DeepAffinity: interpretable deep learning of compound–protein affinity through unified recurrent and convolutional neural networks

Mostafa Karimi, Di Wu, Zhangyang Wang, and Yang Shen

Department of Electrical and Computer Engineering, TEES–AgriLife Center for Bioinformatics and Genomic Systems Engineering and Department of Computer Science and Engineering, Texas A&M University, College Station, TX 77843, USA

Abstract

A semi-supervised deep learning model that unifies recurrent and convolutional neural networks [1] has been developed to exploit both unlabeled and labeled data, for jointly encoding molecular representations and predicting affinities. They are trained over generic protein-ligand data from BindingDB [19] and not fine-tuned for the kinase targets in the challenge.

Introduction

It is critically important to characterize compound–protein interaction (CPI) for drug discovery and development [2]. Considering the enormous chemical and proteomic spaces, computational prediction of CPIs facilitates experimental parallels and accelerates drug discovery. Indeed, computational prediction of CPI has made much progress recently, especially for repurposing and repositioning known drugs for previously unknown but desired new targets [3,4] and for anticipating compound side-effects or even toxicity due to interactions with off-targets or other drugs [5,6].

Computational methods roughly fall in two categories based on input data types: (protein) structure-based and sequence based methods. Structure-based methods can predict compound–protein affinity, i.e. how active or tight-binding a compound is to a protein; and their results are highly interpretable. They are often tackled through energy models [7] or machine learning [8,9]. Their heavy reliance on actual 3D structures of CPI presents a limitation for these methods. Sequence-based methods overcome the limited availability of structural data and the costly need of molecular docking. Rather, they exploit rich omics-scale data of protein sequences, compound sequences. Sequence-based CPI has been tackled through shallow models [10] or deep learning models [11,12] but their predictions lack interpretability.

To overcome limitations of current structure- and sequence-based CPI prediction methods, we have designed informative yet compact data representations that are structurally interpretable. We have also developed semi-supervised deep learning models that unify recurrent and convolutional neural networks, exploit labeled and unlabeled data, and use attention mechanisms for interpretability.

Methods

Data

We used data from three public datasets: all K_d labeled compound-protein binding data (17,819 samples) from BindingDB [19], compound data (500K samples for training and 500K samples for validation) in the SMILES format from STITCH [20] and protein amino-acid sequences from UniRef with 50% sequence identity and length less than or

equal to 1500 amino acids (120,000 samples for training and 50,525 for validation) [21] for training our unified RNN-CNN model.

Input formats

We developed a novel protein representation, Structural property sequence (SPS) by incorporating the predicted protein structural property such as secondary structure elements (SSEs), Solvent accessibility, physicochemical characteristics and length of each secondary structure elements (SSEs). For drug representation, we used SMILE [13] that are short ASCII strings to represent compound chemical structures based on bonds and rings between atoms.

Deep learning methods

First, we encoded compound SMILES or protein SPS into representations, by unsupervised deep learning from unlabeled data from STITCH and UniRef. Specifically, we used a recurrent neural network (RNN) model, seq2seq [14] that has seen much success in natural language processing and was recently applied to embedding compound SMILES strings into fingerprints [15]. We choose gated recurrent unit (GRU) [16] with attention mechanism [17] as our seq2seq model.

Next, with compound and protein representations learned from the above unsupervised learning, we solve the regression problem of compound–protein affinity prediction using supervised learning. For either proteins or compounds, we append a CNN after the RNN (encoders and attention models only) that we just trained. The CNN model consists of a one-dimensional (1D) convolution layer followed by a max-pooling layer. The outputs of the two CNNs (one for proteins and the other for compounds) are concatenated and fed into two more fully connected layers. The entire RNN-CNN pipeline is trained from end to end [18], with the pre-trained RNNs serving as warm initializations, for improved performance over two-step training. More details about how the final models are derived are included in the next subsection.

Lastly, we have also introduced protein and compound attention models in supervised learning to both improve predictive performances and enable model interpretability at the level of letters (SSEs in proteins and atoms in compounds). In the supervised model we just have the encoder and its attention α_t on each letter t for a given string x (protein or compound). And the output of the attention model, A, will be the input to the subsequent 1D-CNN model. Suppose that the length of protein encoder is T and (s_1,..,s_t,..., s_T) are the output of protein encoder and similarly the length of compound encoder is D and $$(m_{1},..,m_{_d},..., m_D) $$ are the output of compound encoder. We parametrize the attention model of unified model with matrix $$$U_a$$$ and the vector $$$v_a$$$. Then, The attention model is formulated as:

$$e_t = v_a tanh(W_a s_t) \quad \forall t = 1, \cdots, T \text{ where: } \alpha_t = \frac{exp(e_t)}{\sum_k exp(e_k)} \quad \forall t = 1 \cdots T \text{ and } A = \sum_t \alpha_t s_t.$$

The attention weights (scores) α_t suggest the importance of the t^{th} "letter" (secondary structure element in proteins and atom or connectivity in compounds) and thus predict the binding sites relevant to the predicted binding affinity.

