

# Answer sheet

## #Day1: Programmatic Access to UniProt Database using KNIME

---

1. There are several alternative ways to process *collection* data types; we can either use the 'Split Collection Column' or the 'Ungroup' node to separate alternative names into single instances. Your task is to probe both nodes and figure out what is the difference and how the output table looks like. After getting familiar with the nodes, please fill in the box below:

ANSWER: The example nodes for splitting the collection column are, e.g., the 'Split Collection Column' or the 'UnGroup' node. The 'Split Collection Column' splits the collection column into separate columns based on the number of values. The output table has additional columns with the added alternative names, each in a single column. The 'UnGroup' node splits the collection column into rows. Creates one additional column with an alternative name, without the collection sign.

2. After execution of your KNIME workflow please fill in number of alternative names per UNIPROT ID into the table below:

Uniprot ID	Number of Alternative Names
P0C6X7	1
P0C6U8	1
P59594	2
Q9BYF1	3
O15393	1
P50052	1
A0A220F1P8	0

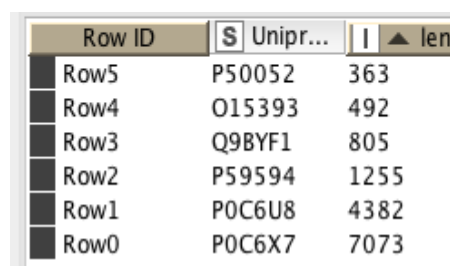
3. For retrieving the sequence length we will need to use a specific attribute of the <sequence> element. For further information look into the KNIME 'Help' section (*hint: you might use the '@' character to specify an attribute in your Xpath query*). Complete Xpath query for sequence length retrieval into the box:

ANSWER: `/dns:uniprot/dns:entry/dns:sequence/@length`

Filter the output table from the previous step to only keep those targets whose sequence is larger than 250 amino acids. You can use the 'RowFilter' node for this. Read the 'Help' section to figure out how to filter a KNIME table on basis of a number range. However, keep in mind that the sequence length retrieved via 'Xpath' node is included as 'string' data type in your table. Figure out how to convert string to number in KNIME and write down possible solutions into the box:

ANSWER: Conversion of a string to a number (integer) can be done by using the 'String To Number' node.

After filtering the table for the sequence length  $> 250$  amino acids, create a screenshot of your final table which includes the 'Uniprot ID' and 'Length' columns only (*hint: all redundant columns can be removed from the table by using the 'Column Filter' node*). Upload a screenshot as a separate file, alongside with this protocol.



Row ID	S Unipr...	I len
Row5	P50052	363
Row4	O15393	492
Row3	Q9BYF1	805
Row2	P59594	1255
Row1	P0C6U8	4382
Row0	P0C6X7	7073

## #Day2: Using Cross-References to Retrieve Structural Data from the Protein Data Bank (PDB)

---

1. In the example workflow we learned how to link UNIPROT entries to the available PDB structures. After retrieval of available PDB structures use appropriate nodes to calculate the number of unique PDB IDs per UNIPROT entry and fill in the table below:

Uniprot ID	Number of Unique PDB IDs
P0C6X7	100
P0C6U8	111
P59594	48
Q9BYF1	16
O15393	0
P50052	3
A0A220F1P8	0

2. After execution of the PDB nodes in your workflow, you should get a list of PDB IDs with their structural properties. Your task is to create a list of available experimental methods ('Experimental Method' column) and calculate the number of unique PDB IDs which were resolved by the respective methods.

Experimental Method	Number of Unique PDB IDs
Electron Microscopy	26
X-ray crystallography	207
Solution NMR	25
Theoretical Model	20

3. Which KNIME nodes would you use to calculate basic statistics, such as minimum, maximum, mean, and median values? Name at least two different nodes:

ANSWER: The 'Statistics' node, the 'GroupBy' node, the 'Spark statistics' node.

4. In case of the PDB structures resolved via X-Ray diffraction, use any available KNIME node to calculate the minimum, maximum, mean, and median values of crystal resolution [Å].

