3D-e-Chem KNIME Nodes for Integrated Structural Cheminformatics Analyses and Computer-Aided Drug Discovery

Chris de Graaf[1], Márton Vass[1], Ross McGuire[2], Stefan Verhoeven[3]

[1]VU University Amsterdam, [2]Radboudumc, BioAxis Research,
[3]Netherlands eScience Center

17 March 2017

The workshop is set up to stimulate participants with varying degrees of experience in chemoinformatics to learn and apply the different structural chemoinformatics tools and workflows developed within the context of the 3D-e-Chem project. You will learn how to construct and apply simple cheminformatics workflows using the 3D-e-Chem KNIME nodes for the exploitation of G protein-coupled receptor and kinase data (two important target classes) to obtain useful information for drug discovery.

For the workshop you will need:

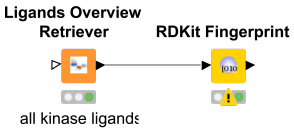
- KNIME 3.3.1 + all free extensions installed
- In KNIME go to Help > Install New Software…, click on **Available Software Sites** link, check the box of the **Stable Community Contributions** site, and click OK
- From the dropdown menu select the **Stable Community Contributions** site, open the **Cheminformatics** folder, check **3D-e-Chem KNIME nodes** and follow the installation instructions, then restart KNIME
- In a web browser go to https://github.com/3D-e-Chem/workflows and **download** the **ZIP**ped workflow bundle, then extract it to your own **knime-workspace** folder

1. Creating your first workflow to access kinase ligand data from KLIFS

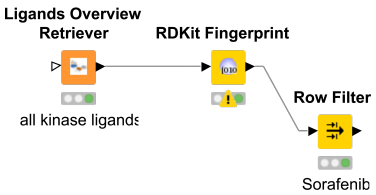
http://klifs.vu-compmedchem.nl/

- In the menu bar click on **File** and **New**…
- Click **Next**, then give the workflow a name, and click **Finish**
- Find the **KLIFS Ligand Overview Retriever** node in the **Node repository** and drop it to the workspace
- Right click on the node and click **Configure**… – you can optionally provide kinase IDs to the node but if you want a full overview, you don't need to change anything now
- Right click on the node and click **Execute**, wait until the status bar becomes green
- Right click on the node and open the **Out-Port table** – you can see the list of kinase ligands found in crystal structures including several different identifiers and chemical structures

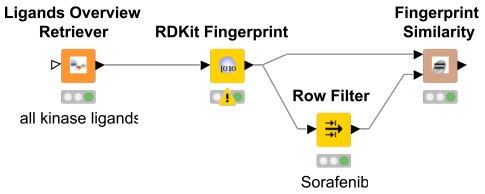
- Find the **RDKit Fingerprint** node in the Node repository, drop it to the workspace and connect the output port of the **KLIFS Ligand Overview Retriever** node to the input port of the **RDKit Fingerprint** node
- Right click on the node and click **Configure…** you can optionally select a different fingerprint type, but for this tutorial we will use the default **Morgan fingerprint**
- Execute the node and open the output table – a new column with fingerprints has been added to the table (no fingerprint for metal containing ligands can be calculated)



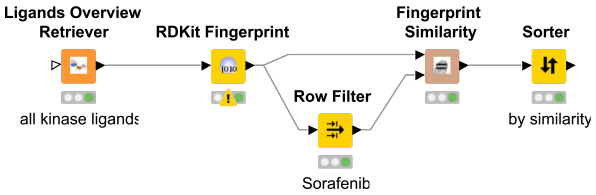
- Drop a **Row Filter** node on the workspace and connect it to the **RDKit Fingerprint** node
- In the configuration window select **PDB-code** as the **Column to test** and type "**BAX**" in the pattern field (this is the PDB code of Sorafenib), then execute and view the result



