EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

MATHEMATISCH-
NATURWISSENSCHAFTLICHE
FAKULTÄT

User's Guide

ROCColoring

A KNIME-based version of the pROC-Chemotype protocol

Alexander Dörr[1]*

October 22, 2015

[1]Center for Bioinformatics Tuebingen (ZBIT)

*Corresponding author: alexander.doerr@uni-tuebingen.de

Contents

1 Prerequisites 2

2 Setting up the ROCColoring node 3

3 Overview of parameters 6

4 Views and output of the node 8

5 Final notes 10

1 Prerequisites

The following prerequisites are similar to the pROC-Chemotype Protocol and should be established to enable a proper function of the ROCColoring KNIME node:

1. **Original active set as a SDF file** (not docked), containing bioactivity information with column name(s): ”Ki_in_nM”, ”Kd_in_nM” and/or ”IC50_in_nM”. This file will be used for clustering and bioactivity annotations. DEKOIS 2.0 active sets automatically possess such information.

2. **Docked actives SDF file**, containing:
   - Bioactivity information with column name(s) as in (1.).
   - Docking information with a defined column name for the docking score.
   - Exactly the same molecule names as in the original active set.

3. **Docked decoys SDF file**, containing:
   - Docking information with a defined column name for the docking score.

4. **Installation of the KNIME workbench** with the KNIME chemistry extension.

5. **Downloaded ROCColoring** KNIME node from ”www.dekois.com”.

2 Setting up the ROCColoring node

The file `org.roccoloring_1.0.0.jar` has to be copied in the folder ”plugins” of the KNIME workbench. After starting KNIME, the node should appear at the bottom of node repository list from which it can be dragged on the workflow environment.

The original active set and the docked data sets are supplied to the first and only input port of the ROCColoring node via an SDF Reader from the KNIME chemistry extension (see Figure 2.1).

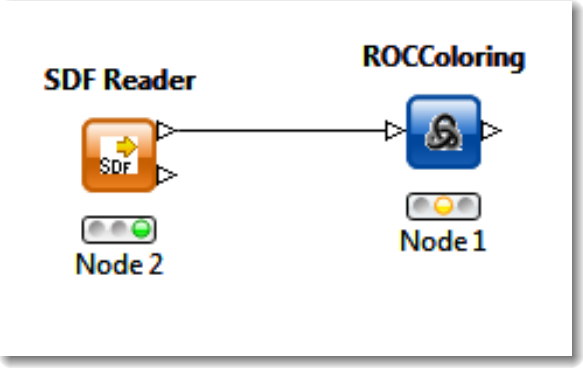

Figure 2.1: SDF Reader attached to the ROCColoring node. In order to supply the ROCColoring node with compounds, an SDF Reader from the chemistry extension is connected to its input port.

In the SDF Reader’s configuration the SDF files have to be presented in a specific order. This is done by means of the ”File selection” tab. The first file should be the original active set which is used for clustering. Afterwards, the docked actives and decoys SDF file are listed with the actives file at first. More docked actives and decoys SDF files can be added in the same alternating order. Additionally, the option ”Add column with source location” has to be enabled to allow a differentiation of the different files (see Figure 2.2).

In order to allow the ROCColoring node to access the docking score and the bioactivity information from the given files, the respective properties have to be enabled for extraction in the ”Property handling” tab (see Figure 2.3).

