EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

MATHEMATISCH-
NATURWISSENSCHAFTLICHE
FAKULTÄT

pROC-Chemotype Protocol

A User Guide – Pipeline Pilot Protocol

Authors

Dr. Tamer M. Ibrahim

and

Prof. Frank M. Boeckler

2015

Contents

I. Prerequisites:

The following prerequisites should be established to enable a proper function of the pROC-Chemotype protocol:

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.
3. **Docked decoys sdf file**, containing:
   - Docking information with a defined column name for the docking score.
4. **Installation of ChemAxon** (JChem, Library MCS version 0.7 or later) standalone program with a proper license.
5. **Installation of ChemAxon Pipline Pilot components**.
6. **Installation of R version 2.3.1**. Later versions may need some code adaptations for a proper pROC-Chemotype plot production.
7. **Downloaded pROC-Chemotype** protocol from “www.dekois.com”.

II. Overview of the pROC-Chemotype protocol:

The downloaded XML protocol can be imported using: **file > Open protocol**, which results in two main connected protocols, one for compiling pROC-Chemotype plots, and the other as an output subprotocol. The pROC-Chemotype protocol consists of three main workflows (as seen in Figure 1): (a) clustering and bioactivity assignment, (b) processing of the docked data, and (c) calculating pROC AUC, EF and pROC-Chemotype plotting.

