Deep Neural Network Architecture for Drug-Target Interaction Prediction

Nelson R. C. Monteiro^(✉), Bernardete Ribeiro, and Joel P. Arrais

Department of Informatics Engineering (DEI),
Center for Informatics and Systems of the University of Coimbra (CISUC),
University of Coimbra, Coimbra, Portugal
nelsonrcm117@gmail.com

Abstract. The discovery of potential Drug-Target Interactions (DTIs) is a determining step in the drug discovery and repositioning process, as the effectiveness of the currently available antibiotic treatment is declining. Successful approaches have been presented to solve this problem but seldom protein sequences and structured data are used together. We present a deep learning architecture model, which exploits the particular ability of Convolutional Neural Networks (CNNs) to obtain 1D representations from protein amino acid sequences and SMILES (Simplified Molecular Input Line Entry System) strings. The results achieved demonstrate that using CNNs to obtain representations of the data, instead of the traditional descriptors, lead to improved performance.

Keywords: Drug repositioning · Drug-Target Interaction ·
Deep learning · Convolutional Neural Network ·
Fully Connected Neural Network

1 Introduction

The discovery of new and potential drugs is declining, as there is an increase of the misuse of the available medicine, causing a resistance effect to these kinds of agents [1]. Therefore, establishing effective computational methods is decisive to find new leads. Computational methods for DTI prediction are divided into 3 main approaches [4], namely ligand based, docking simulation and chemogenomic. Ligand based approaches are built upon the concept that similar molecules have similar properties and therefore should bind to the same group of proteins [6]. Docking Simulation is used for structure based drug design, where the interaction is simulated and scored using 3D structures [5]. Chemogenomic approaches are based on the chemical, genomic and/or the pharmacological space [8]. Due to the amount of available data and computational power, machine learning [3] and deep learning [9] are pursued over the traditional methods. We propose a deep learning approach to predict DTIs using 1D raw data, amino acids sequences and SMILES. We exploit the particular ability of CNNs

to obtain 1D representations, which are features that express local dependencies or patterns, that can then be used in a Fully Connected Neural Network (FCNN), acting as a binary classifier. Coelho et al. (2016) [7] dataset was used to evaluate and validate the model. Additionally, we compared our model with different approaches, specifically random forest (RF), a FCNN architecture and support vector machine (SVM).

2 Methods

2.1 Data

The protein sequences were extracted from UniProt and the SMILES strings were collected from PubChem exclusively, in their canonical format. Since we are using protein sequences and SMILES strings directly, each amino acid and character, respectively, is considered as a feature. Therefore, it was necessary to define a threshold based on their length. An information threshold of 95% was used, resulting in a maximum length of 1205 for the protein sequences and 90 for the SMILES. All entries duplicated or containing missing characters in one of the datasets were also removed, resulting in 16011 (5839 positive and 10172 negative) samples for training and 7926 (3012 positive and 4914 negative) for testing. Table 1 summarizes the amount of unique drugs, targets and drug-target interactions extracted from the databases used to create the datasets and Table 2 the amount of unique proteins, drugs and number of targets for the training and testing datasets. Plus, only Yamanishi et al. (2008) [8] and DrugBank positive entries were used for training and testing, respectively.

Table 1. Unique drugs, targets and DTIs used to create the datasets.

	Positive		Negative	
	DrugBank	Yamanishi et al. (2008)	BioLip	BindingDB
Drugs	1328	790	894	12454
Targets	706	1371	636	404
DTI	3530	7206	1223	14985

Table 2. Unique drugs, targets and number of targets for the training and testing datasets.

	Unique		Number of targets	
	Targets	Drugs	1	>1
Training	1790	9583	8026	1557
Testing	1068	5718	4884	834

2.2 Data Representation

We used Yu et al. (2010) [10] protein substitution table, which organizes amino acids into 7 groups according to their physicochemical properties. Each amino acid was encoded into an integer based on the corresponding group. In the case of SMILES strings, a simple integer encoding was used to transform each character of the strings into an integer.

2.3 Model

The proposed approach is based on a deep learning architecture (Fig. 1) to predict DTIs using directly protein sequences and SMILES (1D raw data). One-Hot Layer was used to assign a binary variable for each unique integer value, converting every integer into a binary vector. Two series of 1D convolutional layers were used, one for the protein sequences and another for the SMILES. A global max pooling layer was applied after each series of convolutional layers to reduce the spatial size of each feature map to its maximum representative feature. The obtained deep representations were concatenated into a single feature vector, characterizing a DTI pair. The resulting feature vectors were then used as the input of a FCNN architecture. Dropout was applied between each fully connected layer to reduce the overfitting by deactivating a percentage of neurons. This architecture was followed by an output layer.

Fig. 1. Drug-Target Interaction model architecture.

2.4 Hyperparameter Optimization Approach

Two simultaneous methods, combined with grid search, were used to determine the best model, early stopping and model checkpoint. Considering the fact that dividing the training set into training and validation and applying cross validation led to high scores for every model architecture and set of parameters in both training and validation, it was not possible to select the best model using this approach, as every model was supposedly good in the validation set but the results were inconsistent when applied to the testing set. Therefore, we decided to use all the training set for training and the testing set to evaluate the model performance in each epoch. Since the testing set is highly imbalanced, F1-score was used for this evaluation (Fig. 2). Table 3 summarizes the hyper-parameters obtained from grid search.

