DeepAffinity: Interpretable Deep Learning of Compound-Protein Affinity through Unified Recurrent and Convolutional Neural Networks

MOSTAFA KARIMI1,3, DI WU1, ZHANGYANG WANG2, YANG SHEN1,3

1Department of Electrical and Computer Engineering, 2Department of Computer Science and Engineering, 3TEES-AgriLife Center for Bioinformatics and Genomic Systems Engineering, Texas A&M University, USA.

Abstract

- High-throughput drug discovery through deep learning.
- Novel representation of structurally-annotated protein sequences.
- We present a semi-supervised deep learning model that unifies recurrent and convolutional neural networks to exploit both unlabeled and labeled data.
- Transfer learning for new protein classes with few labeled data.
- Embedded attention mechanism to gain interpretability.
- Our models outperform conventional options in achieving relative error in IC50 within 5 to 10-fold.

Availability: https://github.com/Shen-Lab/DeepAffinity

Methods

Overall scheme

Data representation
- Compound: SMILES strings
- Protein: We developed Structural property sequence (SPS) based on predicted secondary structure elements (SSEs), solvent accessibility, physicochemical characteristics and lengths of SSEs.

Semi-supervised deep learning model
- Unsupervised learning: Seq2seq autoencoder models with attention mechanism to exploit abundant unlabeled data.
- Supervised learning: Unified recurrent and convolutional neural networks with attention mechanism are jointly trained starting with pre-trained encoder part of seq2seq
- Interpretability through the embedded attention mechanism
- Deep transfer learning

Result

Novel Representations v.s. Baseline Pfam/Fingerprints

SPS representation saves 40% training time and 20% memory while achieving the similar or better performances over test set and lowered RMSE for generalization sets

Shallow Models v.s. Deep Models

Unified RNN-CNN models outperform random forest and separate RNN-CNN models. Averaging ensembles of models lower RMSE by reducing the variance of model.

Deep transfer learning for new classes of protein targets

Deep transfer learning models increasingly improved the predictive performance compared to the original deep learning models, given increasing amount of labeled data. Even few labeled data is enough for significant improvement.

Predicting target selectivity of drugs

Protein-tyrosine phosphatase (PTP) family:

- Random forest using baseline representations cannot tell target specificity within the PTP family as the proteins’ Pfam descriptions are almost indistinguishable.
- Using novel representations, random forest correctly predicted PTP1B selectivity for compounds 1 and 3 but not compound 2, whereas unified RNN-CNN models correctly did so for all three compounds.

Interpretability

How do the compound-protein pairs interact?

Compared to randomly ranking the SSEs, our approach can enrich binding site prediction by 1.6~2.0 fold for the three CPIs.

Human factor Xa–DX9065a interaction:

The binding site was correctly predicted with a high rank 2. And the SSE ranked first, a false positive, was its immediate neighbor in sequence.

How are targets selectively interacted?

- Position 192 has been identified as the source of specificity: it is a charge-neutral polar glutamine (Gln192) in Xa but a negatively-charged glutamate (Glu192) in thrombin.
- The ground-truth segment (black) was ranked the 4th among 50 segments.

References


Acknowledgements

This project is in part supported by the National Institute of General Medical Sciences of the National Institutes of Health (R35GM124952 to YS) and the Defense Advanced Research Projects Agency (FA8750-16-2-0027 to ZW). Most of the computing time is provided by the Texas A&M High Performance Research Computing.