Statistic Value Type	Crystal resolution [Å]
Minimum	1.4
Maximum	3.4
Mean	2.3
Median	2.2

5. After retrieval of co-resolved ligands, look at different ligand properties. What is the minimum, maximum, and median molecular weight for those ligands?

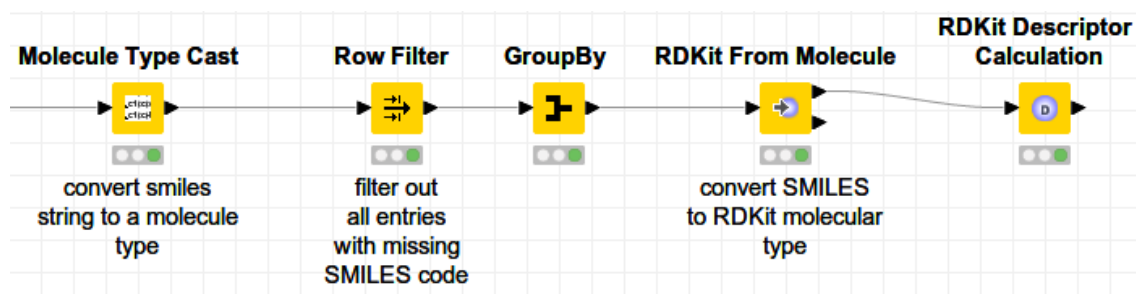
Statistic Value Type	Molecular Weight
Minimum	24.305
Maximum	772.406
Mean	330.894
Median	380.875

6. Calculate the number of unique ligands per protein target and fill in the table.

Uniprot ID	Number of Unique ligands
P0C6X7	41
P0C6U8	39
P59594	6
Q9BYF1	3

O15393	0
P50052	2
A0A220F1P8	0

7. Provide additional ligand properties by calculating RDKit descriptors for the ligands. First, convert the SMILES string format into a molecular type ('Molecule Type Cast' node) format. Now you can visually inspect the ligand structure in a table. As a next step, remove all rows with missing molecular type ('Row Filter' node) and keep only unique ligands ('GroupBy' node using ligand molecular type as a group). Afterwards, convert smiles columns with visualized structures into the RDKit format ('RDKit from Molecule'). Then, calculate RDKit descriptors by using the 'RDKit descriptor calculation' node. Below you can see the visual depiction of such a workflow:



Finally, use any KNIME node of your choice to calculate different statistical measures for the respective RDKit descriptors (table below):

	Rdkit descriptor					
Statistic Value Type	SlogP	SMR	TPSA	Number of Heteroatoms	Number of Rotatable Bonds	Number of Aromatic Rings
<b>Minimum</b>	-3.283	0	0	0	0	0
<b>Maximum</b>	6.723	180.807	407.410	18	30	7
<b>Mean</b>	1.138	85.161	95.540	6.5	6.95	1.323
<b>Median</b>	1.289	87.053	79.345	4.5	6	1

## #Day3: Integrative Data Mining of Ligand Bioactivity Data from ChEMBL and PubChem

---

1. Which number of *bioactivity data points* did we retrieve per single protein target?

ANSWER:

O15393: 6  
P0C6U8: 273  
P0C6X7: 320  
P50052: 1473  
Q9BYF: 376

2. Which number of *unique compounds* did we retrieve per single protein target?

ANSWER:

O15393: 6  
P0C6U8: 109  
P0C6X7: 196  
P50052: 694  
Q9BYF: 167

3. What different types of bioactivity endpoints (e.g., Km, Ki, IC50) did we retrieve? Write down a list of different bioactivities type and include the number of data points per given bioactivity type.

ANSWER:

Activity: 42  
IC50: 1872  
Ki: 530  
Km: 4

4. How many unique compounds do possess a bioactivity mismatch / a clash (activity label = 0.5)?

ANSWER: 328 compounds.

5. Check the compound overlap between PubChem and ChEMBL database (*hint: use an appropriate molecular structural format for checking the data overlap*).

ANSWER: 532 compounds.

