threshold of the correlation coefficient in the range of 0.95 to 0.9999 is shown to be sufficient to reduce the number of descriptors [13]. To achieve as much dimensionality reduction as possible, when creating new descriptor clusters, we set the merge threshold as $s_{uv} \geqq 0.95$, that is, $1/s_{uv} \leqq 1/0.95 \approx 1.05$. In this case, the distance between clusters, $d(C_1, C_2)$, is obtained as follows from Formula (1).

$$d(C_1, C_2) = \max_{u \in C_1, v \in C_2} \frac{1}{s_{uv}} \tag{4}$$

Using these, clustering among descriptors is performed in the following procedure.

1) Merge the pair with the minimum distance among the pairs of descriptors that satisfy $1/s_{uv} \leqq 1.11$.
2) If descriptor pairs satisfying $1/s_{uv} \leqq 1.11$ exist between clusters, merge clusters C_1 and C_2 with minimum distance $d(C_1, C_2)$. Repeat until there are no more descriptor pairs that satisfy the condition.
3) After clustering is completed, a table of correspondence is obtained between cluster numbers and descriptors belonging to the clusters.
4) Standardize and average the sequence of the amount changes in the characteristic value of each descriptor in the cluster, and use the composite descriptor as a new descriptor.

Dimensionality reduction is performed by combining descriptors in the same cluster and replacing them with a composite descriptor, cluster X (X is the cluster number).

4.3. **Clustering of feature vectors by SOM.** SOM [11] is a clustering method that maps multidimensional data to lower dimensions and visualizes them. It clusters the feature vectors of the target and EC chemical equation. Enzymes of EC number whose chemical equation is located near the target are presented as the best enzyme candidates.

The SOM program uses source code created with reference to R language files output by KH Corder [16]. The input data are the feature vectors of each chemical equation after dimensionality reduction and standardized for all of it. The labels of the plot points are the target and the EC number of the chemical equation.

The number of units is 400 (20 × 20) and the shape of them is hexagonal. The training is divided into two stages: a rough ranking phase and a convergence phase [17]. The number of training cycles is 1,000 for the first phase and 200,000 for the second phase. After the SOM is run, feature vectors are mapped onto each winner unit, and agglomerative clustering with color coding is performed. This clustering is done by the Ward's method using Euclidean distance. In this case, the number of clusters is 9.

5. Experimental Results and Discussion.

5.1. **Outline of numerical experiment.** The experimental flow of this study is described below. First, information about each chemical equation is obtained from KEGG [8] and PubChem [18]. Next, the physical and chemical property values of the compounds are calculated using 208 descriptors in RDKit, and the feature vectors of each chemical equation are created by obtaining the change in characteristic value. After dimensionality reduction, the feature vectors of each chemical equation are mapped using SOM.

Target chemical equation and evaluation method of the proposed method

This time, we focus on the chemical equation for the first step of synthesis in the process of producing molnupiravir. The target chemical equation is shown in Figure 5. Target 2 is the original synthesis, which selectively esterifies the primary alcohol of ribose (first term on the left side) [19]. Here, eight enzyme products are screened. The enzyme product with the best yield is the enzyme classified as EC 3.1.1.3. In this experiment, target 1, which is considered to be more reactive, was probably tested before synthesis of target 2. The reaction of target 1 is also a possible reaction when the enzyme of EC 3.1.1.3 is used. Therefore, we evaluate the proposed method based on how closely the feature vector of the EC 3.1.1.3 chemical equation is located to the feature vectors of these two chemical equations in SOM.

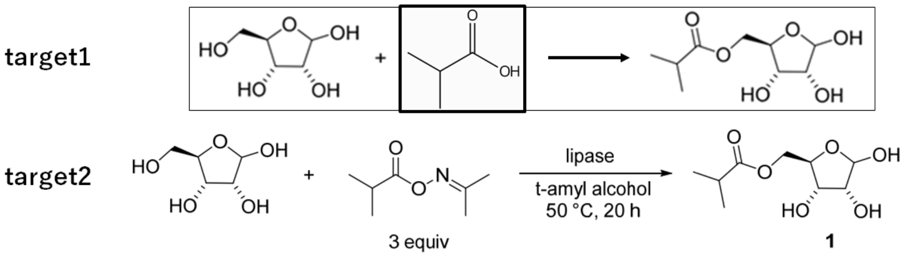


FIGURE 5. Target chemical equation (based on [19])

EC chemical reaction to be compared

The EC chemical equations to be compared with the target using SOM are about 100 types in the EC 3.1.1 class. This is because the target is an esterification reaction, and EC 3.1.1, Carboxylic Ester Hydrolases, are considered appropriate as the enzymes to be predicted.