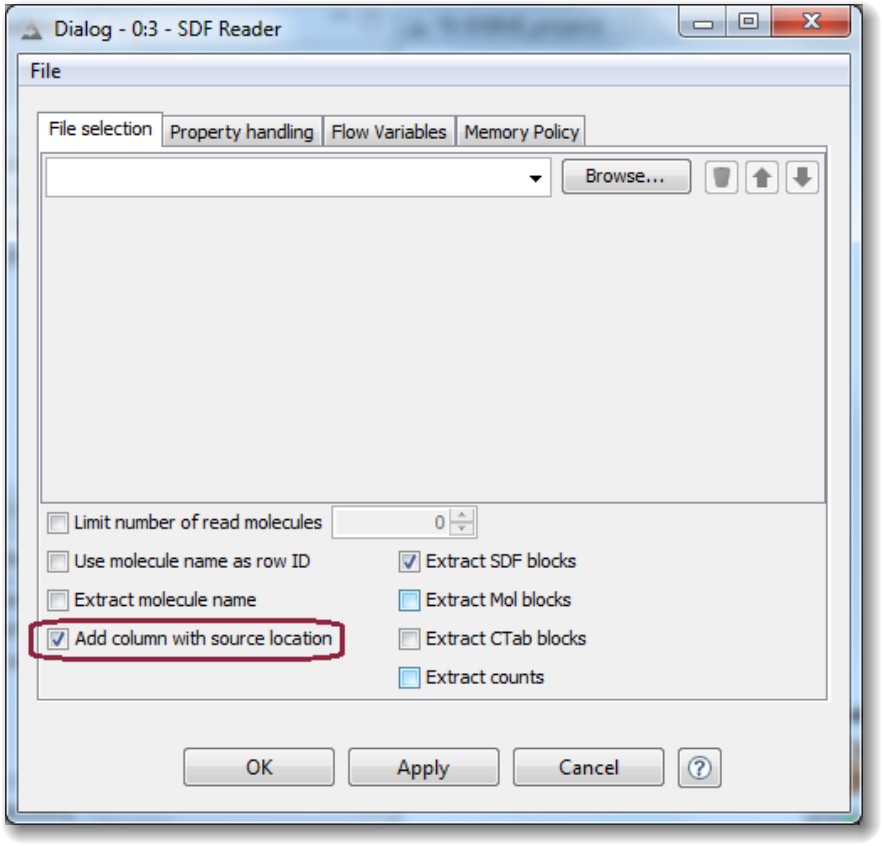

Figure 2.2: Selection of SDF files. The SDF files have to be supplied in the following order: 1. original active set for clustering, 2. docked actives SDF file, and 3. docked decoys SDF file. Additional docked actives and decoys SDF files can be added in same alternating order. The option ”Add column with source location” has to be checked to distinguish the different data sets.

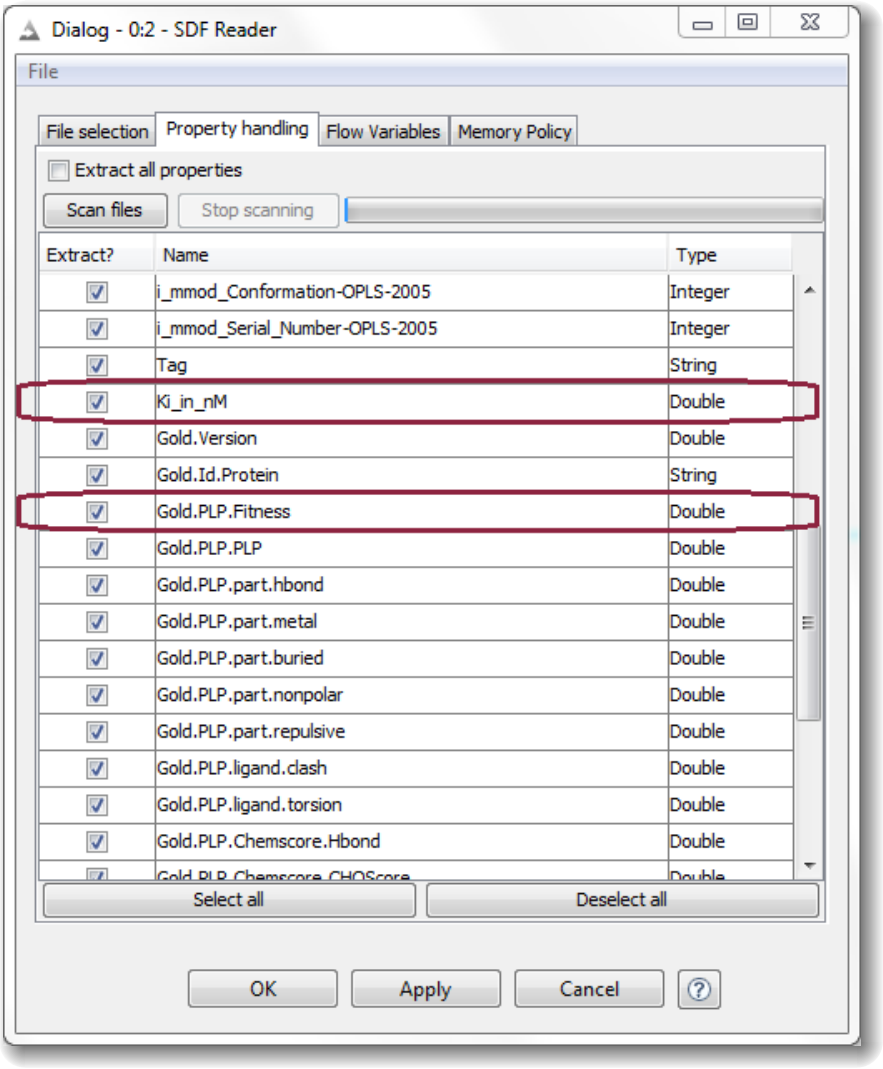

Figure 2.3: Property handling. At least the two properties for docking score and bioactivity information have to be extracted. In this example ”Ki_in_nM” represents bioactivity information and ”Gold.PLP.Fitness” the docking score.

