Development of Machine Learning Models for Comprehensive Prediction of Enzyme Annotations

渡邉, 直暉

(Degree)
博士（工学）

(Date of Degree)
2023-03-25

(Date of Publication)
2024-03-01

(Resource Type)
doctoral thesis

(Report Number)
甲第8647号

(URL)
https://hdl.handle.net/20.500.14094/0100482395

（別紙様式 3）

論文内容の要旨

氏　　名　渡邉　直暉

専　　攻　応用化学専攻

論文題目（外国語の場合は，その和訳を併記すること。）

Development of Machine Learning Models for Comprehensive Prediction of Enzyme Annotations

(酵素アノテーションの網羅的予測のための機械学習モデルの開発)

指導教員　荻野　千秋

（注）2，000 字～4，000 字でまとめること。

Chapter I

Introduction

Enzymes are applied to microorganisms in order to biosynthesize a wide range of industrial chemicals, pharmaceuticals, antibiotics and food additives. The variety of useful substances has increased with the rapid development of various technologies such as genetic engineering, synthetic biology and metabolic engineering. Existing enzymes may be capable of reacting with newly characterized substrates to synthesize new products in addition to known natural reactions. Furthermore, novel enzymes can be harnessed to catalyze previously unknown reactions. Therefore, novel enzyme discovery and the expansion of current metabolic pathways are required to increase the production of target compounds.

Protein sequence information has been registered in various biological databases. The number of unannotated sequences is explosively increasing due to the development of genome sequencing technology. Numerous hypothetical and uncharacterized enzyme functions are accelerating. The number of available enzyme sequences could potentially increase by explosively increasing in the number of unannotated sequences. New enzyme functions exist within the unannotated protein sequences. Novel enzyme discovery is necessary to expand the pathways that can be accessed by metabolic engineering for the biosynthesis of functional compounds. However, experimental verification of all available unknown protein sequences cannot be achieved due to high costs and time limitations. Therefore, a valid computational method to predict enzyme functions from sequence information is needed to discovery novel enzymes within a huge number of unannotated sequences.

The most basic solution in computational methods is to use Basic Local Alignment Search Tool (BLAST) algorithm in which highly similar enzyme sequences to input sequences are search from protein sequence databases and their functions are inferred based on the most similar annotated enzymes. However, the method cannot predict the function of uncharacterized enzymes with low similarity to annotated enzymes. Furthermore, it has recently been reported that machine learnings included in computational methods performed better than BLAST. Therefore, several studies have recently proposed machine learning models for prediction of various biological annotations. Machine learning can process vast amounts of available protein sequences and is suitable for the mass prediction of various biological functions. To build

prediction models for classification tasks, protein sequence information with biological annotation should be transformed to the feature vectors using feature extractions and target classes which want to predict the annotations should be built. The feature vectors and classes are learned by a machine learning method and then a prediction model is built. Most importantly, the use of various feature extractions and machine learning methods are needed because the optimal solutions for the methods depend on the annotation to be predicted.

Conventional enzyme function prediction models built from enzyme sequences using machine learnings have recently been developed to discover novel enzymes. However, this strategy can only predict the information up to the enzymes. In order to synthesize functional compounds using microorganisms, it is necessary to simultaneously predict enzymes and even substrates and products in enzymatic reactions. This prediction will not only discover new enzymes, but also apply to the discovery of novel metabolic pathways.

Chapter II
Exploration and Evaluation of Machine Learning Based Models for Predicting Enzymatic Reactions

In this chapter, enzymatic reaction prediction models are developed to discover novel enzymes and enzymatic reactions using several machine learnings which have the potential to acquire new knowledge from a large number of datasets. First, Enzyme (E) models are built from enzyme sequence information using same strategy as conventional enzyme function predictions. Next, Substrate-Enzyme (SE) and Substrate-Enzyme-Product (SEP) models combined enzyme sequence information with compound chemical structure information predict enzyme-compound combinations in enzymatic reactions.

