RDKit habitation in Pharmaceutical Bioinformatics

Telukunta, KK[a]; Zierep, P[a], Ntie-Kang, F[b]; Purschke, L[a]; Konrad, M[a]; Günther S[a,c]

[a]*Pharmaceutical Bioinformatics, Institute of Pharmaceutical Sciences, University of Freiburg, Germany*
[b]*Dept. of Pharmaceutical Chemistry, Martin-Luther University of Halle-Wittenberg, Germany*
[c]*Freiburg Institute for Advanced Studies (FRIAS, University of Freiburg, Germany)*

kiran.telukunta@pharmazie.uni-freiburg.de

Project 1: What is NANPDB ?

Northern African Natural Products Database (NANPDB)[1] is an online accessible database of Natural products (NPs) which are main sources of drugs and drug leads and play an important role in drug discovery by providing novel scaffolds[2,3]. In NANPDB we have collected natural products from the northern african region which is spreading over 9 million km^2[2]. Data was extracted from literature sources of the period 1962 to 2016. The data consists of ~4500 NPs (Table 1) isolated from plants, animals (e.g. corals), fungi, and bacterial sources (Fig. 1). The website provides browsable lists, compounds & species cards, diverse search options (e.g. by keywords, structural similarity or sub-structure). Similarity and sub-structural searches were implemented by using RDKit.

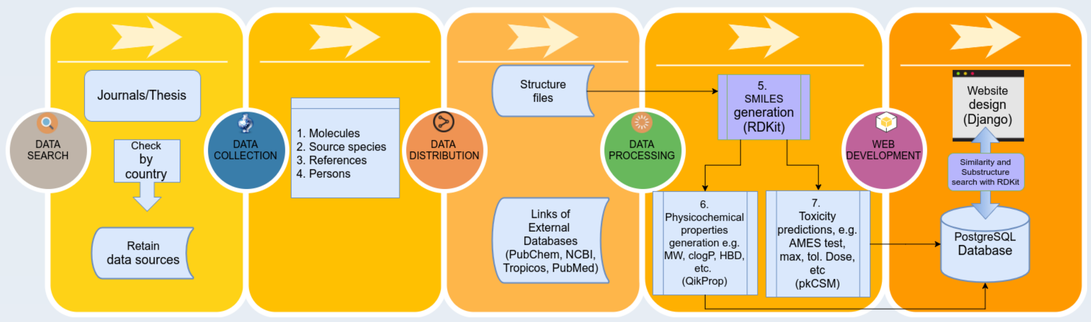

Figure 1: Workflow of data collection and implementation

www.african-compounds.org/nanpdb

Table 1: Current data of NANPDB

unique SMILES	plant families	source organisms	biological activities	modes of action
4469	146	617	98	37
unique PubChem IDs	**kingdoms** (majorly from plantae, animal, bacteria and fungi)	**cited references**	**PubMed references**	**compound classes**
2059	5	787	324	95

Project 2: SeMPI – a genome-based Secondary Metabolite Prediction and Identification web server