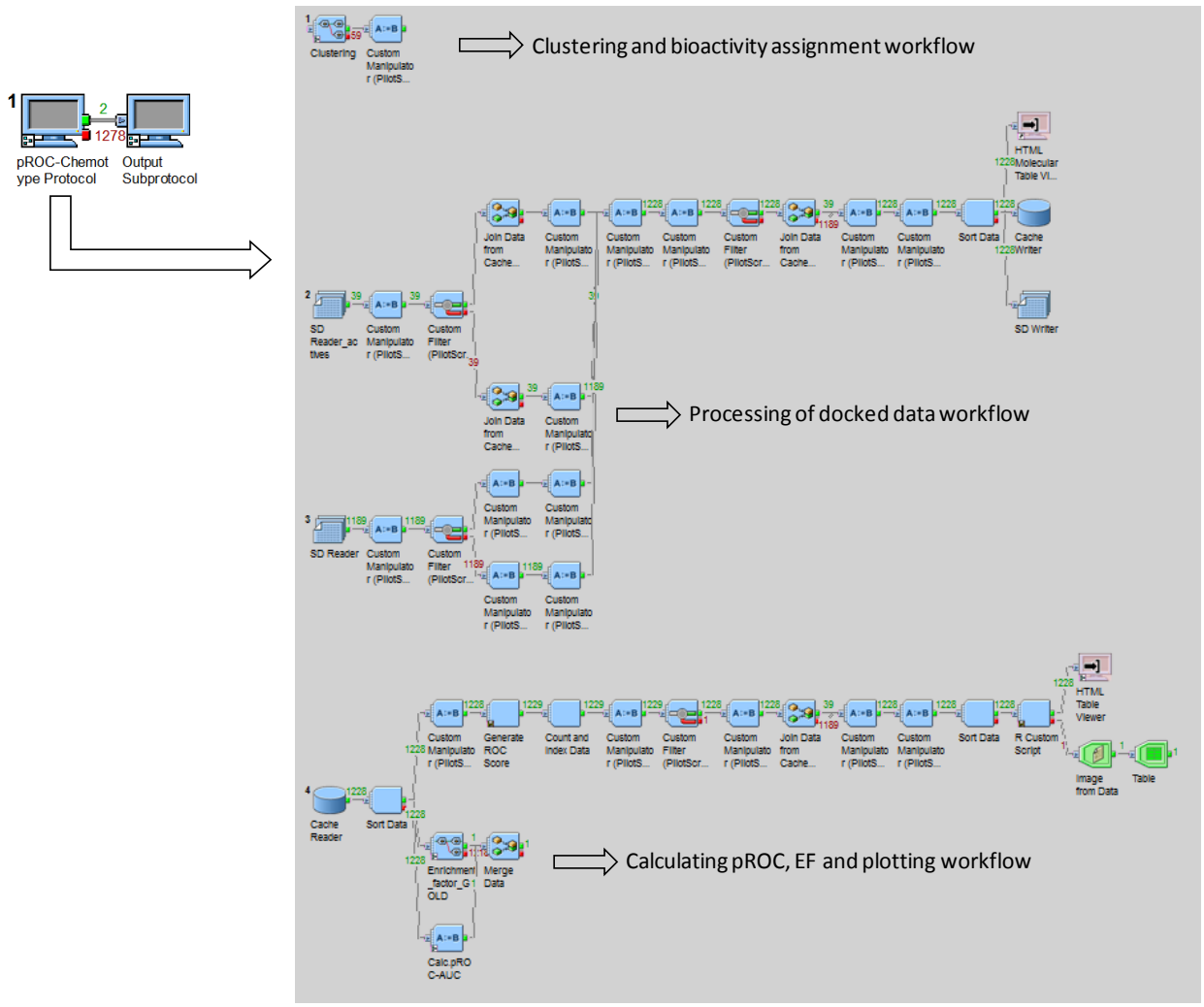

Figure 1. Overview of the pROC-Chemotype protocol. It consists of three main workflows for: a) clustering and bioactivity assignment, b) processing of the docked data, and c) calculating pROC AUC, EF and pROC-Chemotype plotting.

In the first workflow (clustering and bioactivity annotation/assignment), a sdf file of the prepared active set (e. g. prepared by Ligprep/Maestro) is used as an input for clustering. The 3D coordinates are converted into 2D coordinates and non-polar hydrogens are removed for clear 2D representation. Then, the 2D structures are clustered by LibMCS component of ChemAxon (see Figure 2). The output are either saved in a "cache writer" or processed further to retrieve the common substructure per MCS cluster either by FCFP_6#C smiles or from the external sdf clustered file by the standalone version of LibMCS of ChemAxon (more details can be found in the parameters section).

For the bioactivity annotation/assignment, the same sdf file of the active dataset (prepared by Ligprep) is used as a reference for bioactivity tagging. The ligands are ranked according to their K_I, K_D or IC_{50} values. If two ligands share the same bioactivity, the alphabetical order for the name (e.g. BindingDB name) is considered. For ligands with more than one bioactivity data, the priority is considered as K_I, then K_D, and then IC_{50}.

The bioactivitiy data in the sdf file should be given under one or all of the following fields/columns of the sdf file: “Ki_in_nM”, “Kd_in_nM” and/or “IC50_in_nM”. The numeric values should be provided as float in nanomolar range of activitiy/affinity. The protocol will process such fields to produce the Type of data (TOD) and level of activity (LOA) columns.