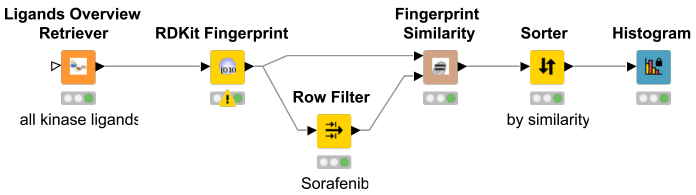
- Drop a **Fingerprint Similarity** node on the workspace and connect it to the **RDKit Fingerprint** and the **Row Filter** nodes
- In the configuration window you can leave everything at default. A chemical similarity coefficient (Tanimoto) column is added to the table

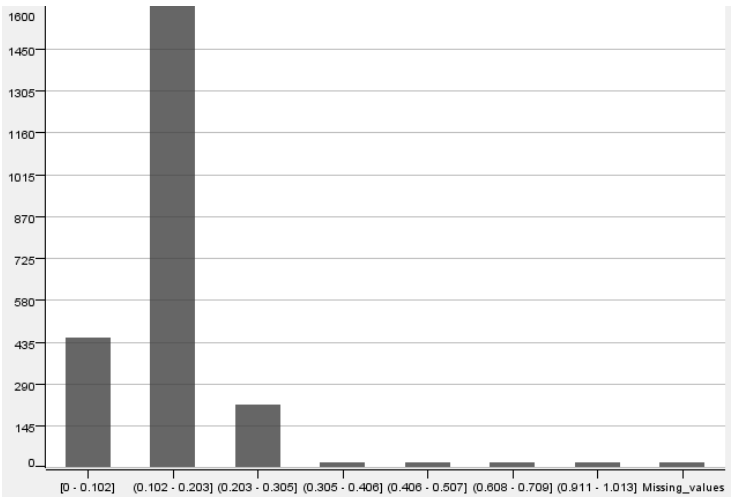


- Drop a **Sorter** node on the workspace and connect it to the **Fingerprint Similarity** node. In the configuration window select the **Tanimoto** column to sort by and check the **Descending** option. Inspect the resulting table and see what are the similar ligands found in kinase crystal structures to the reference Sorafenib



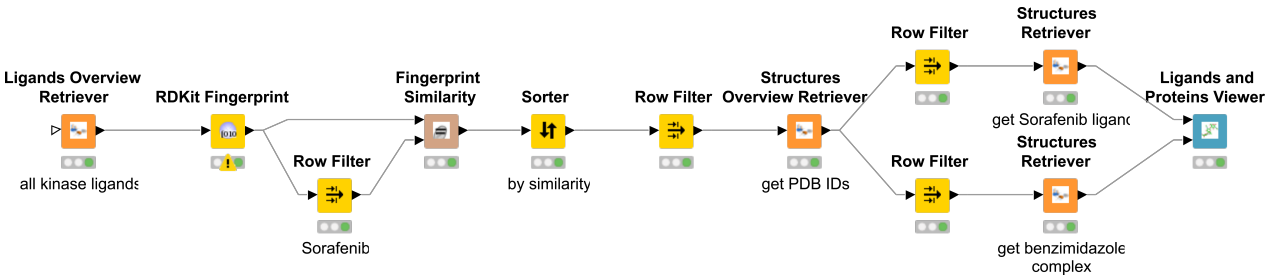
- Drop a **Histogram** node on the workspace and connect it to the **Sorter** node. Open the configuration window, the **Tanimoto** column is automatically selected for binning. Close the window, execute and inspect the results