Data acquisition and creation of feature vectors

First, the information about EC chemical equation was collected using the source code [20] which obtains EC numbers and reactions of KEGG. Next, the structural information files of the compounds in the chemical reaction were obtained from PubChem, and the feature vectors of the EC chemical equation were created using RDKit. The used EC chemical equations consist of two reactants and two products. The feature vectors of the targets were obtained using the structure information files available on SciFinder^n [21].

Table 2 shows the feature vectors for target 1 and each chemical equation. The rows indicate the target and EC number, and the columns contain the names of 208 descriptors. The numbers represent the fourth number of EC 3.1.1. The numbers after the period distinguish the chemical equations which have the same EC number. The numbers after the underscore classify the chemical equations registered in multiple EC numbers. From this table, descriptors with nan-values or divergent values in the elements and with 98% or more of the elements having the same value were excluded. Finally, dimensionality reduction was performed on the feature vector consisting of 113 descriptors.

5.2. Experimental results and discussion.

Dimensionality reduction of feature vectors

For the comparison between the EC chemical equation and the target 1 (target 2) chemical equation, 12 (12) synthetic descriptors were created, and 84 (84) dimensional feature vectors were obtained.

SOM results of EC number prediction for targets

Figure 6 shows the results for target 1. "E" means EC numbers other than EC 3.1.1 class. EC 3.1.1.80 and EC 3.1.1.45 chemical equations are located near target 1 and belong to the same cluster as it. The chemical equations in KEGG are shown in the left part of Figure 7. These are the esterification reaction of 2-Maleylacetate and Norajmaline, respectively. Near target 2, the EC reaction equations shown in the right part of Figure 7 were located. EC 3.1.1.75 and EC 3.1.1.101 are hydrolysis reactions on Poly-beta-hydroxybutyrate and Polyethylene terephthalate, respectively.

Discussion 1

Although five representative chemical equations of EC 3.1.1.3 were used, including duplicates of other EC numbers, all of them were far from the targets and belonged to different clusters from them as shown in Figure 6. There are two reasons for this.

First, the most important descriptors were not weighed while combining highly correlated descriptors allowed to reduce the dimentionality of feature vectors. We would like to consider a method to select only a small number of important descriptors.

Second, it is possible that factors other than the amount of change in characteristic values have affected the EC number prediction. For example, in the target 2 reaction, tert-amyl alcohol is used as the solvent, and the product is produced by shaking at 50°C for 20 hours [19]. On the other hand, the EC chemical equation is a reaction that occurs in nature and organic solvents are not used basically. It is necessary to create features that take account of factors such as the reagents used in the experiment, the solvent, the experimental environment, and the combination ratio.

Discussion 2

While the studies that predicted the fourth digit [3, 4] focus on the evaluation of prediction accuracy, the advantage of this study is not only to predict the EC number for the target given the correct answer, but also to visualize the similarity among other chemical reactions by SOM. Although the prediction accuracy was not satisfactory because the correct answer, EC 3.1.1.3, was far from the target, features such as a cluster of chemical equations containing polymers around target 2 were able to be observed. Investigating these properties of the EC chemical equations located near these targets seems to lead to

Table 2. Feature vector table of target 1 and EC chemical equation

	MaxEStateIndex	MinEStateIndex	MaxAbsEStateIndex	MinAbsEStateIndex	qed	MolWt	HeavyAtomMolWt	ExactMolWt	NumValence
target1	−8.378152	0.949632	−8.378152	−0.144028	−0.003234	0.0	−0.0	0.0	
33	−7.632875	0.794822	−7.632875	−1.064815	−0.015391	0.0	−0.0	0.0	
6	−6.597222	0.946759	−6.597222	−0.949074	−0.081471	−0.0	0.0	0.0	
1	−6.486111	0.972222	−6.486111	−0.675926	−0.003359	1.008	0.0	1.007276	
7_8	−7.085822	0.351574	−7.085822	−0.914074	−0.156941	0.0	0.0	0.0	
···	···	···	···	···	···	···	···	···	
111	−8.902683	0.535378	−8.902683	−0.657129	−0.032571	0.0	0.0	0.0	
118	−8.839073	0.535378	−8.839073	−0.575822	0.010726	0.0	−0.0	−0.0	
2	−7.241293	0.225648	−7.241293	−0.869167	−0.18409	−0.0	0.0	0.0	
5.1	−8.727962	0.779028	−8.727962	−0.513312	0.04305	1.008	0.0	1.007276	
26.1	−7.502687	0.544916	−7.502687	−0.540144	0.010529	27.002	28.01	26.98709	