## #Day4: Substructure Searches In DrugBank

---

### 4.2.1 Generation & Analysis of Murcko scaffolds

1. Have a look at the ligands for which Murcko scaffold generation failed. What could be the reason for this? (*hint: check the definition for Murcko scaffolds<sup>1</sup>*)

ANSWER: Those structures do not possess a ring system which is a definition of the Murcko Scaffold.

2. Which Murcko scaffold is the highest populated one? Provide the number of unique compounds which do possess this particular Murcko scaffold, as well as the canonical smiles of the Murcko scaffold (*hint: use the 'RDKit Canon SMILES' node to convert Murcko Scaffold structures into their corresponding canonical smiles form*)

ANSWER: c1ccc(-c2nn[nH]n2)c(-c2ccc(Cn3cnc(-n4cccc4)c3)cc2)c1 is the highest populated scaffold (30 entries).

### 4.2.2 Scaffold Clustering & Generating Maximum Common Substructures

1. How many distinct scaffold clusters do you get after hierarchical clustering?

ANSWER: 72 clusters.

2. Give the number of the highest populated scaffold cluster:

ANSWER: Cluster 67 (12 scaffolds).

3. Calculate the average number of Murcko scaffolds which are clustered in distinct clusters.

ANSWER: The average number of Murcko scaffolds/cluster ~ 2.472.

---

<sup>1</sup> BEMIS, Guy W.; MURCKO, Mark A. The properties of known drugs. 1. Molecular frameworks. *Journal of medicinal chemistry*, 1996, 39.15: 2887-2893.

### 4.2.3 Substructure Search In DrugBank

1. Calculate the number of identified compounds per target and fill in the table:

Uniprot ID	Number of identified compounds from Drugbank
P0C6X7	253
P0C6U8	440
Q9BYF1	2187
P50052	505

2. Is there a compound overlap between the protein targets? If yes, please write down the number of overlapping compounds per respective protein target:

ANSWER: Yes, there was an overlap between the target proteins:

S Unique concatenate(Uniprot ID)	I Unique count
P0C6U8, P0C6X7	230
P0C6U8, P0C6X7, P50052, Q9BYF1	6
P0C6U8, P0C6X7, Q9BYF1	4
P0C6U8, P50052, P0C6X7, Q9BYF1	2
P0C6U8, P50052, Q9BYF1	142
P0C6U8, P50052, Q9BYF1, P0C6X7	1
P50052, Q9BYF1	7
Q9BYF1	1668
Q9BYF1, P0C6U8, P0C6X7	10
Q9BYF1, P0C6U8, P50052	30
Q9BYF1, P50052	302
Q9BYF1, P50052, P0C6U8	15



#### 4.2.4 Analysis of Identified Compounds from DrugBank

1. Calculate RDkit descriptors for the identified compounds and keep only those which obey Lipinski Rule of Five.<sup>23</sup> Name a few examples of such compounds:

ANSWER: Compounds which obey the rule of five are, e.g., Phentermine, Exisulind, Irbesartan, Ethylmorphine.

2. Did you identify some known antiviral drugs? If yes, could you please provide some examples of such hits, including their DrugBank ID?

ANSWER: Compounds known as antivirals are. e.g., Nesbuvir (DB07238), Delavirdine (DB00705), or Atervirdine (DB12264).

3. Choose one of the antiviral drugs from the previous question and indicate its name in the box below. On which protein target(s) of interest has the chosen drug been predicted to be active? Do a literature search and try to find some information whether the chosen drug has already been suggested for COVID-19 treatment. Is the chosen drug currently tested in clinical trials to fight COVID-19? Provide some reference (e.g. reference to a paper or link to a website). Has the chosen drug been identified via different experimental/computational approaches? Is there something known about the interaction of the chosen drug with COVID-19 target(s)? Write down a short summary (text limit: a single A4 page)

---

<sup>2</sup>Lipinski CA . "Lead- and drug-like compounds: the rule-of-five revolution". Drug Discovery Today: Technologies. 1 (4): 337–341. doi:10.1016/j.ddtec.2004.11.007. PMID 24981612.

<sup>3</sup> LIPINSKI, CA., et al. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced drug delivery reviews*, 1997, 23.1-3: 3-25.