3 Overview of parameters

The ROCColoring node has a single mandatory option which is ”Select property for ranking”. This option displays the extracted properties from the SDF files from which the property representing the docking score of the given compounds has to be selected (see Figure 3.1). It is assumed, that a high score corresponds to a position in the upper part of the ranking. If the score of a compound is inversely proportional to its rank, the option ”Invert docking score” has to be checked (see options group ”Further parameters”).

The remaining options influence the outcome of the clustering and can be left unchanged whereat their standard values will be used. Among the current options the following parameters can be changed:

- Molecule similarity threshold. This threshold sets the minimum score of the given similarity measures for two compounds to be even considered for clustering and computation of their maximum common substructure (MCS).
- Cluster similarity threshold. During the hierarchical clustering process, two clusters are joined only if their corresponding MCSs have similarity score equal or above the given value.
- Minimum MCS size. This parameters specifies the minimum size a MCS has to have to be even considered as valid during clustering.
- Fingerprint bond diameter. This diameter is used for the ECFP fingerprint.
- Preserve rings? It can be chosen whether to break aromatic rings during MCS computation or if they should only be matched as a whole.

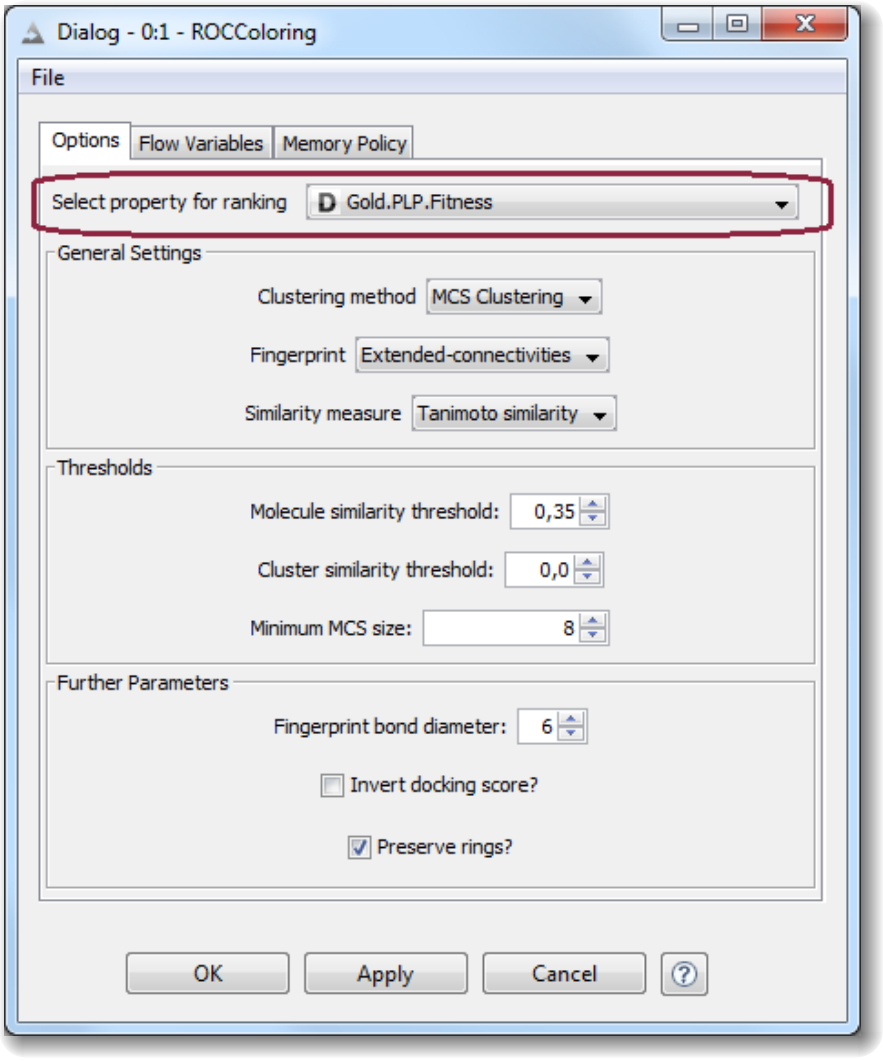

Figure 3.1: Parameters of the ROCColoring node. In this menu, the property with the docking score has to be selected by changing the option ”Select property for ranking”. The other parameters influence the clustering of the original active set.