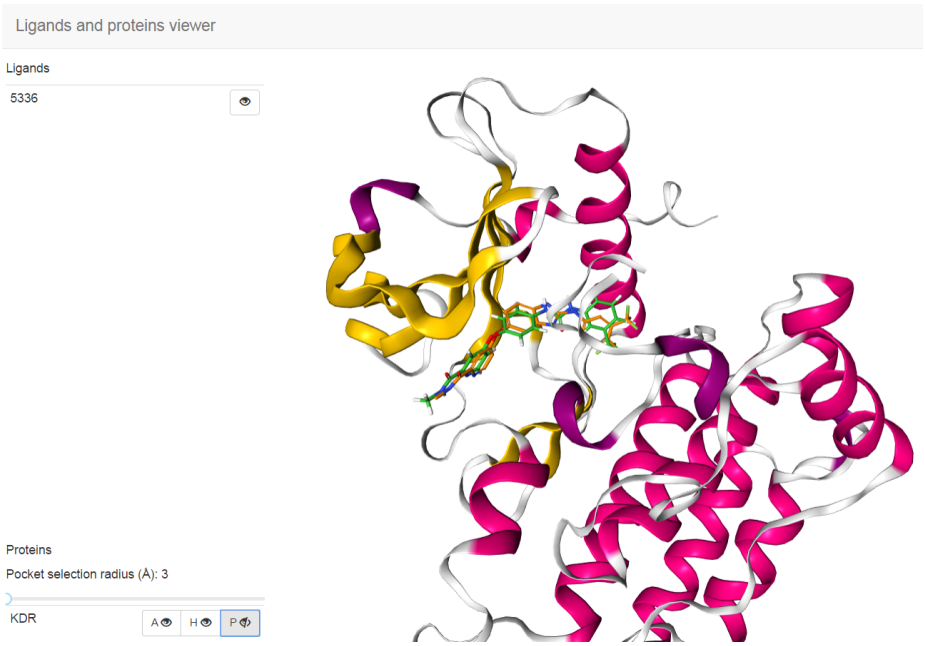


2. Viewing the 3D structures of ligand-protein complexes from KLIFS

- Extend the previous workflow after the **Sorter** node with the nodes shown in the following figure:

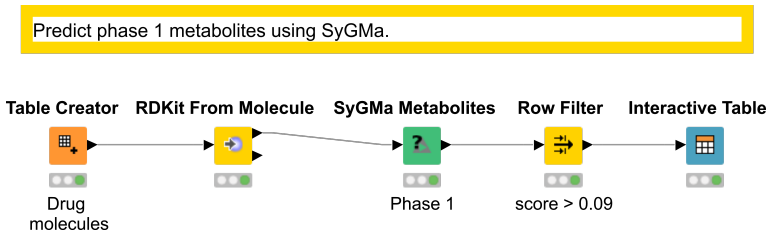


- For the configuration of the **Row Filter** node select **Include rows by number** and filter for the first two rows of the table
- For the configuration of the **KLIFS Structures Overview Retriever** node select the **Ligand ID** column and the **Ligand IDs** input type
- For the configuration of the top **Row Filter** node filter the **PDB** column for "**3wze**"
- For the configuration of the bottom **Row Filter** node filter the **PDB** column for "**2qu5**"
- For the configuration of the top **KLIFS Structures Retriever** node select **Ligand (MOL2)** as the structure type
- For the configuration of the bottom **KLIFS Structures Retriever** node select **Complex (PDB)** as the structure type
- Right click on the **Ligands and Proteins Viewer** node and click **Execute and Open Views** to open the 3D structures in your web browser (you can toggle the protein/ligands/binding pocket on and off using the buttons):



3. Prediction of small molecule metabolites using the SyGMa Metabolites node

- Open the **SyGMa-example** workflow from the **KNIME Explorer**, you should see the following workflow:



- The **Table Creator** node is used to input a list of molecules using their names and SMILES strings for the metabolite prediction, which are then converted to RDKit representation using the **RDKit From Molecule** node (this is required for the **SyGMa Metabolites** node and for the alignment of the metabolites with the parent structure)
- The **SyGMa Metabolites** node performs the specified number of Phase I and II metabolic transformation steps and returns the predicted metabolites with a score associated to them as of how likely a specific metabolite is to form (*ChemMedChem.* 2008, 3, 821.). It also returns the type of transformation the metabolite was generated by
- The **Row Filter** node is used to return only metabolites with a SyGMa score larger than 0.09 in the example.
- The **Interactive Table** node is used to visualize the metabolite structures. The comparison is facilitated by using the same orientation as the parent molecule
- Inspect the results for Sorafenib: the experimentally verified metabolic transformations of Sorafenib are demethylation, hydrolysis to carboxylic acid and N-oxidation, which can be found in the output of the **SyGMa Metabolites** node as the 1st, 2nd and 10th predicted metabolites

