Galaxy Drug Discovery Pipelines

Grüning BA*, Lucas X, Senger C,
Erxleben A, Flemming S, Günther S

Pharmaceutical Bioinformatics,
Institute of Pharmaceutical Sciences, University of Freiburg, Germany
**e-mail: bjoern.gruening@pharmazie.uni-freiburg.de*

Introduction

Motivation: A variety of software tools and components exists for compound analyses and drug discovery research, including tools for ligand- and structure-based *in silico* screenings. Certain processes have to be executed sequentially (pipelines) on sets containing up to several million compounds. Reformatting of data at tools' interfaces is frequently inevitable. For large projects, a collaboration management for researchers working on the same workflow is necessary. Workflows have to be repeatable and traceable and should be executable without programming or computer skills.

Results: We used Galaxy (http://galaxy.psu.edu), a well-established workflow management system, to integrate a toolbox for pharmaceutical researchers. It contains predefined software components allowing for the use of ready-to-use pipelines as well as the creation of new pipelines for drug discovery. The capabilities of the toolbox are demonstrated by a case study including a high-throughput (HT) docking experiment based on the results of one of the implemented workflows.

Methods and Results

Existing and newly developed tools were integrated into a local Galaxy workflow management system. A pipeline was set up enabling researchers to build focused libraries based on drugs' target protein sequences (Fig. 1).

1. DNA or protein sequences are used as input for the target protein.

2. Identification of similar protein sequences based on BLAST searches using
a) the NCBI RefSeq database for sequences
b) the PDB database for 3D structures for subsequent docking.

3. Several different biological and chemical relevant identifiers are assigned to each other (e.g. GI to UniProt accession number).

4. Proteins are searched in Reactome database pathways and all PubMed abstracts which are annotated with all PubChem compounds mentioned in the abstracts (CIL [1]), yielding compounds associated with the proteins.

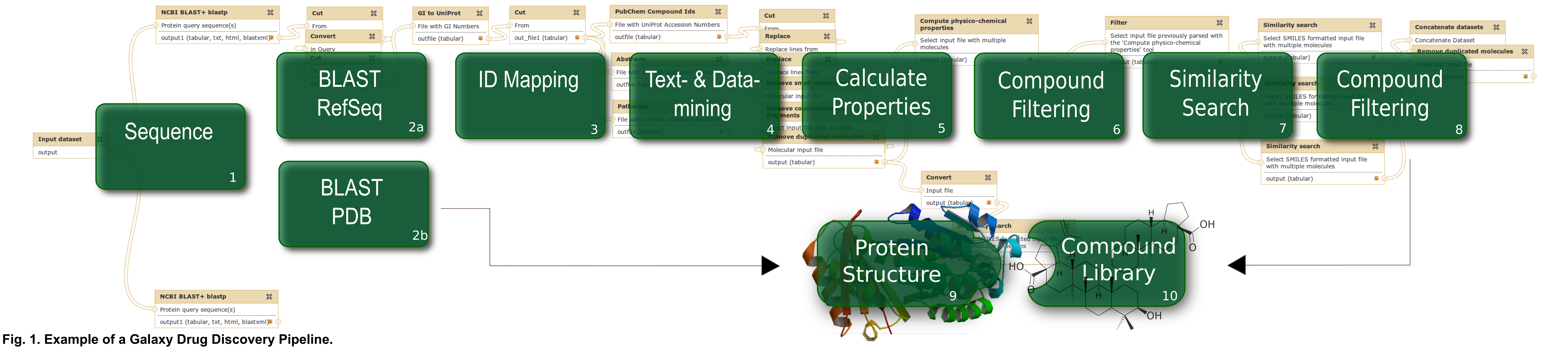

Fig. 1. Example of a Galaxy Drug Discovery Pipeline.

5. For each found compound
- physico-chemical properties are computed (e.g. number of hydrogen-bond donors and acceptors, number of rotatable bonds, octanol/water coefficient, and polar surfaces) and
- canonical representations (e.g. SMILES, InChI) are assigned.