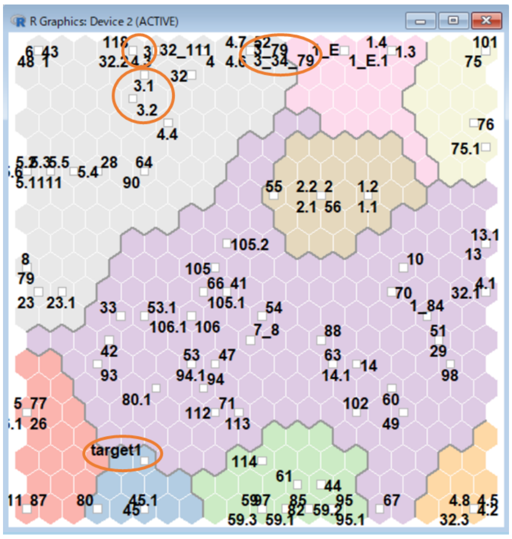


FIGURE 6. Clustering results of target 1 and EC 3.1.1 chemical equations by SOM

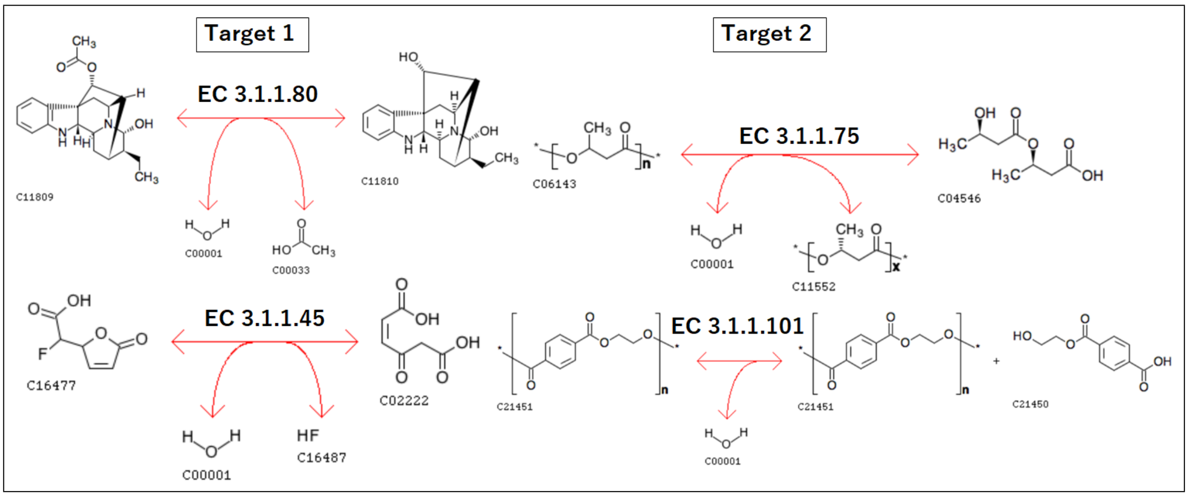


FIGURE 7. EC 3.1.1 chemical equations located near the targets (based on [8])

better enzyme predictions. In addition, these EC numbers are likely to be little-known enzymes, because the number of references for these EC numbers in BRENDA is only about 30, compared to about 370 for EC 3.1.1.3. Therefore, it is necessary to examine whether these enzymes of the EC numbers are better than those of EC 3.1.1.3 through validation experiments.

6. Conclusion.

Data was obtained from KEGG, PubChem, and other sources, and a feature vector was created by calculating the amount of changes in characteristic values consisting of 208 descriptors for each chemical equation using RDKit. The clustering of feature vectors was performed by SOM. The amount of changes in characteristic values of EC 3.1.1.80 was determined to be similar to that of target 1. In the case of target 2, EC 3.1.1.10 and 3.1.1.75 were presented. Although enzymes of EC 3.1.1.3 were not selected as the most suitable enzymes, we expect that further examination of these predicted EC numbers would result in a discovery of better enzymes than those of EC 3.1.1.3.

As a future study, we will 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.

REFERENCES

[1] *Proposed Japanese Name for Newly Established Enzyme Classification of EC7*, https://www.jbsoc.or.jp/notice/ec_translocase.html, Accessed on 2022.1.15 (in Japanese).