While accuracies of E models are not optimal, SE models and SEP models predict enzymatic reactions with high accuracy using all tested machine learning-based methods. In comparison to BLAST, most of SE and SEP models more correctly predict. In this chapter, SEP-Random Forests (RF) model achieves the best performance using *Escherichia coli* K-12 test. Various metrics indicate that the current strategy of combining sequence and chemical structure information is effective at improving enzymatic reaction prediction. However, these models cannot exclude unlikely

enzyme-compound combinations, because they do not learn the combinations in which enzymatic reactions do not occur. The current models are successfully built to provide the basis for predicting enzymatic reactions. However, these models cannot exclude unlikely enzyme-compound combinations, because they do not learn the combinations in which enzymatic reactions do not occur. Moreover, more extensive testing using enzyme and isozyme sequences from additional species is needed because the models are evaluated using only *E. coli* K-12 test.

Chapter III

Comprehensive Machine Learning Prediction of Extensive Enzymatic Reactions

In order to improve the previous prediction models, new E, SP, SE and SEP models are developed using several machine learning algorithms, including Deep Neural Network (DNN) by updating training datasets and feature extractions. Moreover, these SE and SEP models can predict whether or not enzymatic reactions will occur. The models are evaluated using test datasets including the enzyme sequences derived from various species.

Improvements in prediction performances for these SE and SEP models over that of the previous SEP-RF model result of the same test indicate that the updated methods are more effective for prediction of enzymatic reactions. In addition, SE and SEP models do not require rigorous optimization of datasets and feature extractions when comparing the process of building these E models. The SEP-DNN model exhibits the highest prediction accuracy with Macro F_1 scores up to 0.966 using a number of enzyme sequences derived from various species and with robust prediction of unknown enzymatic reactions that are not included in the training data. This model can predict more extensive enzymatic reactions in comparison to previously reported model regardless of the test datasets. The current models will help to discover new enzymes with novel functions, existing enzymes that may react with new substrates and unknown combinations of substrates-enzymes-products that can expand current metabolic pathways in the future.

On the other hand, the enzymatic reaction prediction models need to be further improved in the several points. First, the number of available compounds in the training data is much smaller than the number of available enzyme sequences, and prediction results greatly depend upon the included compound information. All SE and

SEP models show lower prediction accuracy for the test reactions with compounds that do not exist in the training data and for reactions which is low similarity with training data. Therefore, the enzymatic reaction prediction models can be further improved by optimizing feature extractions. It is also necessary to consider reducing the number of dimensions for feature vectors.

Second, negative training datasets for SE and SEP models consist of random SE and SEP combinations in order to prevent the models from relying only on compound feature information. The current models tend to misjudge some reactions as negative because most of the negative samples are similar to positive samples. Thus, improved methods to build negative training data are needed.

Chapter IV
EnzymeNet: Residual Neural Networks based model for Enzyme Commission number prediction

The strategy in the previous chapters has hypothesized that amino acid sequences used in the predictions are enzyme. However, the sequences may not be the enzyme when actually predicting unknown reactions. Excluding non-enzyme proteins is necessary before predicting with the appropriate combinations of sequence and compound using enzymatic reaction prediction models. Moreover, the previous reported models predict the sequences with numerous consecutive identical amino acids, which are found within unannotated sequences, as enzymes.

Therefore, Enzyme Commission (EC) number prediction models named EnzymeNet are developed using Residual Neural Networks (ResNet), which is included in deep learning, to predict enzyme annotations for enzymatic reaction in addition to exclude non-enzyme sequences and the exceptional sequences described above. EC number system is used to classify enzymes using 4 digits based on the reaction type. Deep learning has enabled to predict the 3-dimensional structures of proteins from protein sequences. The results indicate that deep learning methods are capable of capturing the extensive enzyme features within a sequence. Moreover, several studies have reported models for prediction of protein annotations from sequence information using Convolution Neural Network (CNN) which is often used in image recognition. Therefore, EC number prediction models are built using ResNet, which contains the structures of multiple CNN layers, and predict enzyme functions while capturing the

extensive features of enzyme sequences.

In order to reduce the number of steps in EnzymeNet prediction, the models roughly predict the EC number first digits and then determine the full EC numbers. As a result, EnzymeNet models predict with higher accuracy than previously reported models and with robust prediction of the enzymes which are low similarity with training data. The robustness of EnzymeNet models will lead to discover novel enzymes for biosynthesis of functional compounds using microorganisms.