NOTE

To use the SyGMa node later, you will need to install the SyGMa Python package:

Install Miniconda for Python 2.7 from https://conda.io/miniconda.html

Open the Anaconda Prompt and execute the following commands:

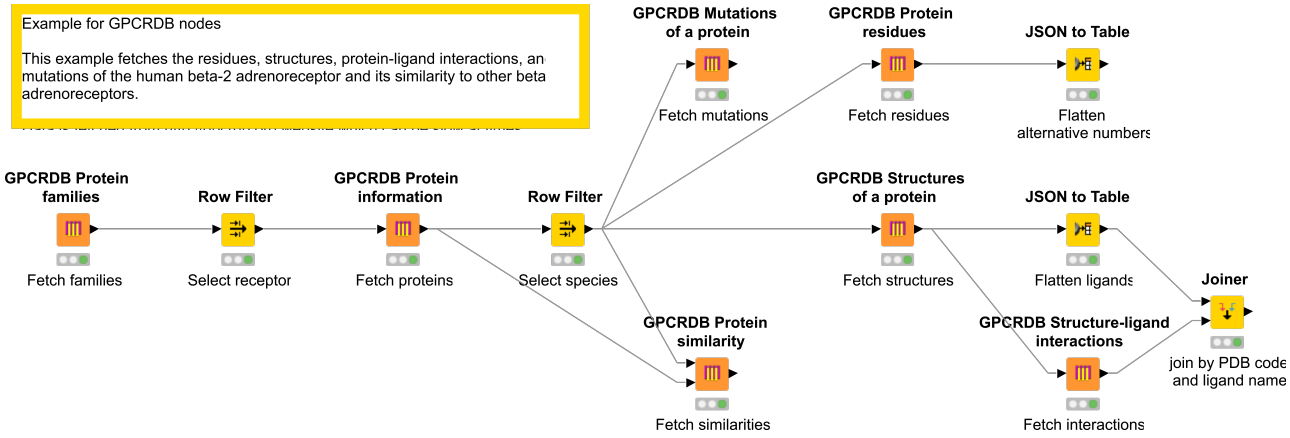
```
conda install -c 3d-e-Chem -c rdkit -c conda-forge pandas protobuf rdkit sygma
```

In KNIME go to File > Preferences > KNIME > Python and select your new Python executable (e.g. C:\ProgramData\Miniconda2\python.exe)

4. Obtaining G protein-coupled receptor data from GPCRdb

http://gpcrdb.org/

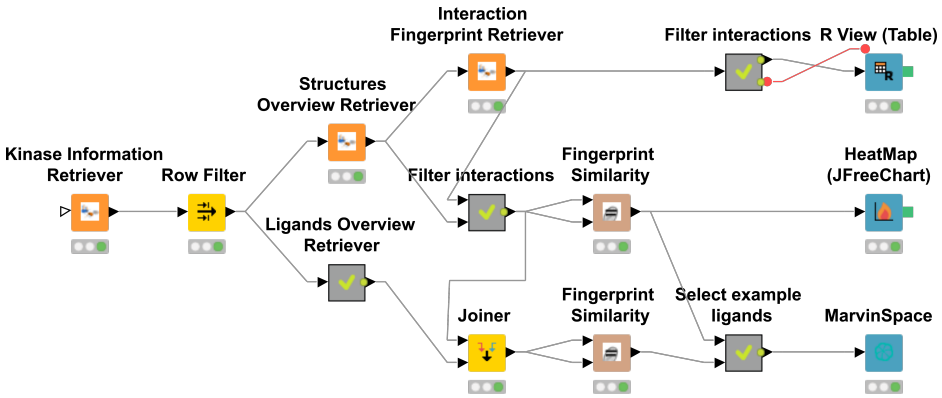
- Open the **GPCRDB_example_full** workflow from the **KNIME Explorer**, you should see the following workflow:



