BITS :: Call for Abstracts 2024 - Oral communication

Type	Oral communication
Session	Bioinformatics AI, Models and Tools
Title	Predicting Novel Molecular Interactions through Compound Signatures and Target Descriptors
All Authors	Eva Viesi(1,2,5), Giada De Simone(4), Ugo Perricone(4,5), Patrick Aloy(2,3), Rosalba Giugno(1,5)

Affiliation

(1) Department of Computer Science, University of Verona, Verona, Italy
(2) Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and
Technology, Barcelona, Catalonia, Spain
(3) Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Catalonia, Spain
(4) Molecular Informatics Unit, Fondazione Ri.MED, Palermo, Italy
(5) NBFC, National Biodiversity Future Center, Palermo, Italy

Motivation

The characterization of small compounds both from a chemical and biological perspective may help to gain a more detailed insight into their behaviour and effect on protein targets, elucidating their impact on complex systems and disorders. Furthermore, the identification of new compound-target interactions (CTIs) through machine learning-based predictive methods can address key challenges associated with experimental in vivo and in vitro approaches, mainly related to time constraints and resource limitations (Zhao et al., 2022). Finally, the investigation of small air pollutant molecules is crucial to understand, and possibly counteract, their toxicological effects on human health and the environment. To this end, the proposed methodology calculates and integrates bioactivity signatures of pollutant compounds with protein sequence features to train machine learning models for predicting novel CTIs.

Methods

From PubChem BioAssay (Wang et al., 2017) we collected protein targets associated with 1830 air pollutants provided by the Environmental Protection Agency (EPA). Based on a wide range of calculated molecular properties, we generated four new Signaturizer models (Duran-Frigola et al., 2020, Bertoni et al., 2021) to derive pollutants' bioactivity signatures and we used the iFeature package (Chen et al., 2018) to compute sequence descriptors of the selected targets. We combined molecule signatures and target features into a single vector representation to capture interactivity between multiple ligands and proteins. Due to class imbalance, we employed a One-Class Support Vector Machine (OCSVM) classifier to select negative samples (Zheng et al., 2019). We trained four machine learning classifiers and evaluated their performances via cross-validation and external validation.

Results

We computed the cross-validation results on five different generated datasets fed to the Logistic Regression (LR), K-Nearest Neighbors (KNN), Random Forest (RF), and Multi-Layer Perceptron (MLP) classifiers in terms of area under the ROC curve, average precision, recall, precision, f1, and balanced accuracy. We obtained that the LR was the best fitting estimator for our combined features. We then compared the performance on a randomly sampled dataset and on the instances selected by the OCSVM, observing that the latter leads to significantly improved results. As a last step, we demonstrated the generalization capability of our final models on external CTIs and we found that molecule features derived from APDB data (Viesi et al., 2023), our resource of small air-polluting compounds, show higher stability in recall score than the other datasets.

Info

References:
Zhao, L. et al. Computational and Structural Biotechnology Journal 20 (2022)
Wang, Y. et al. Nucleic acids research 45.D1 (2017)
Duran-Frigola, M. et al. Nature Biotechnology 38.9 (2020)
Bertoni, M. et al. Nature Communications 12.1 (2021)
Chen, Z. et al. Bioinformatics 34.14 (2018)
Zheng, Y. et al. BMC bioinformatics 20 (2019)
Viesi, E. et al. Database 2023 (2023)

filename	-

Figure

-

Availability	-

Dissemination Material

Social

https://www.linkedin.com/in/eva-viesi

Summary

-

Corresponding Author	
Name, Surname	Eva, Viesi
Email	eva.viesi@univr.it
Submitted on	24.04.2024