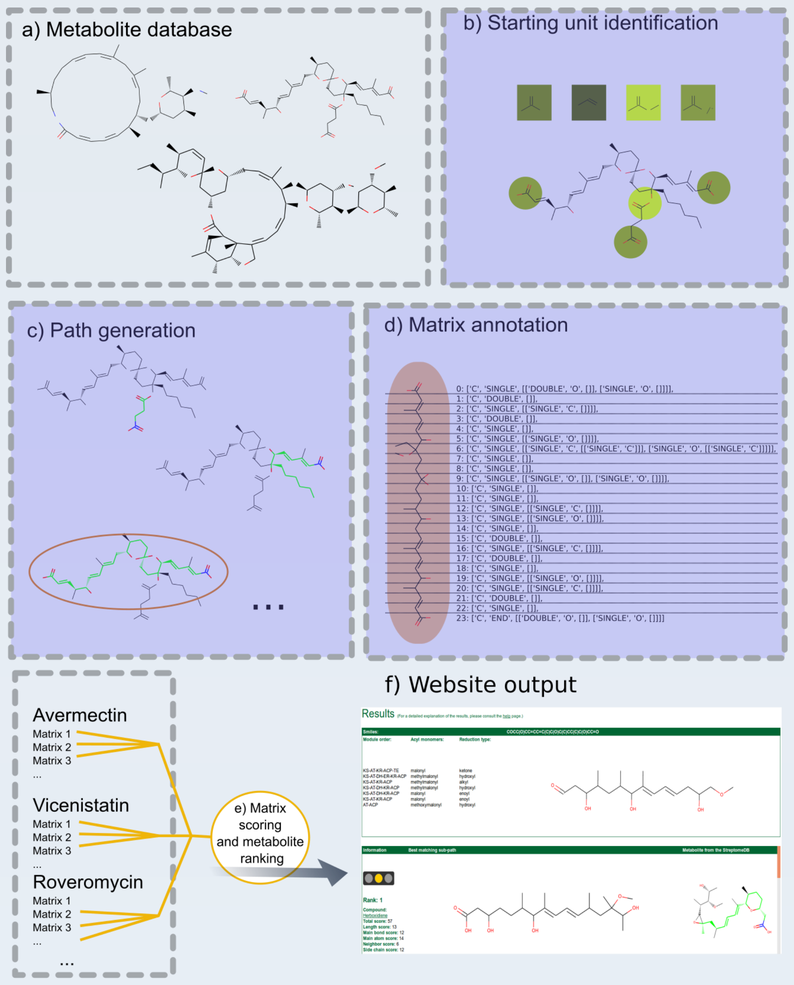

Figure 2: Database screening

SeMPI[4] is a webserver which predicts the structure of a secondary metabolite based on genomic information. Within the prediction pipeline a polyketide chain (Fig. 2a) is compared to 4000 diverse molecules annotated in StreptomeDB 2.0[5]. RDKit was used to identify conserved starting units (Fig. 2b) and identification of paths, representing the putative initial biosynthesized carbon chains of the metabolites (Fig. 2c). All paths are stored in a specific format, represented as matrix (Fig. 2d). Subsequently matrices are compared with predicted initial chains (Fig. 2e). Finally, SeMPI lists matching molecules from the database, based on the ranking of their best scoring paths (Fig. 2f).

sempi.pharmazeutische-bioinformatik.de

Project 3: Fragment based target analyse for fingerprint evaluation & development

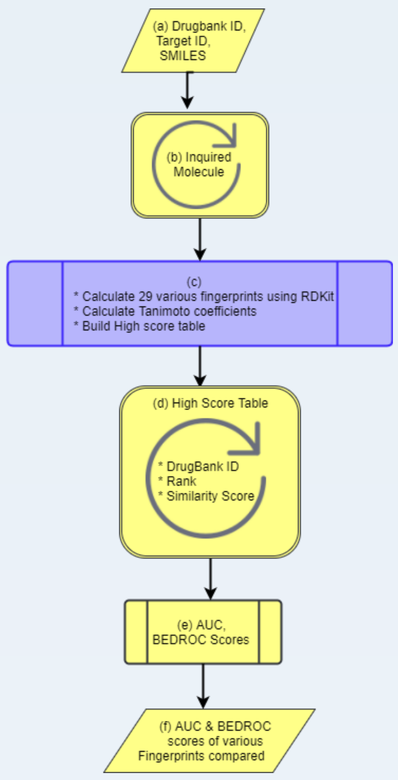

Figure 3: Scores generation workflow

In this project we evaluated available fingerprint (FP) techniques. Furthermore, we extended these fingerprints to develop MoKo fingerprints with additional bits derived from sub-structure patterns which were generated by the software Canvas (Schrödinger). We have taken the Drugbank[6] data consisting of drugs and their targets (Fig. 3-a) for benchmarking the quality by means of target prediction. Each molecule were compared to each other and ranked by their similarity index i.e., tanimoto coefficients (Fig. 3-b,c). The quality of a fingerprint was evaluated by comparing the ranking of molecules with the same target and is reflected by AUC or BEDROC scores (Fig. 4). The results indicated that various fingerprints can be slightly improved by including additional bits encoding for specific sub-structures of molecules.

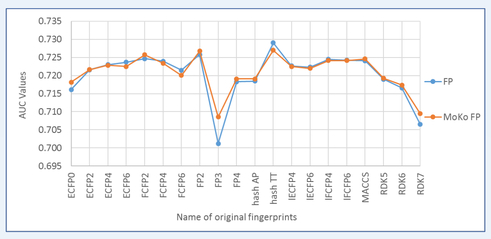

Figure 4a: Comparission of MoKo FP merged +FP with original FPs – AUC values

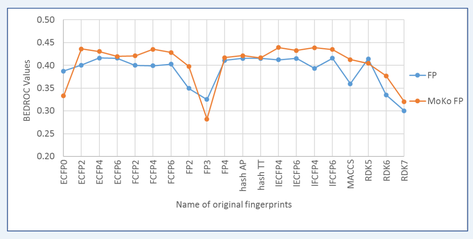

Figure 4b: Comparission of MoKo FP merged +FP with original FPs – BEDROC values

Project 4: Fragment based target prediction (FragPred)

