Predicting compound-kinase binding interactions using convolutional neural networks by combining SMILES and aligned ATP binding sites

AmsterdamUMC-KU-team Team: Georgi K. Kanev[1], Albert J. Kooistra[2], Bart A. Westerman[1]

[1] Department of Neurosurgery, Cancer Center Amsterdam CCA, De Boelelaan 1117, 1081 HZ Amsterdam, The Netherlands [2] Department of Drug Design and Pharmacology, University of Copenhagen, Universitetsparken 2, 2100 Copenhagen, Denmark

Introduction

Convolutional neural networks (CNNs) represent the current state-of-the-art algorithms in image and video recognition. [1] In the recent years, CNNs were applied to predict bioactivity,[2-3,9] learn molecular fingerprints,[4] detect chemical motifs,[5] and predict properties of small molecules[6]. In this work we used SMILES (Simplified Molecular Input Line Entry System) and sequences of aligned ATP binding sites (85 amino acids) of protein kinases downloaded from the KLIFS database containing[7-8] as input for the CNNs to predict the compound-kinase binding interactions. SMILES are widely used for encoding molecular structures and represent compounds in the form of a string over a fixed set of characters, describing all the atoms and structure of small molecules including chirality, bonds, aromaticity and more. The KLIFS (Kinase-Ligand Interaction Fingerprints and Structures)[7-8] database systematically aligns and process all current human and mouse protein kinase structures, focusing on the interactions of ligands in the binding site of protein kinases, assessment of binding pockets, kinase motifs and overall kinase and ligand properties. The representation of the compounds (SMILES string) and protein kinases (aligned ATP binding sites) in the form of a 2D matrix, allows CNNs to identify import motifs and map them to compound-kinase binding interactions.

Methods

Prediction

The data from ChEMBL (v24.1), DrugTargetCommons, IUPHAR/BPS Guide to pharmacology, and literature was integrated and curated. The Kd, Ki and IC50 measurements in combination with a protein kinase were filtered out and used for training of the CNN. The integrated data set comprised of 298,595 compound-kinase measurements, 439 unique kinases and 101,189 compounds. The SMILES of the compounds and the sequences of the ATP binding sites (downloaded from KLIFS) were one-hot-encoded and used as input for the 2D convolutional layers of CNN (Figure 1). The CNN comprised of a single 2D convolutional and 2D max pooling layer for the SMILES (shape=33,156,1) and single 2D convolutional and 2D max pooling layer for the sequences (shape=21,85,1). The convolutional layers used 64 filters. After both max pooling layers, dropout (0.5) was applied. The output of the dropouts was given to dense layers, each with 256 nodes. In addition to these layers, 2 dense layers were used - one received ECFP-4 fingerprints as input and the other received kinase family as input (one-hot encoded). By complementing the one-hot encoding of SMILES with a chemical fingerprint such as the ECFP-4, the chemical information can probably be better encoded into features for the CNN. The output of all described layers was concatenated and given to a dense layer with 32 nodes. A dropout of 0.4 was applied and the output was given to the output layer.