[2] Y. Shirokane, Enzyme classification and nomenclature, *JAS Information*, 2017, http://jasnet.or.jp/4-shuppanbutu/pickup/17.10.pdf, Accessed on 2022.1.15 (in Japanese).

[3] A. Dalkiran, A. S. Rifaioglu, M. J. Martin, R. Cetin-Atalay, V. Atalay and T. Doğan, ECPred: A tool for the prediction of the enzymatic functions of protein sequences based on the EC nomenclature, *BMC Bioinformatics*, vol.19, no.334, https://doi.org/10.1186/s12859-018-2368-y, 2018.

[4] J. Y. Ryu, H. U. Kim and S. Y. Lee, Deep learning enables high-quality and high-throughput prediction of enzyme commission numbers, *Proc. of the National Academy of Sciences*, vol.116, no.28, https://doi.org/10.1073/pnas.1821905116, 2019.

[5] Y. Yamanishi, M. Hattori, M. Kotera, S. Goto and M. Kanehisa, E-zyme: Predicting potential EC numbers from the chemical transformation pattern of substrate-product pairs, *Bioinformatics*, vol.25, no.12, https://doi.org/10.1093/bioinformatics/btp223, 2009.

[6] Q.-N. Hu, H. Zhu, X. Li, M. Zhang, Z. Deng, X. Yang and Z. Deng, Assignment of EC numbers to enzymatic reactions with reaction difference fingerprints, *PLoS ONE*, vol.7, no.12, https://doi.org/10.1371/journal.pone.0052901, 2012.

[7] D. A. R. S. Latino and J. Aires-de-Sousa, Assignment of EC numbers to enzymatic reactions with MOLMAP reaction descriptors and random forests, *Journal of Chemical Information and Modeling*, vol.49, no.7, pp.1839-1846, https://doi.org/10.1021/ci900104b, 2009.

[8] *KEGG: Kyoto Encyclopedia of Genes and Genomes*, https://www.genome.jp/kegg/kegg_ja.html, Accessed on 2022.1.17.

[9] *The RDKit Documentation*, https://www.rdkit.org/docs/GettingStartedInPython.html#list-of-available-descriptors, Accessed on 2022.2.6.

[10] *Clustering (Cluster Analysis)*, https://www.kamishima.net/jp/clustering/#bib_cutting, Accessed on 2022.2.3 (in Japanese).

[11] T. Kohonen, Self-organized formation of topologically correct feature map, *Biological Cybernetics*, vol.43, pp.59-69, 1982.

[12] M. A. Johnson and G. M. Maggiora, *Concepts and Applications of Molecular Similarity*, Wiley, 1990.

[13] A. Rácz, D. Bajusz and K. Héberger, Intercorrelation limits in molecular descriptor preselection for QSAR/QSPR, *Molecular Informatics*, vol.38, nos.8-9, https://doi.org/10.1002/minf.201800154, 2019.

[14] [*With Python Code*] *Variable Selection by Correlation Coefficients and Clustering of Variables*, https://datachemeng.com/variable_selection_and_clustering_based_on_r/, Accessed on 2022.1.29 (in Japanese).

[15] *sklearn.cluster.AgglomerativeClustering*, https://scikit-learn.org/stable/modules/generated/sklearn.cluster.AgglomerativeClustering.html, Accessed on 2022.2.3.

[16] M. Fukushima, *Similarity and Event Detection by Behavior Pattern Analysis from Environmental Awareness Life Logs*, Bachelor Thesis, Toyama Prefectural University, 2018 (in Japanese).

[17] *KH Coder 3 Reference Manual*, https://khcoder.net/en/manual_en_v3.pdf, Accessed on 2022.2.3.

[18] *PubChem*, https://pubchem.ncbi.nlm.nih.gov/, Accessed on 2022.1.17.

[19] *Supporting Information for: Evolving to an Ideal Synthesis of Molnupiravir, an Investigational Treatment for COVID-19*, https://chemrxiv.org/engage/api-gateway/chemrxiv/assets/orp/resource/item/60c75348469df4bbbff44d17/original/mk-4482-gen-2-supporting-information-final.pdf, Accessed on 2022.2.6.

[20] *Data Acquisition Using KEGG API*, https://rstudio-pubs-static.s3.amazonaws.com/472676_97a2c135b5704dc1b52f7759b73466e8.html, Accessed on 2022.12.28 (in Japanese).

[21] *CAS SciFinder^n*, https://scifinder-n.cas.org/, Accessed on 2022.1.23.