A. A. Nikitina[1]
A. A. Orlov[1,2]
D. I. Osolodkin[1,2]
V. A. Palyulin[1]
N. S. Zefirov[1]

CHEMBL ANTIVIRAL DATA ANALYSIS

[1] Department of Chemistry, Lomonosov Moscow State University, 1 bd. 3 Leninskie Gory, Moscow 119991, Russia;
[2] Chumakov Institute of Poliomyelitis and Viral Encephalitides, Chumakov FSC R&D IBP RAS, 8 bd. 1, Poselok Instituta Poliomielita, Poselenie Moskovsky, Moscow 108819, Russia;

nikitina_a@qsar.chem.msu.ru

Despite great progress made in the field of antiviral drug discovery during the past 50 years, only several human viral infections are manageable with vaccination or specific treatment. Two main challenges exist for modern antiviral drug discovery: (1) emergence of resistant strains for well-studied viruses and (2) absence of approved therapeutics for numerous viral species, such as Ebola virus and Zika virus. Thus, there is an urgent need for novel antiviral chemotypes.

A well-established approach for the lead compound identification is similarity search in large compound databases. Among numerous chemical databases the data sources containing bioactivity information are especially important to provide the basis for thorough structure-activity relationships analysis and reliable compound repurposing. The volume of information in such databases is enormously large and cannot be manually checked for inaccuracies in the data. Therefore, it is important to develop automated procedures of information mining and systematization. In this study we focused on public bioactivity database ChEMBL [1] as the basis for chemical space analysis for the compounds tested on antiviral activity.

The mySQL edition of ChEMBL version 20 was used. Advanced searching and data curation procedures for virus-related information were developed to cover as much bioactivity data as possible. Two approaches were exploited to find the relevant information: the first one was based on manual creation of lists of allowed values of *assays.assay_organism*, *target_dictionary.organism* fields, for the second one the substring dictionary containing official and historical viral species names, alternative species names and their abbreviations, and other words related to antiviral activity was applied for text query in *assays.description* field.

These search techniques allowed us to retrieve 1.5-fold more antiviral activity data points as compared to default ChEMBL Web interface search using Taxonomy Tree. Specific annotation procedure was used to map extracted data to ICTV viral taxonomy ranks wherever possible. Activity values and chemical structures were standardized, finally forming the subset of ChEMBL named ViralChEMBL. Physicochemical descriptors and scaffold distributions were analyzed. Chemical space of ViralChEMBL was visualized using self-organizing maps, PCA and RubberBend Scaling algorithms. Specific features of antiviral compounds were compared to the rest of ChEMBL and other chemical spaces present in public small molecules databases such as ZINC, DrugBank, etc.

1. Bento A.P. et al. *Nucleic Acids Res.*, 2014, **42**: 1083-1090.

Free academic software licenses were kindly provided by ChemAxon and OpenEye Scientific Software, Inc.

This study was supported by Russian Science Foundation (Grant Number 16-15-10307).