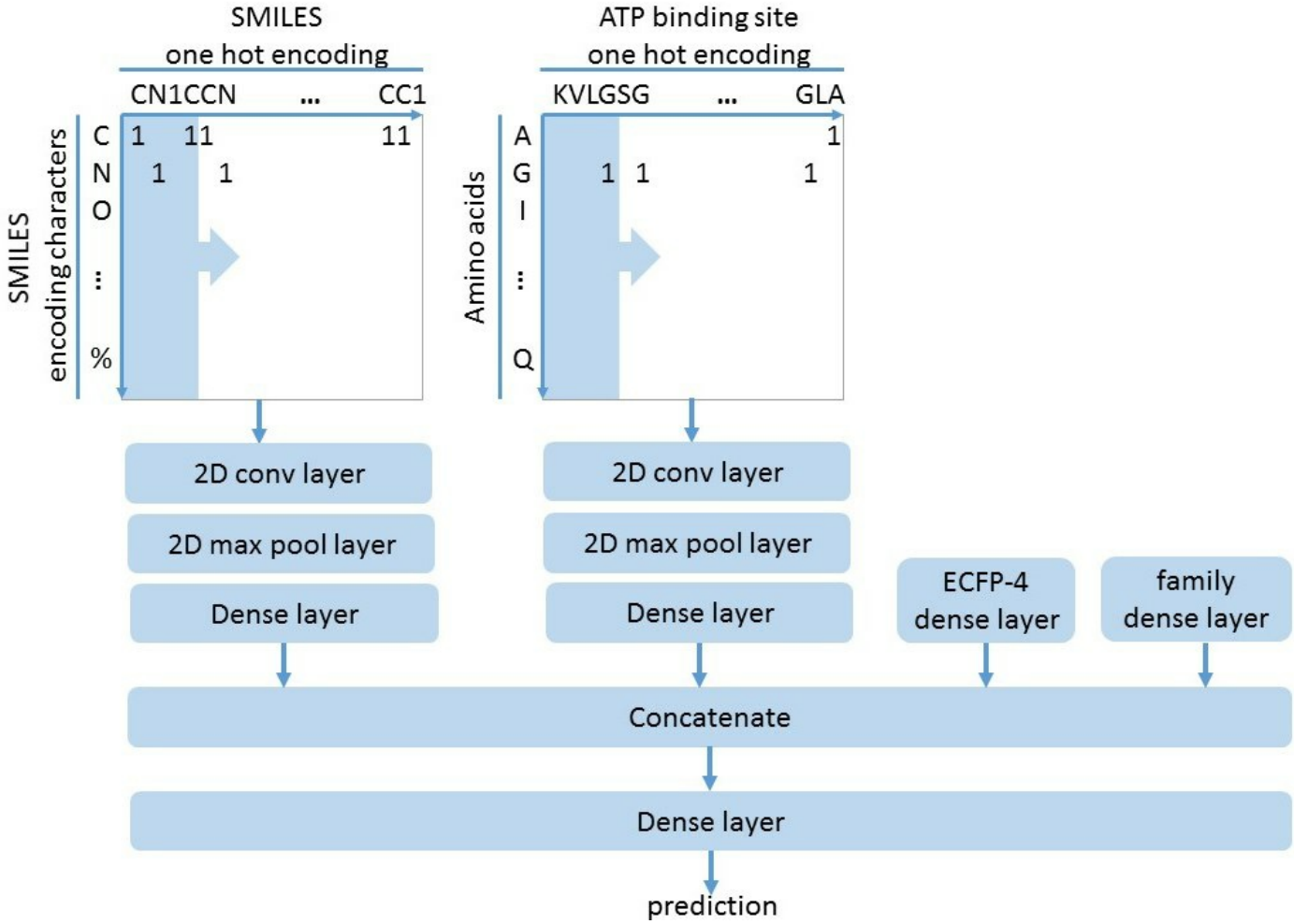

Figure 1. The convolutional neural network used to predict the compound-kinase binding interactions.

Refinement

After applying the model the predicted pKd values were further refined using literature data from both DTC/ChEMBL and the KiEO (Kinase Experiments Omnibus (http://tanlab.ucdenver.edu/KIEO/KIEOv1.0/). This simplistic refinement approach was applied as follows for each kinase-compound pair:

1. If one or more fully characterized (p)Kd/(p)Ki/(p)IC50 values were available from literature the median value was used instead of the prediction.
2. If step 1 was not applied , then all (p)Kd/(p)Ki/(p)IC50 values for highly similar compounds for the same kinase target (Tanimoto score 0.6 using the Morgan fingerprint RDKit) were collected and the minimum, maximum and average were calculated. If no data for similar compounds was available, then the following step was applied. If the difference between the maximum and the minimum value was smaller than 0.1, the minimum and maximum values were changed to the average value - 0.5 and + 0.5, respectively to account for potential larger variations in the data. Subsequently, if the predicted value was outside the range of the current minimum and maximum value, the predicted value changed to the average value.
3. If none of the previous steps were applied and a minimum expected pKd value was available (from the single concentration KINOMEscan for the PKIS2 dataset [10]), then the predicted value was increased with 0.5 if the predicted value was below the minimum expected pKd value.
4. If none of the previous steps were applied and the predicted pKd value was higher than 6 the prediction was scaled up: predpKD + (predpKD-6)*0.5. This step performed, as we noticed that the predictions for which literature data was available (step 1) the predicted value was overall lower than the literature values.

Finally, to correct for outliers values above > 9.5 were scaled down to 9.5. As no values below 5 were present, not lower limit was applied.

The literature data applied in steps 1-3 is available here: https://www.synapse.org/#!Synapse:syn18635344

Conclusion

Here we show that by using an unbiased approach to train a CNN network with 2D convolutional layers, we are able to predict the bioactivity of compounds with a RMSE of 1.125 and a rounded average AUC of 0.658. Despite this, the low spearman correlation (0.259) indicates that the model was probably overfitted and requires further optimization and validation.

References

[1] LeCun et al. "Deep learning", Nature volume 521, pages 436–444 (2015) [2] Wallach et al. "AtomNet: A Deep Convolutional Neural Network for Bioactivity Prediction in Structure-based Drug Discovery", arXiv preprint arXiv:1510.02855 (2015) [3] Jiménez et al. "KDEEP: Protein–Ligand Absolute Binding Affinity Prediction via 3D-Convolutional Neural Networks", Journal of chemical information and modeling 58.2 (2018): 287-296 [4] Duvenaud et al. "Convolutional Networks on Graphs for Learning Molecular Fingerprints", Advances in Neural Information Processing Systems , 2015, pp. 2224–2232 [5] Hirohara et al. "Convolutional neural network based on SMILES representation of compounds for detecting chemical motif", BMC Bioinformatics. 2018; 19: 526 [6] Goh et al. "Using Rule-Based Labels for Weak Supervised Learning: A ChemNet for Transferable Chemical Property Prediction", Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. ACM, 2018 [7] van Linden et al. "KLIFS: A Knowledge-Based Structural Database To Navigate Kinase–Ligand Interaction Space", Journal of Medicinal Chemistry, 2014, 57 (2), pp 249–277 [8] Kooistra et al. "KLIFS: a structural kinase-ligand interaction database", Nucleic Acids Research 44.D1 (2015): D365-D371 [9] Stepniewska-Dziubinska et al. "Development and evaluation of a deep learning model for protein-ligand binding affinity prediction", Bioinformatics. 2018;34(21):3666-74. [10] Drewry et al. "Progress towards a public chemogenomic set for protein kinases and a call for contributions", PloS one. 2017; 12(8):p.e0181585.

Authors Contribution Statement

GKK, data integration, data analysis and development models; AJK, data integration, data analysis and prediction refinement; BAW, project initiation and supervision