BITS :: Call for Abstracts 2019 - Oral communication

Type	Oral communication
Session	Machine Learning in Bioinformatics
Title	Virtual Screening of Molecular Fingerprint through Convolutional Neural Networks
All Authors	Mendolia I (1), Contino S (1), Perricone U (2), Pirrone R (1), Ardizzone E (1)

Affiliation

(1) Dipartimento di Ingegneria - Università degli Studi di Palermo
(2) Gruppo Drug Design, Fondazione Ri.MED, Palermo

Motivation

Deep Learning (DL) gained more and more impact on drug design because it allows a huge increase of the prediction accuracy in many stages of such a complex process. In this paper a Virtual Screening (VS) procedure based on Convolutional Neural Networks (CNN) is presented, that is aimed at classifying a set of candidate compounds as regards their biological activity on a particular target protein. The model has been trained on a dataset of active/inactive compounds with respect to the Cyclin-Dependent Kinase 1 (CDK1) a very important protein family, which is heavily involved in regulating the cell cycle. One qualifying point of the proposed approach is the use of molecular fingerprints as a suitable embedding for describing molecules; up to our knowledge there is no Deep Learning approach for VS that makes use of such descriptor. Several kinds of fingerprints are reported in the scientific literature to address different aspects of both the structure and the local properties of a molecule.

Methods

Modern approaches in Chemoinformatics have focused on the use of ML techniques applied to fingerprints instead of classical molecular descriptors. The reason is that fingerprints contain information on chemical groups and paths, and they give a more complete information about molecular complexity thus allowing a more robust comparison between two or more structures than descriptors. SMILE descriptors also convey information on molecular structures but their inherent string form needs the cycles to be cut, and the description of the same molecule is not unique thus a ``SMILE canonicalization'' is also needed.
Molecular fingerprints are generated analyzing each atom together with each neighborhood till 6 or 7 bonds away. Such a neighborhood is searched for a set of predefined molecular substructure, the so called patterns, and a string of 4 to 5 bits is generated for each pattern. From ChEMBL site has been downloaded activity data for CDK1 target, particularly, CHEMBL308, that are referred to single protein, and CHEMBL1907602, that is referred to protein complex. Data is chosen for the high number of samples (1830 samples). With RDKit KNIME extension, eight kind of molecular fingerprints have been generated (RDKit, Morgan, AtomPair, Torsion, FeatMorgan, Layered, Pattern, Avalon) with three different size (1024, 512, 256 bit).
For training-set 1432 samples are used and selected from two different datasets. Different rate of active and inactive samples are used for testset: 100%, 90%, 80%, 50%, 20%, 10%, 0%.
For greater clarity it is reported only global data referred to all the molecules present in the test-set (175 inactives and 100 actives molecules). The neural network architecture build for this study is a convolutional neural network (CNN) that use convolution in place of general matrix multiplication. This is an operation of two real-valued function which includes an input, a kernel and an output tipically called feature map. In deep learning, input and kernel is generally a matrix of features, also called tensor, that will used by learning algorithm, and will be stored separately.
In this study two CNN (mono-dimensional,1D and bi-dimensional, 2D) was build, to analyze different class of input: (i) vector of molecular fingerprints and (ii) matrix of molecular fingerprint.

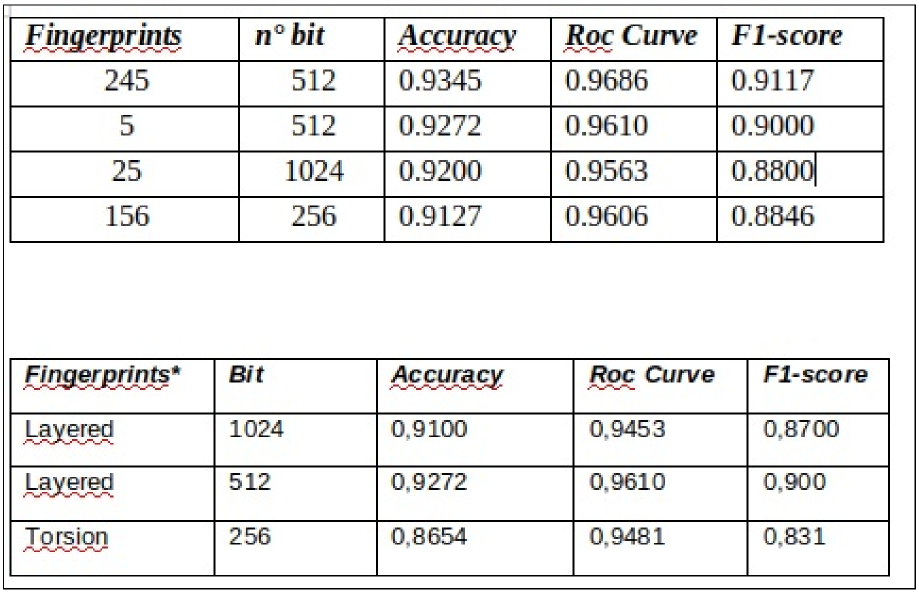
Results

The trial process consists of two different steps to evaluate the biological effect of samples molecule. In the first one, the study focused on the classification problem using mono-dimensional convolutional layers. One fingerprint at time has been provided as input data and that's allowing us to get good results.
In the second step, two-dimensional convolutional layers it has been used, and it has been thought to use all the fingerprints together arranged in a matrix, in such a way that, besides the direct information given by the individual fingerprints, the network could also acquire the indirect ones given by the combination of the different types of fingerprints.

Info

-

Figure



Fingerprints	n° bit	Accuracy	Roc Curve	F1-score
245	512	0.9345	0.9686	0.9117
5	512	0.9272	0.9610	0.9000
25	1024	0.9200	0.9563	0.8800
156	256	0.9127	0.9606	0.8846

Fingerprints*	Bit	Accuracy	Roc Curve	F1-score
Layered	1024	0,9100	0,9453	0,8700
Layered	512	0,9272	0,9610	0,900
Torsion	256	0,8654	0,9481	0,831

Availability	-

Corresponding Author

Name, Surname	Isabella, Mendolia
Email	isabella.mendolia@unipa.it
Submitted on	24.04.2019