6./8. Compounds can be filtered using sets of physico-chemical properties and pre-defined filtering rules like
- Lipinski's Rule of Five,
- lead-like properties [2],
- drug-like properties [3],
- fragment-like properties [4], and
- user-defined properties.

7. Based on attributes of the input molecules (e.g. fingerprints), similarity searches on several small molecule databases are performed in parallel. Resulting datasets are combined and duplicates are removed. Options for analyses and manual post-processing are provided.

9/10. Available protein structures obtained from a search in PDB and resulting compound libraries can be used for HT docking experiments.

Case Study

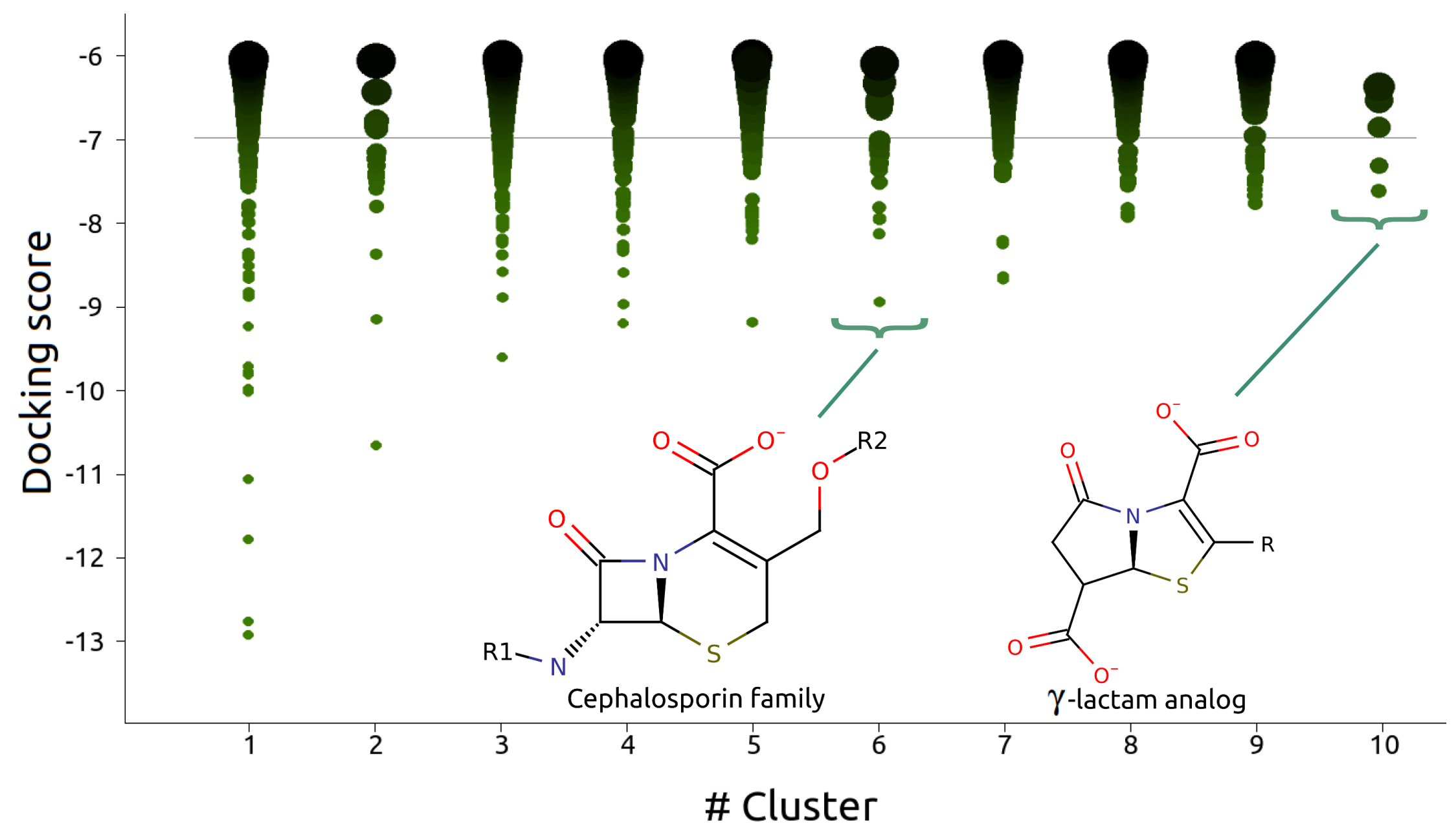

Fig. 2. Docking scores obtained for the different structural clusters identified among the best 500 ranked results. Several novel scaffolds with predicted high affinity for the receptor were identified.

