BIOTRANS
LA ROCHELLE 2023

N°877 / OC / PC
TOPIC(s) : Enzyme discovery and engineering / Artifical intelligence / computational methods

Discovery enzyme-substrate interactions via numerical representations strategies and machine learning algorithms

AUTHORS

David MEDINA / UNIVERSIDAD DE MAGALLANES, AV. PDTE. MANUEL BULNES 01855, PUNTA ARENAS
Roberto URIBE-PAREDES / UNIVERSIDAD DE MAGALLANES, AV. PDTE. MANUEL BULNES 01855, PUNTA ARENAS
Alvaro OLIVERA-NAPPA / UNIVERSIDAD DE CHILE, BEAUCHEF 851, SANTIAGO
Marcelo NAVARRETE / UNIVERSIDAD DE MAGALLANES, AV. PDTE. MANUEL BULNES 01855, PUNTA ARENAS

PURPOSE OF THE ABSTRACT

Several computational models have been proposed to predict the affinity between an enzyme and its substrate. In most cases, structural information of both the enzyme and the substrate, together with molecular dynamics, are used as input for constructing predictive systems based on complex deep learning architectures. This strategy hampers its widespread application due to the high computational cost involved. In addition, in many cases, obtaining the secondary structure of the enzymes requires the use of computational systems for its prediction, adding up to the need of computation. Methods based on linear representation can address this issue. Representation via amino acid sequence for enzymes and SMILE for substrates have provided promising results, that nevertheless lack generalization capability and only allow classification recognition systems, limiting the application for the estimation of affinity between enzyme and substrate.

This work explores alternatives to solve the efficient affinity prediction between an enzyme and a substrate. To this aim, we collected all the reactions reported in the KEGG database. Next, we build a data set of the affinities between all the enzymes and the substrates through virtual screening. This step enabled the calculation of a numerical value representing affinity. Lastly, we explored computational strategies for the numerical representation of the enzymes, substrates, and the interaction complex. Since our overall strategy aims to save computational power we included linear information, i.e., amino acid sequences and SMILES. Different amino acid codings were explored, including methods based on physicochemical properties, applications of transformations through convolutions, like the Fast Fourier Transform, and strategies based on pre-trained models via natural language processing methods. In the case of SMILES, methods based on learning numerical representations such as Mol2Vec, graph junction tree, or similar were explored. The interaction complexes were approached using concatenation methods, non-linear convolution, or combination strategies. First, the concatenation strategies were explored to build the dataset by combining the different methods explored for enzymes and substrate and building the predictive models using classic machine learning algorithms and deep learning architectures, like convolutional neural networks (CNN) and long short-term memory (LSTM). In this case, the best performance was 0.35 for the combination of secondary structure-property to represent the enzymes and Mol2Vec for representing the smiles with a random forest algorithm. Then, the convolutional or combination methods were explored, employing PCA strategies and Kernel-PCA techniques for all combinations of representation between enzyme and substrate representation. In this case, the best performance was achieved by the application of PCA techniques to the dataset built by the esm1b pre-trained model (in the case of enzymes) and Mol2Vec (in the case of SMILES) in the predictive model trained using LSTM architectures, achieving a performance of 0.56 Pearson coefficient. The model developed with natural language strategies, LSTM, and linear combinations strategies achieved better results than the concatenation methods, demonstrating synergy between the NLP and LSTM architectures. The generated model has been developed for usability purposes to correlate mutational changes in enzymes with the affinity between substrates. In addition, this model allows simulating interactions, which facilitates the discovery of new reactions between enzymes and substrates, becoming a relevant and advantageous strategy for landscape

navigation and the discovery of new interest molecules. In future work, the model will be the improvement of the architecture and compare the results with methods available in the literature.

FIGURES

FIGURE 1

FIGURE 2

KEYWORDS

Machine learning algorithms | Numerical representations | enzyme-substrate affinity prediction | Protein languange models

BIBLIOGRAPHY