Vol.2020-BIO-64 No.2
2020/12/7

深層学習を用いた酵素機能予測法の開発

川崎悠二 [1,a)]　里村和弘 [1]　長田直樹 [1]　遠藤俊徳 [1]

概要：本研究では深層学習によってアミノ酸配列から酵素の機能予測を行う方法の開発を目的としている．深層学習のモデルのハイパーパラメータ（層の数，ユニット数）は，学習を複数回繰り返し行い，その中で最も正解率の高い場合の値を採用した．深層学習の入力には，次の２つの特徴量を用いた．1 つ目は各アミノ酸の共起関係を基にした配列のベクトル表現である．これらのベクトル表現は，固有の配列パターンを文脈として捉えるために Doc2Vec を用いて作成した．2 つ目は PSSM(Position Specific Scoring Matrix)である．PSI-BLAST によって獲得した PSSM を用いることで，進化情報を考慮し，相同性の低い配列に対してもより正確な機能予測を行うことができると考えた．EC 番号の main classes の予測を行った場合 99.5%，sub classes では 95.4%，sub-sub classes では 85.7%の正解率で機能予測を行うことができた．

キーワード：深層学習，酵素，doc2vec

1. 背景

現在，様々な生物においてゲノム解読が進んでおり，新規のアミノ酸配列のデータが急速に増加している．これらの配列の機能を実験的に決定するには，多くの時間とコストが掛かってしまう．そのため，計算機による機能予測の役割がますます重要になっていると言える．既に，酵素のアミノ酸配列から立体構造を予測し，機能を解明しようとする試みが多数行われているが，正確に立体構造を予測することはとても困難な課題となっている．そこで，アミノ酸配列から直接機能を予測することができれば，立体構造予測の困難を回避して新規の配列に対しても機能を決定することができると考えられる．

機能の分類基準としてはＥＣ番号が広く用いられている．ＥＣ番号は酵素の階層的分類であり，1 つの酵素に 1 つの番号が割り当てられている．酵素の反応の種類や基質，補酵素などに基づいて 4 桁の数字で機能分類がされており，桁が増えるごとにより詳細な分類となっている．

本研究では，ＥＣ番号を基に深層学習を用いて酵素の機能を予測する方法の開発に取り組んだ．

2. 方法

2.1 深層学習

深層学習は与えられたデータセットの特徴的なパターンを認識して，他クラス分類を行うことができる．今回用いた学習モデルは隠れ層が 4 層からなるニューラルネットワークを用いた．モデルのハイパーパラメータは，隠れ層の数は 1～8，ユニット数は 50～600 の範囲で試行を行い，その中で最も正解率の高かった場合の値を採用した．

2.2 Doc2Vec

Doc2Vec [1] は任意の長さの文章を固定長のベクトルに変換することができる技術であり，文章の id と単語を学習モデルに入力して，次に出現する単語を予測するというタスクを解きながら文章のベクトル表現を獲得する．これにより，固有な配列パターンを文脈として捉えることができると考えられる．

2.3 PSSM(Position Specific Scoring Matrix)

タンパク質の進化的情報を組み込むために，PSSM を特徴量として用いた．各アミノ酸配列の PSSM は PSI_BLAST [2] を用いて作成し，20×20 行列の二次元データを 400 次元の 1 次元データに変換してデータを作成した．

3. 材料

3.1 全データ

酵素のデータベースである BRENDA [3] から 53,169 個のアミノ酸配列を取得した．酵素機能の分類クラスは 180 種類である．なお，ある機能に属する配列数が 10 個に満たない場合はデータから除いた．

3.2 評価用データセット

相同性を 40%以下にカットしたデータセットを作成した．各 EC 番号に属するデータ数は，EC1 が 1,618 個，EC2 が 3,450 個，EC3 が 2,791 個，EC4 が 679 個，EC5 が 518 個，

1 北海道大学大学院情報科学院
a) Kawasaki.y@ibio.jp

ⓒ2020 Information Processing Society of Japan

EC6 が 776 個の合計 9,832 個である．

4. 結果

4.1 Doc2vec によって作成された特徴量を用いた予測結果

全データ 52,169 個を 8:2 の割合で学習用データとテスト用データに分割して学習を行った．EC 番号 1 桁目では 99.65%，2 桁目では 95.43%，3 桁目では 85.94%の正解率で予測を行うことができた．

表 1 学習結果(Doc2Vec)

Layer	Data	Accuracy(%)
First Layer	Train	99.68
	Test	99.65
Second Layer	Train	98.70
	Test	95.43
Third Layer	Train	94.09
	Test	85.94

4.2 相同性 40%以下のデータセットに対する予測結果

PSSM と Doc2Vec によって作成された特徴量を組み合わせた入力データを用いて予測を行った結果，相同性を 40%以下でカットしたデータセットに対して EC 番号 1 桁目まで 87.3%の正解率で予測を行うことができた．

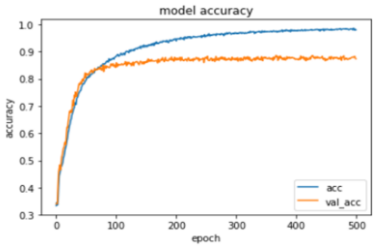

図 1 学習結果(相同性 40%以下)