Chapter V
General Conclusion and Future work

Combining the EC number prediction models with enzymatic reaction prediction models enables to predict comprehensive enzyme annotations related to enzymatic reactions. First, the EC number prediction models select only enzyme from the amino acid sequences and roughly estimate a reaction catalyzed by the putative enzyme. Next, based on the EC number prediction results, the enzymatic reaction prediction models predict the substrate that is likely to react with the enzyme and the product that is likely to be synthesized. This system which combines two prediction steps is evaluated only on annotated data, and therefore must be optimized depending on the targets to be predicted. The current system will help to select enzyme sequences and discover novel enzymatic reactions including missing links in metabolism and biosynthesis pathways for the production of useful substances using microorganisms.

（別紙１）　　論文審査の結果の要旨

氏名	渡邉　直暉		
論文題目	Development of Machine Learning Models for Comprehensive Prediction of Enzyme Annotations （酵素アノテーションの網羅的予測のための機械学習モデルの開発）		
審査委員	区　分	職　名	氏　　名
	主　査	教　授	荻野　千秋
	副　査	教　授	山地　秀樹
	副　査	教　授	南　秀人
	副　査	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・プロジェクトリーダー・統括研究員	荒木　通啓
	副　査		印

要　　旨

近年、様々な生物学的アノテーションを予測するための機械学習法がいくつかの研究により報告されている。機械学習は膨大な量の利用可能なタンパク質配列を処理することができ、様々な生物学的機能の大量予測に適している。予測モデルを構築するためには、生物学的アノテーションを持つタンパク質配列情報を、特徴抽出により特徴ベクトルに変換し、ターゲットアノテーションをクラスとして構築する必要がある。最も重要な点は、予測するアノテーションによって、様々な特徴抽出と機械学習の手法が用いられることである。従来、新規酵素を発見するために、酵素の配列から機械学習を用いて構築された酵素機能予測モデルが最近開発されている。しかしながらこの戦略では、酵素までの情報しか予測できない。微生物を用いて機能性化合物を合成するためには、酵素はもちろん、酵素反応における基質や生成物まで同時に予測することが必要である。新しいこの予測手法は、新しい酵素の機能発見だけでなく、新しい代謝経路の発見にも応用できる。予備検討委員会では、学位論文草稿の各章について以下の様に確認を行った。

第 1 章では、これまでのタンパク質、酵素機能予測に関して、その背景から、現在至る現状を、網羅的に総括し、候補者の用いた手法に関して、他の手法との比較を行い、その優位性、問題点などを洗い出している。第 2 章以下では、洗い出した問題点を解決するために、どのようなことができるかについて検証を行った。

第 2 章では、多数のデータセットから新たな知識を獲得する可能性を持つ複数の機械学習を用い、新規酵素や酵素反応を発見するための酵素反応予測モデルを開発した。まず、酵素の配列情報から、従来の酵素機能予測と同じ手法で酵素(E)モデルを構築する。次に、酵素の配列情報と化合物の化学構造情報を組み合わせた基質-酵素（SE）モデルおよび基質-酵素-生成物（SEP）モデルにより、酵素反応における酵素と化合物の組み合わせ予測を行った。E モデルの精度は最適とは言えないが、SE モデルおよび SEP モデルは、テストしたすべての機械学習ベースの手法を使用して、高い精度で酵素反応を予測した。また、BLAST と比較すると、SE モデル、SEP モデルの多くで正解率が高かった。ここで、SEP-Random Forests モデルは、*Escherichia coli* K-12 テストを用いて最高のパフォーマンスを達成した。このように、配列情報と化学構造情報を組み合わせる戦略は、酵素反応予測の向上に有効であることが示された。
この研究内容に関しては、英文雑誌に以下の様に掲載されたことを確認した。**Watanabe, N.**; Murata, M.; Ogawa, T.; Vavricka, C. J.; Kondo, A.; Ogino, C.; Araki, M. Exploration and Evaluation of Machine Learning-Based Models for Predicting Enzymatic Reactions. Journal of Chemical Information and Modeling. 2020, 60, 1833–1843.

