Using a Novel Unbiased Dataset and Deep Learning Architectures to Predict Protein-Protein Interactions

Luís Silva
University of Coimbra
Pólo II - Pinhal de Marrocos
Coimbra 3030-290
Email: lspedro97@gmail.com

Carlos Pereira
Polytechnic Institute of Coimbra
Coimbra
Email: cpereira@dei.uc.pt

Joel P. Arrais
University of Coimbra
Pólo II - Pinhal de Marrocos
Coimbra 3030-290
Email:jpa@dei.uc.pt

***Abstract*—Proteins are indispensable to the living organisms and are the backbone of almost all the cellular processes. However, these macromolecules rarely act alone, forming the protein-protein interactions. Given their biological significance it should come as no surprise that their deregulation is one of the main causes to several disease states.**

The sudden surge of interest in this field of study motivated the development of innovative *in silico* methods. Despite the obvious advances in recent years, the effectiveness of these computational methods remains questionable. There is still not enough evidence to support the use of just *in silico* techniques to predict protein-protein interactions not yet experimentally determined. It is proved that one of the primary reasons leading to this situation is the non-existence of a "gold-standard" negative interactions dataset. Contrary to the high abundance of publicly available positive interactions, the negative examples are often artificially generated, culminating in biased samples.

In this paper a new unbiased dataset is presented, that does not overly constraint the negative interactions distribution. Beyond the novel dataset, also distinct deep learning models are proposed as a tool to predict whether two individual proteins are capable of interacting with each other, using exclusively the complete raw amino acid sequences. The obtained results firmly indicate that the proposed models are actually a valuable tool to predict protein-protein interactions, mainly when compared with the existing approaches, while also highlighting that there is still some room for improvement when implemented in unbiased datasets.

1. Introduction

Proteins play vital roles in diverse biological processes, but rarely act as individuals, producing elaborate complexes with other proteins to perform the function they were designed to. The existent established laboratory techniques that predict Protein-Protein Interaction (PPI)s are not efficient, as they are too time and cost consuming. This situation propelled the development of robust and efficient *in silico* prediction methods.

Recently, the increasing amount of available data and computational power paved the way to the implemention of Deep Learning (DL) techniques to predict PPIs [1]. However, despite all the existing prediction techniques, the performance of a model is also heavily dependent on the quality of the PPIs datasets incorporated.

2. Data

2.1. PPIs datasets

Due to the lack of datasets of negative interactions, these are generally computationally generated. Most of the existing datasets incorporate a non co-localization technique, which conceives as a negative interaction a random protein pair whose individual proteins are associated with distinct cellular localization annotations. Nevertheless, this approach can lead to over-optimistic estimates of the accuracy [2].

2.1.1. Unbiased dataset. To limit the number of possible false positives only the human multi-validated physical PPIs available in BioGRID [3] were selected. The negative random sampling method was implemented with a slight variation from the original work [4]. The approach consisted on randomly sampling two distinct proteins from the unique proteins of the multi-validated positive interactions. Finally, the pair sampled was compared with the multi-validated and not multi-validated positive interactions and if it was not already labeled as one of them, then it was considered a new negative interaction. This simple twist yielded a small contamination of the negative set, if even existent, as protein pairs with the slight evidence of potentially being considered positives interactions were taken into account and not constituted a negative interaction.

2.1.2. Pan et al. dataset. It is a dataset exclusively composed by human PPIs. Afterwards the negative interactions were computationally generated by the non co-localization method.

2.1.3. Du et al. dataset. It is a *Saccharomyces cerevisiae* PPIs dataset, in which the negative interactions were also generated with the non co-localization method.

TABLE 1: Results of the best models

	Convolution Neural Network (CNN) Model			Fully Convolutinal Neural Network (FCNN) Model		
	Unbiased dataset	**Pan et al. dataset**	**Du et al. dataset**	**Unbiased dataset**	**Pan et al. dataset**	**Du et al. dataset**
Accuracy	62.5%	98.2%	90.3%	60.7%	98.8%	90.8%
Sensitivity	64.5%	97.6%	89.0%	67.3%	98.7%	89.5%
Specificity	60.4%	98.8%	91.7%	53.9%	98.9%	92.2%
F1-Score	63.3%	98.1%	90.2%	63.2%	98.7%	90.6%

3. Model Architectures

In this paper 2 unique model architectures were explored: CNN and FCNN, which were incorporated with a common architectural aspect between them, the multiple input architecture that processes the proteins of an interaction as separate entities. This approach has seen some use in various works [1] and has been regarded as the one that promotes better performances.