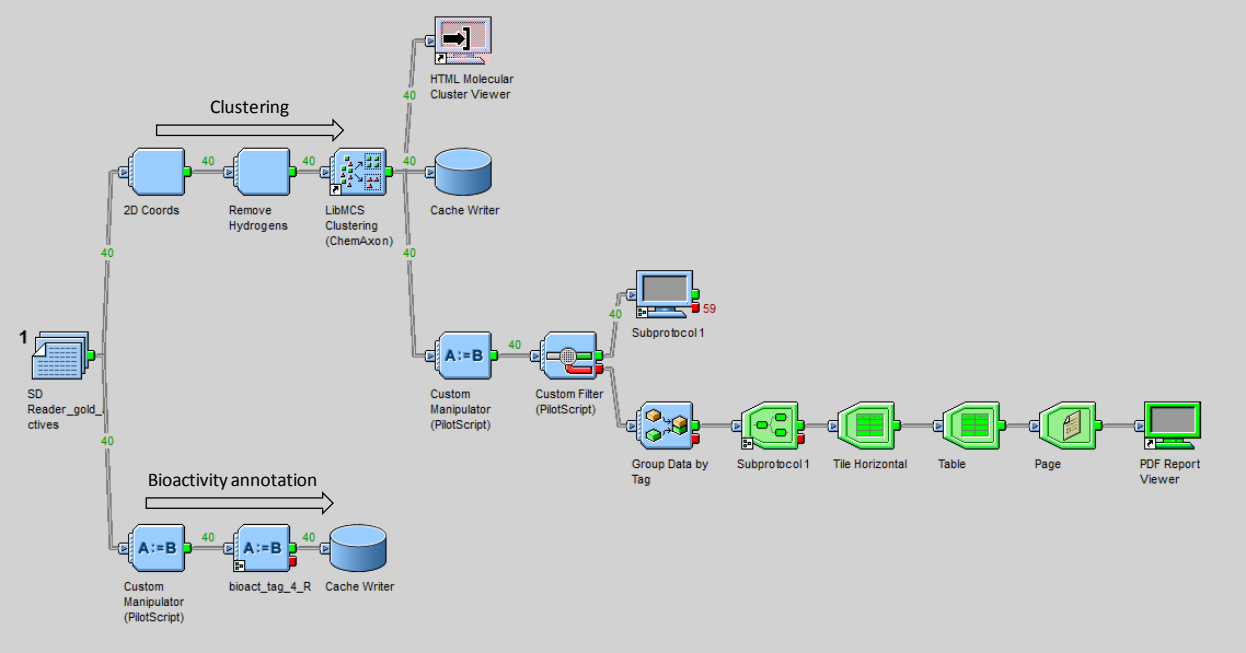

Figure 2. Overview of the clustering and bioactivity assignment workflow.

The docked actives and decoys files (two files) are processed for calculating the logROC AUC, EF calculations. The user should define the docking column in the sdf files. In addition, the expression of

the best scores should be considered. For instance, GOLD (ChemPLP) produces docking scores in a column called “GOLD.PLP.Fitness” and the best scores are expressed as maximum positive values. While for Glide (SP), the docking scores are annotated in the column “r_i_docking_score”, and the best scores are expressed as lowest values (negative values mimicking the binding affinity). By defining the previous parameters, the pROC-chemotype protocol can process any docking program.

The clustering and bioactivity information are joined from the previous workflow, and used in the bioactivity legend in the pROC-Chemotype plot.

III. Overview of pROC-Chemotype parameters:

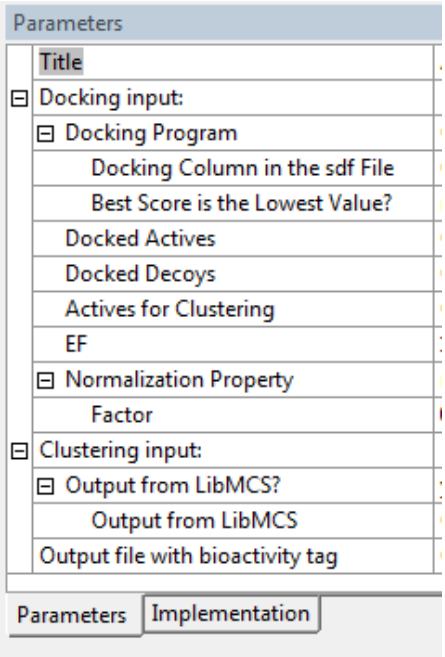

Figure 3. pROC-Chemotype parameters for the user interface.

The general parameters for the user interface are aimed to ease the production of the pROC-Chemotype plots. For instance, the destination URLs for the docking files, prepared active set … etc., link directly the processing components in the different workflows and subprotocols (Figure 3 and Figure 5). These general parameters can be summarized as:

- **Title**: is usually a “string” consisting of the target (dataset) name and preparation protocol.
- **Docking input:**
  - ***Docking Program***: is the name of the docking program used.

- ***Docking Column in the sdf File***: is a "string" directing for the column/field/property name in the sdf file where the docking scores are.
- ***Best Score is the Lowest Value***: is either "**yes**" or "**no**". For instance, the value is "yes" for Glide and "no" for GOLD (ChemPLP). .