第 3 章では、更なる予測モデルの改良のため、学習データセットと特徴抽出を更新し、Deep Neural Network を含む複数の機械学習アルゴリズムを用いて、新しい E、SP、SE、SEP モデルでの予測アルゴリズムを開発した。これらの SE および SEP モデルは、同じテストにおける従来の SEP-RF の結果よりも予測性能が向上しており、更新された手法が酵素反応の予測に有効であることが示された。また、E モデルの構築プロセスを比較すると、SE モデルや SEP モデルはデータセットや特徴抽出の厳密な最適化が不要であることがわかる。その結果、SEP-DNN モデルにおいて、様々な生物種に由来する多数の酵素配列を用い、Macro F1 スコアが最大 0.966 と、最高の予測精度を示した。このモデルは、テストデータセ

氏名	渡邉　直暉

ットによらず、従来報告されているモデルよりも広範な酵素反応を予測することができ、学習データに含まれない未知の酵素反応も堅牢に予測することが可能となった。

更に特徴抽出を最適化することにより、反応予測モデルをさらに改善することが可能となった。複合特徴情報のみに依存したモデルを防ぐために、SE、SEP モデルの負の学習データセットは、ランダムな SE、SEP の組み合わせで構成され、モデルの予測精度を向上することに大きく寄与した。

この研究内容に関しては、英文雑誌に以下の様に掲載されたことを確認した。Watanabe, N.; Yamamoto, M.; Murata, M.; Vavricka, C. J.; Ogino, C.; Kondo, A.; Araki, M. Comprehensive Machine Learning Prediction of Extensive Enzymatic Reactions. The Journal of Physical Chemistry B. 2022, 126, 36, 6762-6770.

これまでの章では、予測に用いるアミノ酸配列が酵素であると仮定してきた。しかし、実際に未知の反応を予測する際には、その配列が酵素でない可能性がある。酵素反応予測モデルを用いて適切な配列と化合物の組で予測する前に、非酵素タンパク質を除外することが必要である。そこで第 4 章では、ディープラーニングに含まれる ResNet を用いて EnzymeNet という酵素委員会（EC）番号予測モデルを開発し、非酵素配列の除外に加えて、酵素反応に対する酵素アノテーションの予測を行った。深層学習により、タンパク質配列からタンパク質の 3 次元構造を高い精度で予測することが可能となった。その結果、ディープラーニングは配列内の 3 次元構造の特徴を捉えることが可能であることが示された。さらに、画像認識でよく使われる CNN（Convolution Neural Network）を用いて、配列情報からタンパク質のアノテーションを予測するモデルもいくつかの研究で報告されている。そこで、EC 番号予測モデルは、複数の CNN の構造を含む ResNet を用いて構築し、酵素配列の構造的特徴を捉えながら予測することを試みている。EnzymeNet 予測では、ステップ数を減らすため、EC 番号の 1 桁目を大まかに予測し、その後、完全な EC 番号を決定するモデルとなっている。その結果、EnzymeNet は従来報告されているモデルよりも高い精度で予測し、学習データとの類似性が低い酵素もロバストに予測することができた。

この結果については、現在、WEB アプリの作成が完成し、論文投稿を行っていることを確認した。

本学位論文では、EC 番号予測モデルと酵素反応予測モデルを組み合わせることで、酵素反応に関連する包括的な酵素アノテーションを予測することが可能になった。まず、EC 数予測モデルは、アミノ酸配列から酵素配列のみを選択し、その酵素が触媒する反応を大まかに推定する。次に、EC 予測結果に基づいて、酵素反応予測モデルが、その酵素と反応しそうな基質と合成されそうな生成物を予測する。このシステムは、アノテーションされたデータのみで評価されるため、予測するターゲットによって最適化する必要がある。本システムは、微生物を用いた有用物質生産において、酵素配列の選択、代謝・生合成経路のミッシングリンクを含む新規酵素反応の発見などに役立つと考えられる。

以上のように、本論文は、酵素のアミノ酸配列、および基質と生成物の特徴抽出を行う事で、その酵素の機能を予測することができるアルゴリズムを構築できる可能性を示している。更に、構築されるアルゴリズムは、酵素の機能予測に留まらず、酵素の機能改変にも大きく寄与する可能性があり、今後の酵素工学、およびその産業において大きく寄与するものであり、重要な知見を得たものとして価値ある集積である。従って、提出された論文は工学研究科学位論文評価基準を満たしており、学位申請者の　渡邉　直暉　は、博士（工学）の学位を得る資格があると認める。