Models submitted

We give more details about the training process for final models as follows. We trained three unified RNN-CNN models with different neurons (300,100), (400,200), and (600,300) at their fully connected layers. For each of these unified RNN-CNN model, we at first pre-trained the RNN encoder part from the encoder part of our seq2seq model and fixed the encoder parts. We trained the rest of the architecture with Adam optimizer [22] with an initial learning rate of 0.001 for 100 epochs. Later, we jointly trained all the architecture with Adam optimizer with an initial learning rate of 0.0001 for another 100 epochs. Finally, motivated from ensemble methods, we consider the last 10 epochs of each model as a predictor. Finally, we take an average of all 30 predictors to calculate the final prediction. Our docker image and src directory provides the 3 unified models with 10 checkpoints (epochs) each.

Conclusion

We have developed accurate and interpretable deep learning models for predicting compound–protein affinity using only compound identities and protein sequences. By taking advantage of massive unlabeled compound and protein data besides labeled data in semi-supervised learning, we have jointly trained unified RNN-CNN models from end to end for learning context- and task-specific protein/compound representations and predicting compound–protein affinity. Given the novel representations with better interpretability, we have included attention mechanism in the unified RNN-CNN models to quantify how much each part of proteins, compounds, or their pairs are focused while the models are making the specific prediction for each compound–protein pair. Noting that our models submitted were trained over generic data, improvements can be made by tailoring and tuning the models for kinase targets.

How to run our image

```
docker run \-\-privileged=true -it --rm -v \$\{PWD\}/io:/input -v $\{PWD\}/io:/output
docker.synapse.org/syn17051692/deepaffinity
```

References

[1] Mostafa Karimi, Di Wu, Zhangyang Wang, Yang Shen, DeepAffinity: interpretable deep learning of compound–protein affinity through unified recurrent and convolutional neural networks, *Bioinformatics*, , btz111, https://doi.org/10.1093/bioinformatics/btz111 [2] Santos R.*et~al*. (2017) A comprehensive map of molecular drug targets. *Nat. Rev. Drug Discov.*, 16, 19–34. [3] Keiser M.J .*et~al*. (2009) Predicting new molecular targets for known drugs. *Nature*, 462, 175. [4] Power A. *et~al*. (2014) Genomics-enabled drug repositioning and repurposing: insights from an IOM Roundtable activity. *JAMA*, 311, 2063–2064. [5] Chang R.L .*et~al*. (2010) Drug off-target effects predicted using structural analysis in the context of a metabolic network model. *PLoS Comput. Biol.*, 6, e1000938. [6] Mayr A. *et~al*. (2016) Deeptox: toxicity prediction using deep learning. *Front. Environ. Sci.*, 3, 80. [7] Gilson M.K., Zhou H.-X.(2007) Calculation of protein–ligand binding affinities. *Annu. Rev. Biophys. Biomol. Struct.*, 36, 21–42. [8] Wallach I.*et~al*. (2015) Atomnet: a deep convolutional neural network for bioactivity prediction in structure-based drug discovery. arXiv Preprint arXiv: 1510.02855. [9] Gomes J.*et~al*. (2017) Atomic convolutional networks for predicting protein–ligand binding affinity. arXiv Preprint arXiv: 1703.10603. [10] Shi Y.*et~al*. (2013) Protein–chemical interaction prediction via kernelized sparse learning svm. In: *Pacific Symposium on Biocomputing*, pp. 41–52. [11] Tian K.*et~al*. (2016) Boosting compound–protein interaction prediction by deep learning. *Methods*, 110, 64–72. [12] Wan F., Zeng J. (2016) Deep learning with feature embedding for compound–protein interaction prediction. bioRxiv, 086033. [13] Weininger D. (1988) Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. J. *Chem. Inf. Comput. Sci.*, 28, 31–36. [14] Sutskever I.*et~al*. (2014) Sequence to sequence learning with neural networks. In: *Advances in Neural Information Processing Systems*, pp. 3104–3112. [15] Xu Z.*et~al*. (2017). Seq2seq fingerprint: an unsupervised deep molecular embedding for drug discovery. In: Proceedings of the 8th ACM *International Conference on Bioinformatics, Computational Biology, and Health Informatics*. ACM, pp. 285–294. [16] Cho K.*et~al*. (2014) On the properties of neural machine translation: encoder–decoder approaches. In: Proceedings of SSST-8, Eighth Workshop on *Syntax, Semantics and Structure in Statistical Translation*, Doha, Qatar, pp. 103–111. [17] Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. "Neural machine translation by jointly learning to align and translate." arXiv preprint arXiv:1409.0473 (2014). [18] Wang Z.*et~al*. (2016b) Studying very low resolution recognition using deep networks. In: Proceedings of the IEEE Conference on *Computer Vision and Pattern Recognition*, pp. 4792–4800. [19] Liu T.*et~al*. (2006) Bindingdb: a web-accessible database of experimentally determined protein–ligand binding affinities. *Nucleic Acids Res.*, 35, D198–D201. [20] Kuhn M.et al. (2007) Stitch: interaction networks of chemicals and proteins. *Nucleic Acids Res.*, 36, D684–D688. [21] Suzek B.E.*et~al*. (2015) Uniref clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*, 31, 926–932. [22] Kingma, Diederik P., and Jimmy Ba. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014).

Authors Statement

All conceived the learning schemes; MK implemented all the deep learning models and trained them; DW managed data cleaning, and made the docker; YS, MK and DW wrote the wiki.