3 Results

We applied grid search for all the models in order to accurately compare and evaluate the performance. The descriptors used were the same as the original

Fig. 2. Hyperparameter optimization model based on grid search.

Table 3. Parameter settings for the proposed model.

Parameters	Value
Number of convolutional layers	3
Number of dense layers (FC)	3
Number of filters	[128, 256, 384]
Filter length (proteins)	[3,4,5]
Filter length (compounds)	[3,4,5]
Epochs*	500
Hidden neurons	[128,128,128]
Batch size	256
Dropout rate	0.5
Optimizer	Adam
Learning rate	0.0001
Loss function	Binary cross entropy
Activation function (CNN)	ReLU
Activation function (FC)	ReLU
Activation function (output)	Sigmoid
Class weights (imbalanced classes)	{0: 0.36, 1: 0.64}

Table 4. Prediction results of testing set.

		Model					
		CNN + FCNN		Random forest		SVM RBF	
		CNN representations	Descriptors	Descriptors	CNN representations	Descriptors	CNN representations
Metric	Sensitivity	0.861	0.827	0.809	0.821	0.739	0.769
	Specificity	0.961	0.963	0.989	0.992	0.989	0.993
	F1-Score	0.895	0.876	0.886	0.896	0.842	0.864
	Accuracy	0.923	0.911	0.921	0.927	0.894	0.908

work [7], which contains a total of 432 protein descriptors and 323 drug descriptors, and collected using PyDPI package [2]. The protein descriptors are divided into amino acid composition, Moran autocorrelation and CTD (Composition, Transition, Distribution) descriptors. On the other hand, drug descriptors are divided into molecular constitutional, molecular connectivity, molecular property, kappa shape and charge descriptors, molecular access system (MACCS) keys and E-state fingerprints.

Due to the fact that the traditional split of the training set into training and validation led to inconclusive results, as mentioned in Sect. 2.4, the results obtained are related to “internal validation”. The testing set was used to discover the best set of parameters, thus there is not an external validation set. Nonetheless, given the disparity of the training and testing set and the low similarity of the drug pairs that constitute them, the results are considered as valid and relevant.

The differences in performance between all models can be interpreted as a result of the difference between using deep representations, obtained from pro-

tein sequences and SMILES strings, and global descriptors. Besides, it's also possible to highlight the difference between applying traditional machine learning and deep learning approaches (Table 4). The results obtained validate the effectiveness of convolutional neural networks as a feature engineering tool and their capacity to automatically surmise and identify important sequential and structural regions for drug-target interactions. Another observation is that using an end-to-end deep learning method resulted in a high sensitivity (0.861) and specificity (0.961) when compared to the other models, which obtained a high specificity and a low sensitivity.

Funding. This research has been funded by the Portuguese Research Agency FCT, through D4 - Deep Drug Discovery and Deployment (CENTRO-01-0145-FEDER-029266).

References

1. Aslam, B., et al.: Antibiotic resistance: a rundown of a global crisis. *Infect. Drug Resist.* **11**, 1645–1658 (2018). <https://doi.org/10.2147/IDR.S173867.30349322>[pmid]
2. Cao, D.S., Liang, Y.Z., Yan, J., Tan, G.S., Xu, Q.S., Liu, S.: PyDPI: freely available python package for chemoinformatics, bioinformatics, and chemogenomics studies. *J. Chem. Inf. Model.* **53**(11), 3086–3096 (2013). <https://doi.org/10.1021/ci400127q>
3. Cao, D.S., et al.: Computational prediction of drug target interactions using chemical, biological, and network features. *Mol. Inform.* **33**(10), 669–681 (2014). <https://doi.org/10.1002/minf.201400009>
4. Chen, R., Liu, X., Jin, S., Lin, J., Liu, J.: Machine learning for drug-target interaction prediction. *Molecules* **23**(9), 2208 (2018). <https://doi.org/10.3390/molecules23092208>
5. Cheng, A.C., et al.: Structure-based maximal affinity model predicts small-molecule druggability. *Nat. Biotechnol.* **25**(1), 71 (2007). <https://doi.org/10.1038/nbt1273>
6. Cheng, F., Zhou, Y., Li, J., Li, W., Liu, G., Tang, Y.: Prediction of chemical-protein interactions: multitarget-QSAR versus computational chemogenomic methods. *Mol. BioSyst.* **8**(9), 2373–2384 (2012). <https://doi.org/10.1039/C2MB25110H>
7. Coelho, E.D., Arrais, J.P., Oliveira, J.L.: Computational discovery of putative leads for drug repositioning through drug-target interaction prediction. *PLOS Comput. Biol.* **12**(11), 1–17 (2016). <https://doi.org/10.1371/journal.pcbi.1005219>. 11
8. Gutteridge, A., Araki, M., Kanehisa, M., Honda, W., Yamanishi, Y.: Prediction of drug-target interaction networks from the integration of chemical and genomic spaces. *Bioinformatics* **24**(13), i232–i240 (2008). <https://doi.org/10.1093/bioinformatics/btn162>
9. Tian, K., Shao, M., Wang, Y., Guan, J., Zhou, S.: Boosting compound-protein interaction prediction by deep learning. *Methods* **110**, 64–72 (2016). <https://doi.org/10.1016/j.ymeth.2016.06.024>
10. Yu, C.Y., Chou, L.C., Chang, D.T.H.: Predicting protein-protein interactions in unbalanced data using the primary structure of proteins. *BMC Bioinform.* **11**(1), 167 (2010). <https://doi.org/10.1186/1471-2105-11-167>

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.