Fragment based target prediction (FragPred) is a method to improve the prediction of target interactions using sub-structure and target information. Initially, compounds from ChEMBL with binding affinity lower than 20µM were extracted as learning dataset (Fig. 5-a,b). These molecules were then fragmented according to common reaction rules (RECAP rules[7] (Fig. 5-c). Subsequently fragments were clustered with RDKit applying different cluster algorithms (Butina, k-means) (Fig. 5-d). Clusters and compound targets were stored with support of the RDKit database cartridge (Fig. 5-f). Based on cluster-target pairs an enrichment analysis was performed to identify fragments that interact with specific targets more frequently than expected by random selection (Fig. 5-g,h). Based on a calcualted enrichtment score a ranking of the targets were done for each cluster. This ranking information was integrated in the target prediction method for queried small molecules (Fig. 5-i,j).

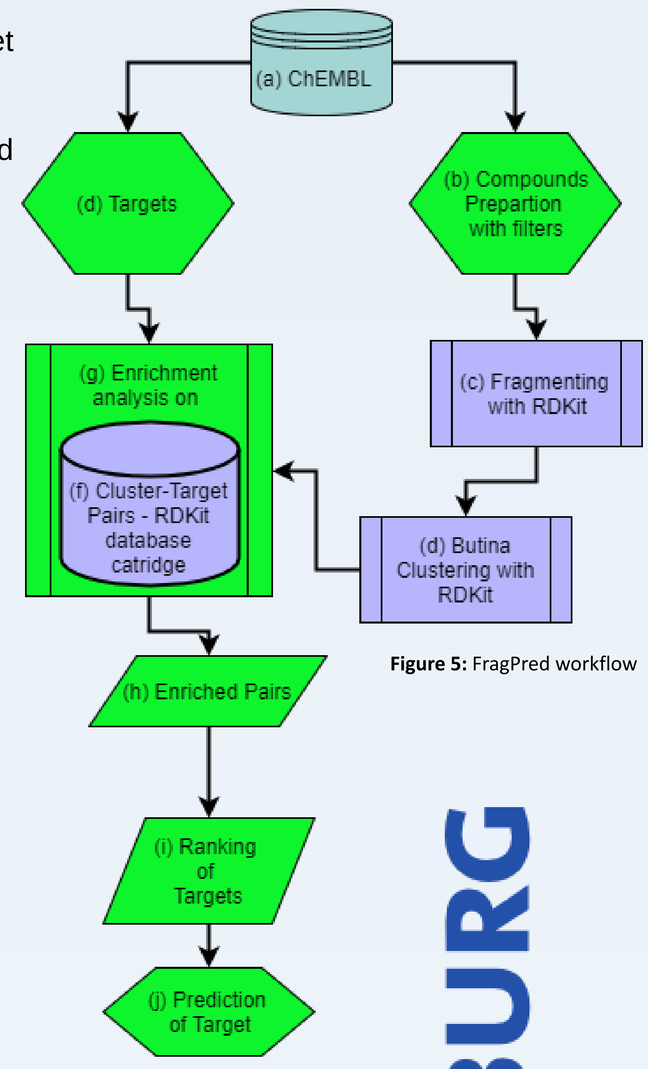

Figure 5: FragPred workflow

[1] Ntie-Kang, F, et al.: *J. Nat. Products* **2017**, PMID: 28641017
[2] Harvey, A. L, et al.: *J. Nat. Rev. Drug Discovery* **2015**, 14, 111-129
[3] Rodrigues, T, et al.: *G. Nat. Chem*. **2016**, 8, 531-541
[4] Zierep, P.F, et al.: *Nucleic Acids Res.* **2017**, 45, W64-W71
[5] Klementz, D., et al.: *Nucleic Acids Res.* **2016**, 44, D509-D514
[6] DrugBank version 5.0
[7] Lewell, XQ., et al.: *J. Cheminf.* **1998**, 38(3), 511-522

Pharmazeutische Bioinformatik
http://www.pharmazeutische-bioinformatik.de

Alexander von Humboldt Stiftung/Foundation

FRIAS FREIBURG INSTITUTE FOR ADVANCED STUDIES

UNI FREIBURG