A DL model expects a fixed size input, regardless of the varying lengths of the protein sequences. In this regard, we defined a common length for the protein vectors, the value obtained by the 90 percentile of the distribution of the proteins length of the main datasets [6]. Any protein longer than the determined value was removed from the dataset, and the remaining ones were zero padded to the determined value and were also encoded using one-hot encoding technique.

3.1. CNN

This architecture is composed by an initial block of 3 convolutional layers. The output is then submitted to a flatten layer. From each protein results a vector and these two vectors are then merged into a single one, specific to each interaction, and posteriorly fed to the final block of 3 fully connected layers to learn a function that is capable of distinguish between interacting and non-interacting pairs.

3.2. FCNN

Notwithstanding the significance of the work conducted by Springenberg et al. [7], the architecture incorporated in this paper does not fully match the one used in the original study. Our model is composed solely by convolutional layers that replace the non-convolutional layers of the CNN. The pooling layers of the CNN perform fixed operations on the feature maps with the intention of subsampling them, but this is also easily achievable just by using convolutional layers with no padding and a higher stride, as it was concluded in the original work [7]. Applying convolutional layers to flatten the input is not as destructive as using a flatten layer, because during the convolutional process the model attempts to create a representation that incorporates all the spatial and structural information and compresses all the knowledge of the feature maps on a single neuron, rather than simply linearly flattening the input. Finally the fully connected were also replaced, as their main difference lies on the connectivity of the neurons between layers. Consequently, using a convolutional layer with filters with dimensions that match the dimesions of the output of the previous layers is in essence the same as using fully connected layers.

4. Results and Discussion

4.1. Results

From each of the datasets we produced two models, one for each of the two neural networks architectures considered, which totalizes six different models established and optimized with a grid search technique. It was considered as the best model the one that achieved the highest result on the validation set, which corresponds to 20% of the training set. The results are displayed in Table 1.

4.2. Discussion

Even though benchmark datasets were used as the starting point, the subsequent elimination of some interactions, attributable to the padding constraints, results in the creation of slightly smaller datasets. Nevertheless, the PPIs representation and distribution of each dataset is similar to its respective benchmark dataset, as the larger part of the dataset remains the same. Just by itself this simple fact does not completely invalidate the comparison of the proposed models with the state of the art techniques built upon the respective benchmark datasets. It simply hinders the comparison task, as is is not possible to clearly claim that the models proposed in this work surpass the already established ones, it is only justifiable to assess whether their performances are at a similar level or not, which just corroborates the reliability of the models and validates their predictive potential.

TABLE 2: Accuracy results of some state of the art models built upon one of the benchmark datasets

	Pan et al. dataset	**Du et al. dataset**
Gui et al. [8]	98.1%	-
Wang et al. [9]	97.1%	-
Sun et al. [2]	98.1%	-
Pan et al. [5]	97.9%	-
Wang et al. [10]	97.8%	-
Zhang et al. [11]	-	95.2%
Du et al. [1]	-	92.5%
You et al. [12]	-	89.2%
Zhou et al. [13]	-	88.8%

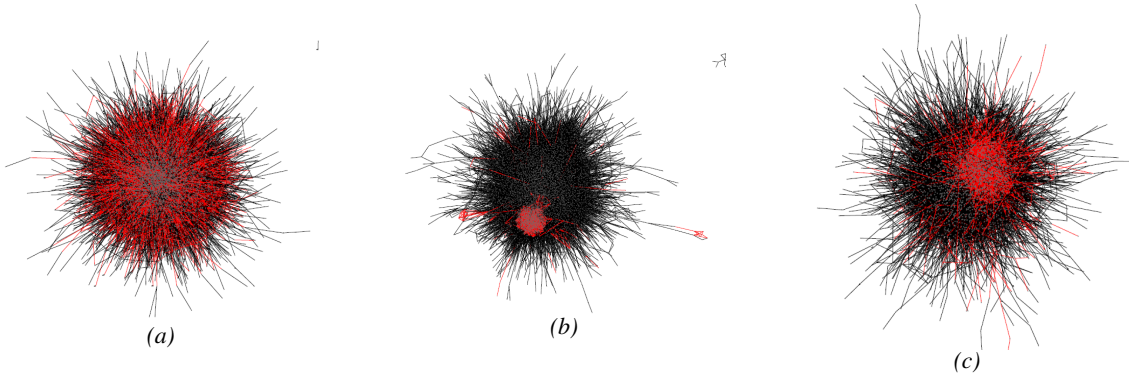


Figure 1: Distribution of the interactions in each of the PPIs domain of the three evaluated datasets. (a) unbiased dataset, (b) Pan et al. dataset, (c) Du et al. dataset. Proteins are represented by the grey nodes, the positive interactions by the black edges and the negative interactions by the red edges.