- ***Docked Actives***: is a "URL" for the docked file of the active set. The docked file should be a sdf file with a defined "***Docking Column in the sdf File***".
- ***Docked Decoys***: is a "URL" for the docked file of the decoy set. The docked file should be a sdf file with a defined "***Docking Column in the sdf File***".
- ***Actives for Clustering***: is the "URL" for the prepared active set to be used for the ChemAxon-based LibMCS clustering. In our study we use the active sets prepared by Ligprep (Maestro).
- ***EF***: is the enrichment factor calculation. The value 1 here means 1% of the given ranked dataset. The value can be changed if desired.
- ***Normalization Property***: is a "string" for the calculatable property that will be used for normalizing the docking score for calculating the AUC and EF. For instance, "num_atoms" refers to the number of heavy atoms.
  - ***Factor***: is a "float" number which used as the power of the given normalization property. For instance, normalization property N with a factor 0.5 is $N^{0.5}$ ($N^{1/2}$). To retrieve original docking results with no normalization factors, then factor parameter should be set to 0. (N.B. the factor should only be given as a float number, e.g. 0, 0.5, 0.6667 …etc; and not as ½, 2/3).

- **Clustering input:**
  - ***Output from LibMCS?***: is either "**yes**" or "**no**". If yes, the subsequent parameter "***Output from LibMCS***" which is a "URL" for the clustered sdf file by the standalone version of the Library MCS-ChemAxon, should be given. In other words, two possibilities are available:

    A) *Output from LibMCS?* is **"yes"** when a clustering of the active set has been performed by the user with the standalone program of Library MCS of ChemAxon independently from the LibMCS component of Pipeline Pilot, and the output file of such standalone clustering is retrieved. Then, the "***Output from LibMCS***" URL should be directed to such output file in order to represent the substructures of the MCS per cluster (i.e. HierarchyID with no branches) from such file.

    To produce such clustered file by the standalone Library MCS program, a simple protocol should be established by the user. This can be simply done by feeding the original active set sdf file to a sdf reader to convert the molecular data into 2D coordinates followed by removal of non-polar

hydrogens. The written output can then be used an input for the standalone Library MCS program of ChemAxon (e.g. see Figure 4). Similar settings of the standalone Library MCS and the LibMCS component of Pipline Pilot usually produce similar results.

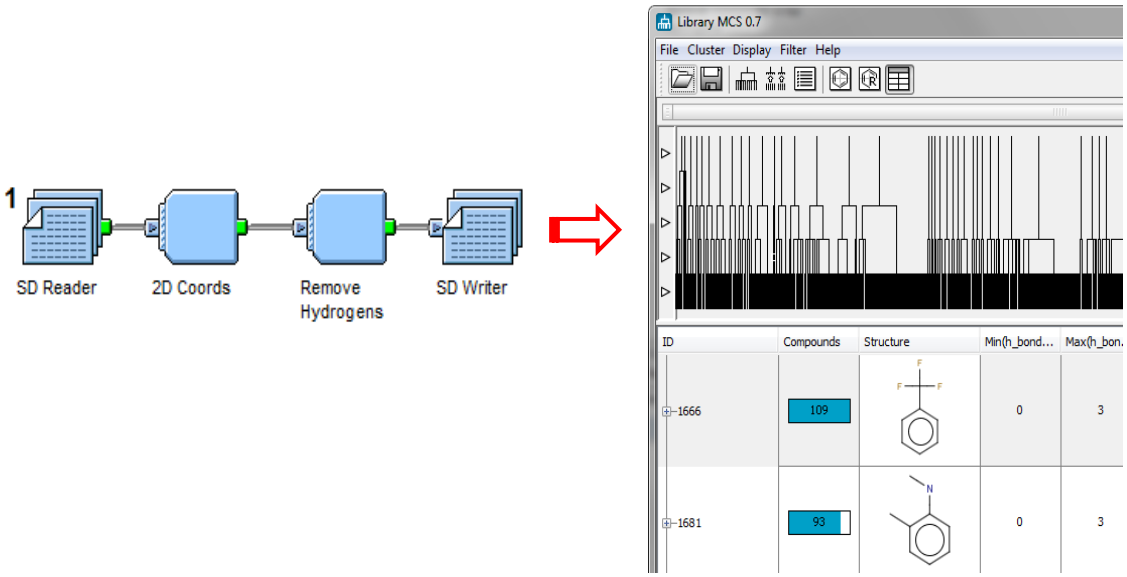

Figure 4. A simple schematic description for producing clustered active set by the standalone Library MCS program of ChemAxon.