4.3 各特徴量の評価

表 2 特徴量別の正解率

特徴量	Accuracy(%)
Doc2Vec	39.63
PSSM	73.82
Doc2Vec ＆ PSSM	87.34

相同性を 40%以下にカットしたデータセットに対する各特徴量別の正解率を求めた．この結果から，PSSM は相同性の低いデータに対して有用であることが分かる．しかし，PSSM はデータの整形によって配列中の各アミノ酸の位置情報が失われてしまうといった問題がある．そのため，PSSM と Doc2Vec を組み合わせた特徴量を導入することで，位置情報を損なうことなく相同性の低いデータセットに対しても 87%の正解率で予測を行うことができた．

5. 考察

5.1 深層学習による予測結果の傾向

EC1.1.2 の予測結果から分かるように，予測を間違う傾向として，EC 番号の 2 桁目までは共通であり，3 桁目が異なる EC 番号に誤って予測されてしまっていることが分かる．3 桁目までの予測の正解率が低い理由としては，酵素が結合する化合物の特徴をうまく学習することができていないからだと考えられる．例えば，酸化還元酵素（EC1）では 2 層目のドナーの種類による分類では正解率 80%を上回るが，3 層目のアクセプターの種類による分類では 50%を下回る分類群が存在する．この傾向は酸化還元酵素以外の他の機能群にもみられた．

表 3 例）EC1.1.2 の予測結果

EC number	分類数
1.1.2	484
1.1.1	66
1.1.3	38
1.1.5	3

また，判別が困難である酵素の特徴として構造的に類似した結合に作用することが考えられる．EC3.1.31 の加水分解酵素(ヌクレアーゼ)は RNA または DNA のホスホジエステル結合に対して作用するが，予測の際に EC3.1.26 , EC3.1.22 に予測されることが多かった．これらの酵素は，DNA に作用するか RNA に作用するか，または 3’末端か 5’末端かの違いがある．つまり，核酸のどの位置の結合に作用するかの小さな違いである．そして，これらの酵素とは別に EC3.2.1 に誤って予測してしまうこともあった. EC3.2.1 は *O*-グリコシド結合に作用する酵素である．例としてアミラーゼがあり，六単糖同士の結合を加水分解する機能をもつ．ヌクレアーゼとは五炭糖か六単糖かの違いはあるが，酵素全体の大きさに比べて小さな構造の違いだといえる．以上のことから，学習モデルによって予測された酵素間には構造的類似性があることが示唆された．

5.2 データ数と正解率の関係

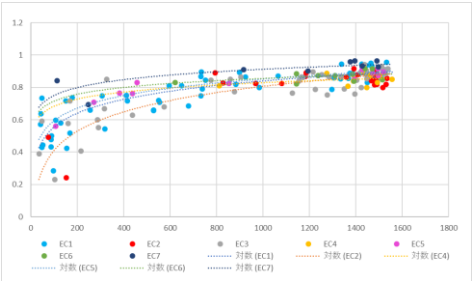

図 2 データ数と正解率

データ数と正解率の間の相関係数は 0.81 であり，正の相関が見られた．従って，正解率を向上させるためには十分なデータ数を確保することが必要であるが，新たに酵素の機能を決定しデータ数を増やすことは現実的ではない．そ

こで，予測結果を 1 つの機能に絞るのではなくて，複数の機能を候補として予測を行うことで確実に機能の選択肢を絞ることができる．

6. 結論

Doc2Vec で作成した特徴量を用いて予測を行った結果，EC 番号のメインクラス予測で 99.65%の正解率を達成し，Doc2Vec による配列のベクトル表現が機能予測に効果的であることが示唆された．相同性が低いデータセットに対しては PSSM が有用であるが，配列中の各アミノ酸の位置情報が失われてしまうという問題を改善するために Doc2Vec によって作成した特徴量を組み合わせることで正解率をさらに改善することができた．また，より詳細な機能を決定するためには，ＥＣ番号 3 桁目までの予測精度を向上させる必要がある，これは酵素のデータ数の増加もしくは，モデルのパラメータ最適化を進めることで改善できる．

予測結果の傾向から酵素が作用する結合の構造的類似性が示唆された．このことから，学習モデルの予測結果に含まれる酵素の間には，少なからず機能的類似性があると考えられ，類似性の指標として活用することが期待される．

7. 参考文献

[1] Quoc Le , Tomas Mikolov
Distributed Representations of Sentences and Documents
In Proceedings of ICML 2014**.**

[2] A.A. Schaffer, L. Aravind, T.L. Madden, S. Shavirin, J.L. Spouge, Y.I. Wolf, E.V. Koonin, S.F. Altschul
Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements
Nucleic Acids Research, 29 (2001), pp. 2994-3005

[3] Ida Schomburg, Antje Chang, Christian Ebeling, Marion Gremse, Christian Heldt, Gregor Huhn, Dietmar Schomburg
BRENDA, the enzyme database: updates and major new developments
Nucleic Acids Research, Volume 32, Issue suppl_1, 1 January 2004, Pages D431–D433

ⓒ2020 Information Processing Society of Japan