

Subject Section

This is a title

Corresponding Author^{1,*}, Co-Author² and Co-Author^{2,*}

¹Department, Institution, City, Post Code, Country and

²Department, Institution, City, Post Code, Country.

*To whom correspondence should be addressed.

Associate Editor: XXXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation:

Results:

Availability:

Contact: name@bio.com

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Abstract

The accurate prediction of enzyme commission numbers (EC numbers) is not only crucial for the classification and understanding of newly discovered enzymes but also for completing the annotation of already known enzymes. Therefore, developing a reliable method for predicting EC numbers is of great importance.

However, due to insufficient data, enzyme function prediction using machine learning is an ongoing challenge. In this paper, we propose several methods for predicting enzymes in three different problem categories (Table 1). Throughout the developing of our models, we used a variety of different input features and machine learning algorithms, of which the best will be thoroughly reviewed in this paper.

2 Introduction

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetur id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

3 Approach

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetur id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Table 1. Description of subproblem categories

Level	Description	Best performing method	F1 score
0	Binary classification	Random Forest	score
1	Main class classification	Feedforward neuronal network	score
2	Subclass classification	Feedforward neuronal network	score

4 Methods

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetur id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

5 Discussion

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetur id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

6 Conclusion

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetur id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

7 Supplementary Information

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetur id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

7.1 K-nearest neighbors algorithm using ncd vectors

A less popular approach of transforming string like input features into numerical values is the normalized compression distance (ncd) algorithm. The ncd algorithm is based on the idea that the similarity of two strings can be measured by the amount of information needed to describe one string given the other string. The ncd of two strings x and y is defined as follows:

$$ncd(x, y) = \frac{C(xy) - \min(C(x), C(y))}{\max(C(x), C(y))}$$
 (1)

where $C(x)$ is the length of the compressed string x and $C(xy)$ is the length of the concatenated string xy .

We implemented this algorithm in python using `gzip`, which is a loss less compression algorithm based on a combination of LZ77 and Huffman encoding. Rigler et al. (2007)

The ncd algorithm was used to transform the amino acid sequences into numerical vectors by comparing each sequence to all other sequences in the training dataset. This resulted in a n -dimensional numerical vector for each sequence, where n is the amount of sequences in the training dataset where each position in the vector represents the ncd of the sequence to the corresponding sequence in the training dataset. These vectors were then used as input for the k-nearest neighbors algorithm. Due to the exponential computational complexity of the ncd algorithm, we had to under sample the non enzyme dataset to match the amount of samples in our enzyme dataset, meaning the positiv in the train dataset were balanced.

When inferring unseen data, the ncd input vector was calculated by comparing it to all sequences in the training data set, thus also resulting in a n -dimensional numerical vector. This means that the performance on new data is largely dependent on the training data set.

The performance of the k-nearest neighbors algorithm using ncd vectors compared to a random baseline is shown in figure 1. Although the mean F1 score of the k-nearest neighbors algorithm using ncd vectors lies at 0.728, it did not perform better than the random baseline, which had a F1 score of 0.843. This indicates that the ncd approach has a worse precision and recall than the random baseline. At the same time both classifiers have a low MCC score of 0.2 and 0.01 respectively, which indicates that both classifiers are not better than random guessing.

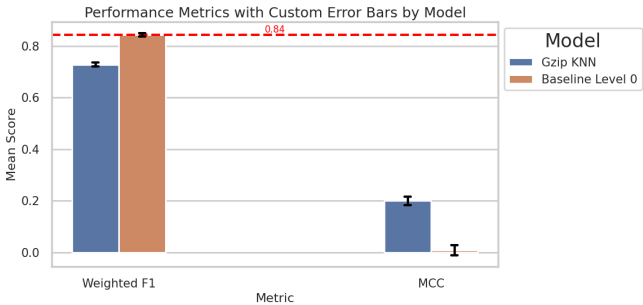


Fig. 1. Performance on test dataset compared to random baseline

The reason for the poor performance of the k-nearest neighbors algorithm using ncd vectors is most likely due to the ncd algorithm not being suited for protein sequences as shown in Matsumoto et al. (2000) as well as the test dataset not being balanced, while the training dataset was.

References

Matsumoto, T., Sadakane, K., and Imai, H. (2000). Biological sequence compression algorithms. *Genome Informatics*, 11, 50–51.

References

Rigler, S., Bishop, W., and Kennings, A. (2007). Fpga-based lossless data compression using huffman and lz77 algorithms. In *2007 Canadian Conference on Electrical and Computer Engineering*, pages 1235–1238.