- The output of the **GPCRDB Protein families** node is a hierarchical list of GPCRs with their identifiers and names in the table. The **Slug** is a hierarchical identifier of GPCRs comprising four digits encoding the class, the ligand type, the subfamily and the subtype, e.g. 001_001_003_008 means class A, aminergic, adrenoceptors, β_2 adrenoceptor
- The **Row Filter** node is used to filter the table for the β_2 adrenoceptor
- The **GPCRDB Protein information** node returns a table with additional information about the selected receptors including the UniProt (http://www.uniprot.org/) entry names and accession codes as identifiers of the various GPCR subtypes and species variants
- The **Row Filter** node is used to filter the table for the human β_2 adrenoceptor
- The **GPCRDB Mutations of a protein** node returns a list of annotated mutations for the selected receptor(s) including sequence data, ligand and radioligand data, measurement type, and wild type and mutant activity (https://zenodo.org/record/58104)
- The **GPCRDB Protein residues** node returns a list of the amino acids of the selected receptor(s) including the UniProt numbering, protein segment information, the GPCRdb numbering and the class specific numbering schemes (*Trends. Pharmacol. Sci.* 2015, 36, 22.) as a JSON field, which can be extracted using the **JSON to Table** node
- The **GPCRDB Structures of a protein** node returns a list of experimental crystal structures of the selected receptor(s) including metadata, references and ligand data in a JSON field, which can be extracted using the **JSON to Table** node
- The **GPCRDB Structure-ligand interactions** node returns the observed protein-ligand interactions from the specified crystal structures (using the **PDB code** identifier). The information includes the ligand name, sequence data, and interaction type annotation
- Finally the **GPCRDB Protein similarity** node returns sequence identity and similarity percentages between the selected sets of receptors taking into account the protein segments specified in the configuration

5. Combining protein-ligand interaction and chemical similarity data from KLIFS

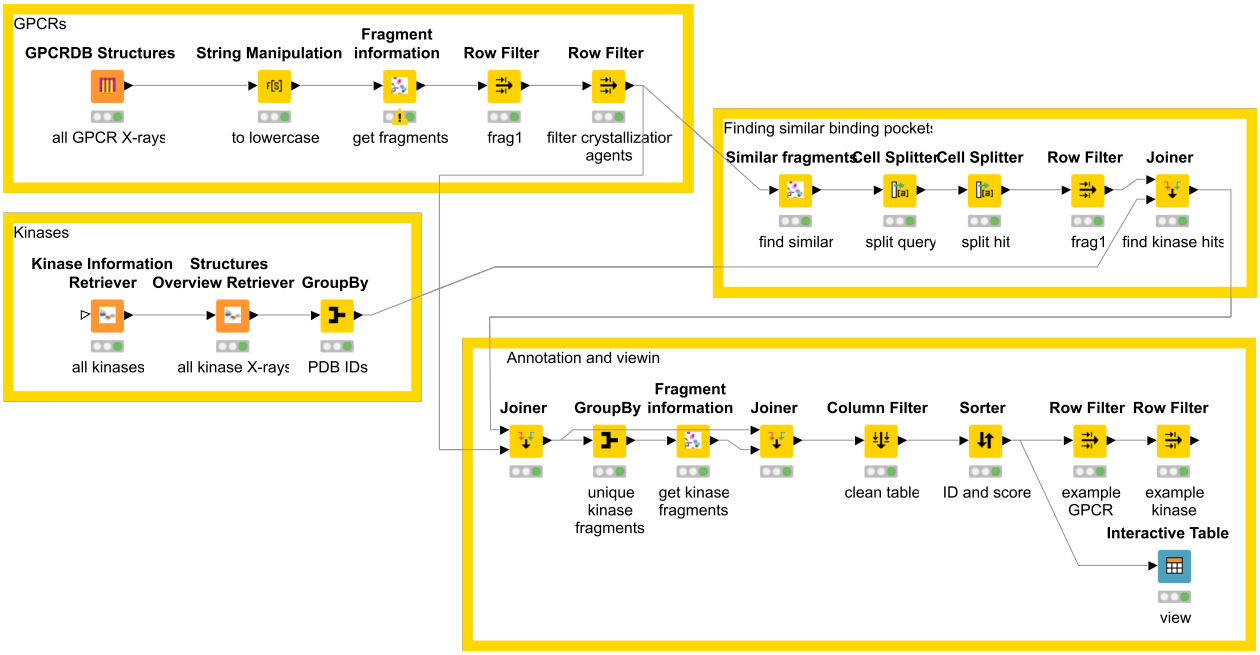
- Open the **KLIFS_example_workflow_small** workflow from the **KNIME Explorer**, you should see the following workflow:



- The **KLIFS Kinase Information Retriever** node is used to obtain information on all human kinases along with their various identifiers
- The **Row Filter** node is used to filter for the MAPK kinases
- The **KLIFS Structures Overview Retriever** node is used to obtain a list of all crystal structures of the selected MAPK kinases
- The **KLIFS Ligands Overview Retriever** metanode is used to obtain the co-crystallized ligands in the MAPK crystal structures and to calculate chemical fingerprints for them
- The **KLIFS Interaction Fingerprint Retriever** node is used to obtain the interaction fingerprints for the predefined pocket residues of the selected kinase structures
- The **Filter Interactions** metanodes are used to select/annotate structures in which the ligand forms hydrogen bond with the hinge residues of the kinase and the **R View** node is used to construct a histogram of the number of such structures
- The top **Fingerprint Similarity** node is used to perform an all vs. all comparison of the protein-ligand interaction fingerprints of the selected kinases with a specific hinge binding pattern (in which the ligand has an H-bond donor for residue hinge.46 and an H-bond acceptor for residue hinge.48)
- The **HeatMap** node is used to visualize the result of the comparison showing that overall IFP similarity is relatively low despite the shared hinge interaction pattern
- The bottom **Fingerprint Similarity** node is used to perform an all vs. all comparison of the ligand chemical fingerprints and the **Select example ligands** metanode is used to identify structures with a high IFP similarity (Tc > 0.75) but low (Tc < 0.3) structural similarity of the ligands and to obtain the 3D structures of these ligands in mol2 format
- Finally the **MarvinSpace** node is used to visualize an overlay of the selected ligands

6. Comparing kinase and GPCR binding pockets using the KRIPO nodes

- Open the **GPCR-kinase** workflow from the **KNIME Explorer**, you should see the following workflow:



- The top left branch of the workflow uses the **GPCRDB Structures** node to fetche all known GPCR X-ray structures and returns their PDB ID (http://www.rcsb.org)
- The **KRIPO Fragment Information** node is used to fetch all ligand fragments from these crystal structures
- **Row Filter** nodes are used to select only the full ligands for this example (frag1) and to filter out crystallization agents
- The bottom left branch of the workflow uses the **KLIFS Kinase Information Retriever** and **KLIFS Structures Overview Retriever** nodes to obtain a list of kinase X-ray structures and to return their PDB ID
- The **KRIPO Similar fragments** node is then used to find binding pockets from the PDB with a similar pharmacophore to those of the GPCR structures (modified Tanimoto coefficient > 0.5 and max. 1000 hots per query), for an explanation of the KRIPO method see *J. Chem. Inf. Model.* 2012, 52, 2031.
- **Joiner** nodes are used to select the kinase targets among the hits from the full PDB and to join the GPCR fragment information from the output of an earlier node to the results
- The **KRIPO Fragment Information** node is used to fetch all fragment structures for the kinase hits
- The table is cleaned up and sorted by the query fragment ID and the modified Tanimoto similarity score, and finally an experimentally verified example target pair is selected