4.2.1. Comparison with existing techniques. The meticulous analysis of the Table 2 and Table 1 recognizes that the individual models established in this work are on par with the state of the art algorithms, as both techniques achieve high results in the benchmark datasets. It is clear that in some cases the CNN and FCNN models even surpass the state of the art alternatives, but again, those differences may be a consequence of the small variances of the datasets, and not from a performance standpoint. With the results obtained it is only fair to declare that the models are in fact valid alternatives. These findings also dispute several studies that solely rely on protein descriptors to obtain valid protein representations.

The results approve the FCNN architecture as it yields, for the most part, better performances than the CNN. After all, the homogeneous take on the architectures is at least intriguing and deserves further examination.

4.2.2. Impact of the main datasets. Despite the overall excellent results on the benchmark datasets, the same is not verified in the models trained on the unbiased dataset. The models are fundamentally the same, so, the inconsistency has to come from the datasets and the way they are built. The obervation of Figure 1 easily illustrates that the benchmark PPIs datasets are in fact biased. Obviously, when the number of positive interactions is the same as the negative interactions, the negative interactions should present a distribution at least as sparse as the distribution of the positive interactions. Figure 1 does not illustrate such scenario across all datasets. The unmissable biased distribution of the negative interactions is the main cause of the discrepancy between the results of the proposed models, as the more restrained the negative interactions are, the better the results.

A more rigourous examination of the interactions was implemented, for which the degree distribution was integrated. In a protein-protein network the degree can be interpreted as the number of possible interactions a specific protein has. From observing Figure 2 we clearly identify that in the unbiased dataset both the positive and the negative interactions present a similar degree distribution, which legitimizes their equivalent distributions of the interactions. On the other hand, in the benchmark datasets, there is a clear difference between the distributions, since the negative interactions severely deviate from the power law curve.

5. Conclusion

In this paper an unbiased independent dataset and two innovative DL architectures were established. At the same time, the reliability of the innovative FCNN architecture was authenticated, which motivates its growth into a standard architecture, not disregarding the need for further investigation.

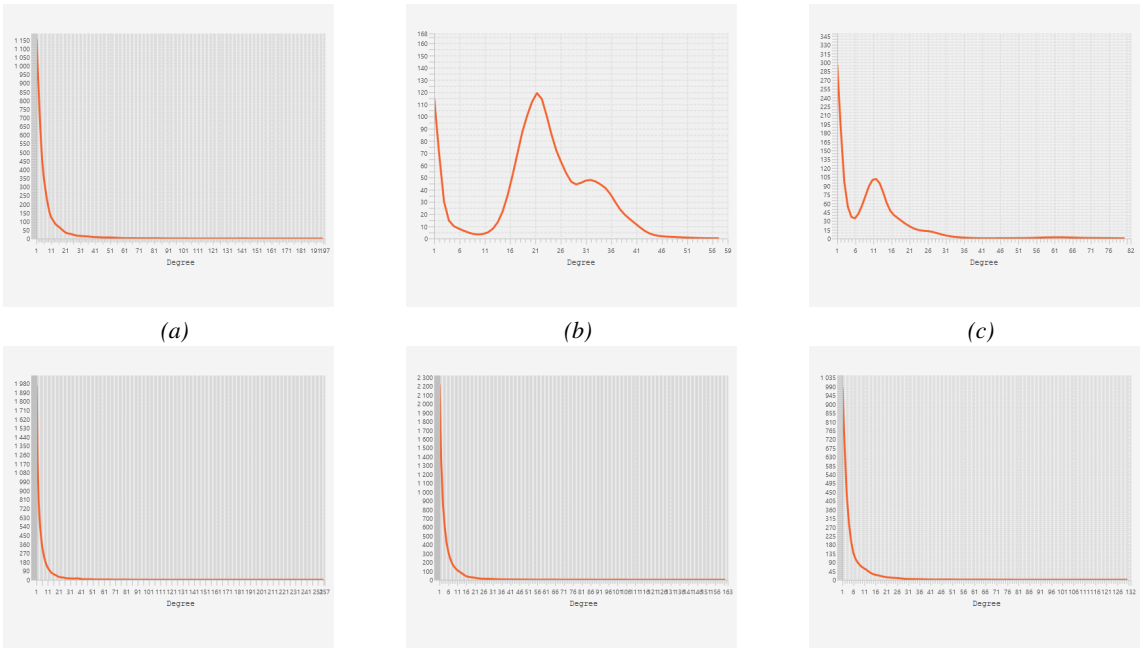
On the other hand, we also concluded that the dataset is a valuable element of a DL experiment, as the data strongly influences the performance. A model is only as good as the insights it can extract from the dataset fed to it, in the end the final word will come from the quality of the dataset. The exclusive analyze of the results of several published papers can incorrectly lead us to believe that the large-scale PPIs prediction problem is well addressed and close to be considered a solved problem. However, with the work developed in this paper, the viability of several state of the art PPI prediction algorithms built upon biased datasets is questioned.

