Using proteochemometrics for bioactivity prediction: A deep learning approach

Rutwik D, Sidhartha A

January 18 2019

1 Introduction

In the last few years, a lot of bioactivity data has become publicly available. However, we lack the methods when it comes to interpret and extract all information from it. Proteochemometrics modelling (PCM) is a computational method to model the bioactivity of multiple related proteins/targets simultaneously. PCM makes use of the additional knowledge of structural similarity between compounds to predict bioactivity. In the past PCM has worked well with ML techniques like SVM, RF etc. Recently, in 2017, in a paper published by Lenselink et al. it was demonstrated that using a combination of PCM with deep neural nets (DNNs) yielded the best results compared to other variations when it comes to predict bioactivity. With the fast progress of Deep learning, implementing the current state of the art with PCM is a promising area for the future of bioactivity prediction and drug design in general.

2 Justification

PCM needs some statistical/computational methods to be able to construct a single model that can output the output variable of the interaction (like affinity). For this, methods like SVM, Random forests, naïve Bayes has been used in the past. However, because of the abstract feature extraction ability of deep learning neural nets, its integration with PCM can be beneficial for extracting better features from the input data. A good PCM model can be used to:

1. Mine drug affinity databases to create multi target and multi species models
2. To combine phenotype data in predictive models
3. To to identify ligands (drug compounds) for 'orphan' target proteins Design personalized medicine for a certain strain of the virus or a certain cancer type using genotypic data

3 Objectives

Implementing the PCM-DNN with the current state of the art in AI, and comparing results to previous publications.

4 Methodology

The project is carried out in into following parts

- Extraction of data from the ChEMBL bioactivity benchmark dataset. Activities that are selected will be passed through certain criteria, based on source of the data, assay confidence, and the activity comments
- Redundancy in the data is removed, and for multiple values of the same data point, the mean is considered
- For ligand descriptors, the RDKit Morgan fingerprints are used. More physicochemical descriptors will also be added.(fingerprints are a string of bits that uniquely encode molecules.They are trained using the molecular data)
- For protein descriptors, physicochemical descriptors are added.
- The neural net is connected the following way:
  - An input layer consisting of the bits required to describe a fingerprint, is connected to 3 hidden layers of 4000, 2000 and 1000 ReLu nodes and an output layer with as many nodes as no. of target proteins is created
  - Based on the output values, they are classified as active(greater than or equal to 6.5) or inactive.
  - For weight updation, stochastic gradient descent is used with Nesterov momentum, which leads to fast convergence. The momentum rate is also modified after each epoch.
  - MCC and Bedroc will be used as the validation metrics to evaluate the accuracy of our testing results.

5 Project schedule

- January 2019
  Literature research
- February , March 2019
  Implementation of base algorithm and lit. research to apply SOTA algorithms
- April 2019
  Evaluation and testing of method
- May 2011
  Documentation