4 Views and output of the node

After everything is set up, the user can start the ROCColoring node. As soon as the indicator switches from yellow to green, the node finished is computation and provides three views for visualization of its results (see Figure 4.1) :

- ROC Curve. Creates a pROC curve for each docked actives SDF file and annotates it with the maximum common substructures computed with the original active set.
- Ranking Box Plot. Depicts the ranks of each docked actives SDF file as a box plot.
- Inter-Cluster Similarity. Shows the inter-cluster similarity of each cluster compared to all clusters in form of a heat map based on the Tanimoto similarity.

Furthermore, the ROCColoring node creates a table at output port 0 containing the computed maximum common substructures.

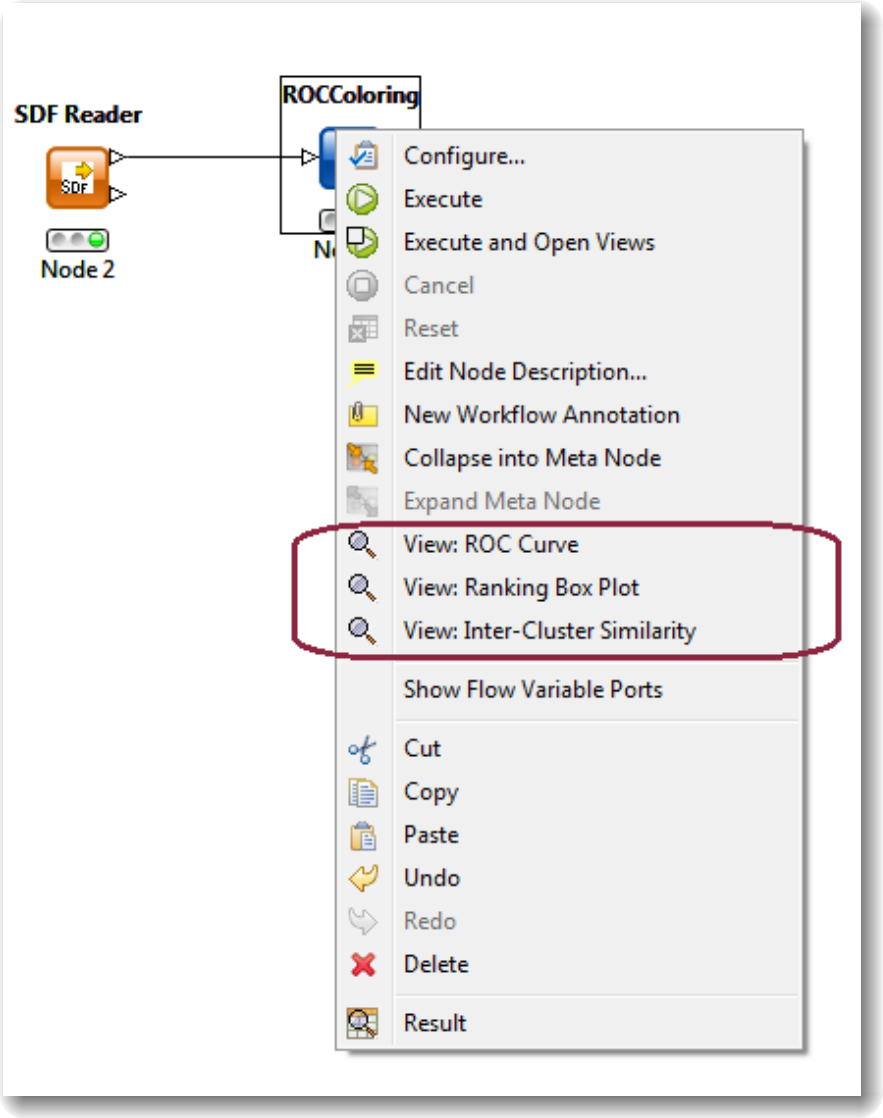

Figure 4.1: Available views after a completed run. Once the ROCColoring node has finished is calculations, the user has access to the three views ”ROC Curve”, ”Ranking Box Plot”, and ”Inter-Cluster Similarity”.

5 Final notes

A standalone version of this node is currently under development with additional features and options. It will be available in Java™ and as KNIME node.