References

[1] Xiuquan Du, Shiwei Sun, Changlin Hu, Yu Yao, Yuanting Yan, and Yanping Zhang. Deepppi: Boosting prediction of protein–protein interactions with deep neural networks. *Journal of Chemical Information and Modeling*, 57(6):1499–1510, 2017.

[2] Tanlin Sun, Bo Zhou, Luhua Lai, and Jianfeng Pei. Sequence-based prediction of protein protein interaction using a deep-learning algorithm. *BMC Bioinformatics*, 18(1), 2017.

[3] C. Stark. Biogrid: a general repository for interaction datasets. *Nucleic Acids Research*, 34(90001), Jan 2006.



(a) *(b)* *(c)*

(d) *(e)* *(f)*

Figure 2: Degree distributions, (a) and (d) are, respectively, the negative and the positive interactions of the unbiased dataset, (b) and (e) also the negative and positive interactions of Pan et al. dataset and ,finally, (c) and (f) are the negative interactions and positive interactions of Du et al. dataset.

[4] Asa Ben-Hur and William Noble. Choosing negative examples for the prediction of protein-protein interactions. *BMC Bioinformatics*, 7(Suppl 1), 2006.

[5] Xiao-Yong Pan, Ya-Nan Zhang, and Hong-Bin Shen. Large-scale prediction of human proteinprotein interactions from amino acid sequence based on latent topic features. *Journal of Proteome Research*, 9(10): 4992–5001, 2010.

[6] Hakime Öztürk, Arzucan Özgür, and Elif Ozkirimli. Deepdta: deep drug–target binding affinity prediction. *Bioinformatics*, 34(17):i821–i829, Jan 2018.

[7] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, 2014.

[8] Yuan-Miao Gui, Ru-Jing Wang, Xue Wang, and Yuan-Yuan Wei. Using deep neural networks to improve the performance of protein–protein interactions prediction. *International Journal of Pattern Recognition and Artificial Intelligence*, page 2052012, 2020.

[9] Xue Wang, Yuejin Wu, Rujing Wang, Yuanyuan Wei, and Yuanmiao Gui. A novel matrix of sequence descriptors for predicting protein-protein interactions from amino acid sequences. *Plos One*, 14(6), Jul 2019.

[10] Lei Wang, Hai-Feng Wang, San-Rong Liu, Xin Yan, and Ke-Jian Song. Predicting protein-protein interactions from matrix-based protein sequence using convolution neural network and feature-selective rotation forest. *Scientific Reports*, 9(1), Aug 2019.

[11] Long Zhang, Guoxian Yu, Dawen Xia, and Jun Wang. Protein–protein interactions prediction based on ensemble deep neural networks. *Neurocomputing*, 324: 10–19, 2019.

[12] Zhu-Hong You, Keith C. C. Chan, and Pengwei Hu. Predicting protein-protein interactions from primary protein sequences using a novel multi-scale local feature representation scheme and the random forest. *Plos One*, 10(5), Jun 2015.

[13] Yu Zhen Zhou, Yun Gao, and Ying Ying Zheng. Prediction of protein-protein interactions using local description of amino acid sequence. *Communications in Computer and Information Science Advances in Computer Science and Education Applications*, page 254–262, 2011.

Funding

This research has been funded by the Portuguese ResearchAgency FCT, through D4 - Deep Drug Discovery and De-ployment (CENTRO-01-0145-FEDER-029266).

Availability

The code and datasets are available at: https://github.com/larngroup/Agnostic_DL_Protein_Protein_Interaction.git