β-lactamase enzymes account for the resistance to current life-saving β-lactam antibiotics (e.g. penicillins, cephalosporins). Thus, the development of antibiotic-accompanying inhibitors for those enzymes that enable the drugs to reach the therapeutic target is of major interest in antibiotic research [5]. The β-lactamase sequence of *E. cloacae*, an organism responsible for many nosocomial infections with high mortality [6,7], was used to analyse the capabilities of our workflow preceding HT docking on its crystallographic structure.
The β-lactamase sequence was used as the pipeline's input. The RefSeq database was queried for closely related homologs. Via text-mining, 480 compounds were identified which are mentioned in abstracts also referencing the homologs. 275 compounds complied with Lipinski's rules. Filtered compounds were used for similarity searches in several small molecule databases. 6,630 unique compounds were identified, prepared, and docked on the binding pocket of the target *E. cloacae* β-lactamase using Glide (Schrödinger Inc.).
After performing HT docking on the identified compounds and rescoring of the resulting poses, the best ranked 500 results were selected and clustered by structural similarity (Fig. 2).
Cephalosporins are the preferred substrate of class C β-lactamases [5]. A cephalosporin analog bicyclo scaffold containing a less constrained γ-lactam ring was identified. This ring has been recently proposed as a novel reversible β-lactamase inhibitor [8].
Compounds identified in the present HT docking approach will be further analysed and candidates will be selected for experimental validation to assess their *in vitro* activity.

Future Prospects and Availability

For the near future, the creation of new ready-to-use workflows is planned. Filtering tools based on statistical learning approaches will be implemented. Subsequently, their application to drug discovery tasks will be investigated. Additional software for protein-ligand docking will be included into the toolbox. The Galaxy Drug Discovery Pipelines toolbox is available via internet for third parties on request.

References
[1] Grüning BA, *et al.* (2010). Compounds In Literature (CIL): screening for compounds and relatives in PubMed. Bioinformatics. 271341-2.
[2] Teague SJ, *et al.* (1999). The Design of Leadlike Combinatorial Libraries. Angew Chem Int Ed Engl. 38:3743-3748.
[3] Lipinski CA, (2000). Drug-like properties and the causes of poor solubility and poor permeability. J Pharmacol Toxicol Methods. 44:235-49.
[4] Carr RA, *et al.* (2005). Fragment-based lead discovery: leads by design. Drug Discov Today. 10:987-92.
[5] Drawz SM and Bonomo RA (2010). Three decades of beta-lactamase inhibitors. Clin Microbiol Rev. 23:160-201.
[6] Maheshwari N and Shefler A (2009). *Enterobacter cloacae*: an "ICU bug" causing community acquired necrotizing meningo-encephalitis. Eur J Pediatr. 168:503-5.
[7] Juanjuan, *et al.* (2007). Retrospective analysis of bacteremia because of *Enterobacter cloacae* compared with *Escherichia coli* bacteremia. Int J Clin Pract. 61:583-8.
[8] Brown T, *et al.* (2010). Structural Basis for the Interaction of Lactivicins with Serine β-Lactamases. J Med Chem. 53:5890-4.

The working group of Pharmaceutical Bioinformatics at the Institute of Pharmaceutical Sciences develops algorithms and software for pharmaceutical research. Our fields of research include the modeling of molecular interactions, prediction of biological effects of molecules, and identification of potential new drug agents. The working group is part of the University of Freiburg's Research Group Program of the Excellence Initiative of the federal and state governments.