B) Alternatively, if the user did not perform the LibMCS independently by the standalone program, then the parameter ***"Output from LibMCS"*** should be **"no"**, and the 2D representation of the common substructure(s) per LibMCS cluster is presented by extracting the largest and common FCFP_6#C smiles per LibMCS cluster. Since the output of the LibMCS component and the standalone program are identical when performing identical options, we recommend the user to perform the separate clustering with the standalone program of ChemAxon (i.e. option A) for getting more accurate representative MCS per LibMCS cluster.

- ***Output file with bioactivity tag***: is a "directory" for locating the processed sdf file with all information of the docked actives and decoys with the respective Cluster, TOD, LOA, fitness (docking score), Rank_DB (docking score-ordered rank), Rank_bioact (bioactivity-ordered rank), …etc. The output file is by default has the name "output_<title>.sdf".
- It is noteworthy that the LOA is expressed as the –log(LOA), for instance, if there is an active with a micromolar range of activity, its LOA will be 6, nanomolar will be 9, and so forth.

Below (Figure 5) is a shown example with full parameters:

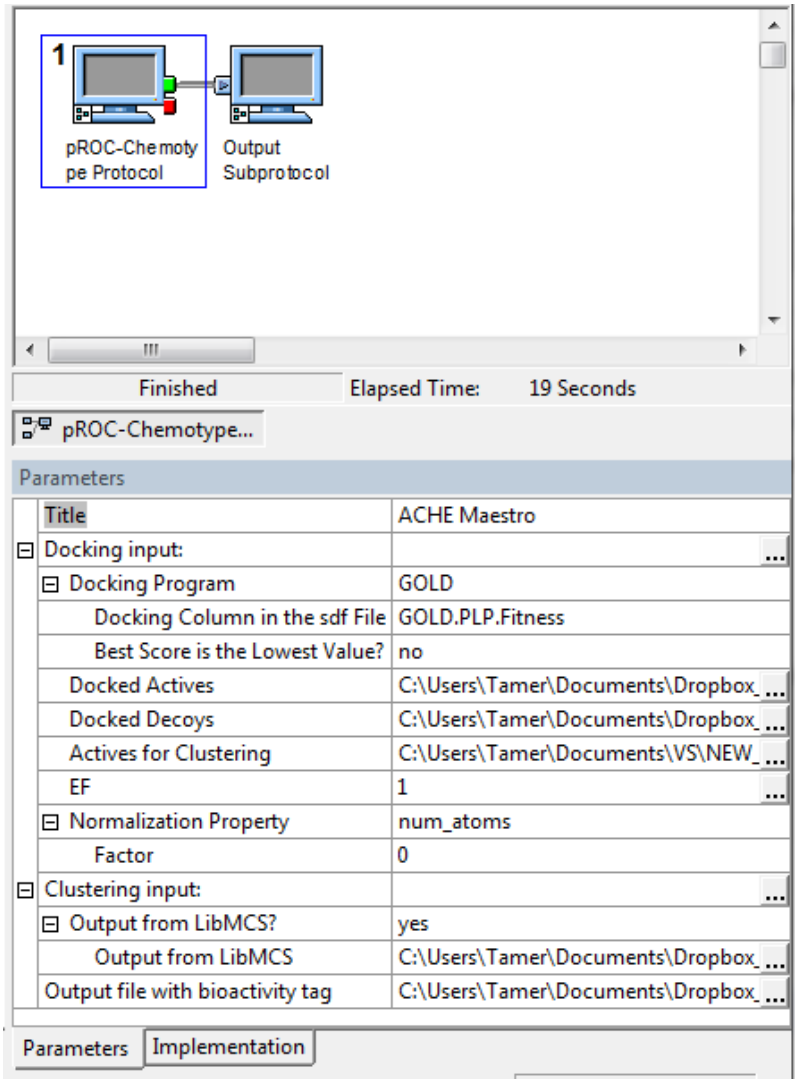

Figure 5. Overview of the pROC-Chemotype protocol and interface with given examples of the parameters.

IV. Additional notes and output examples:

The output of the protocol can be summarized as:

1- Clustering output:

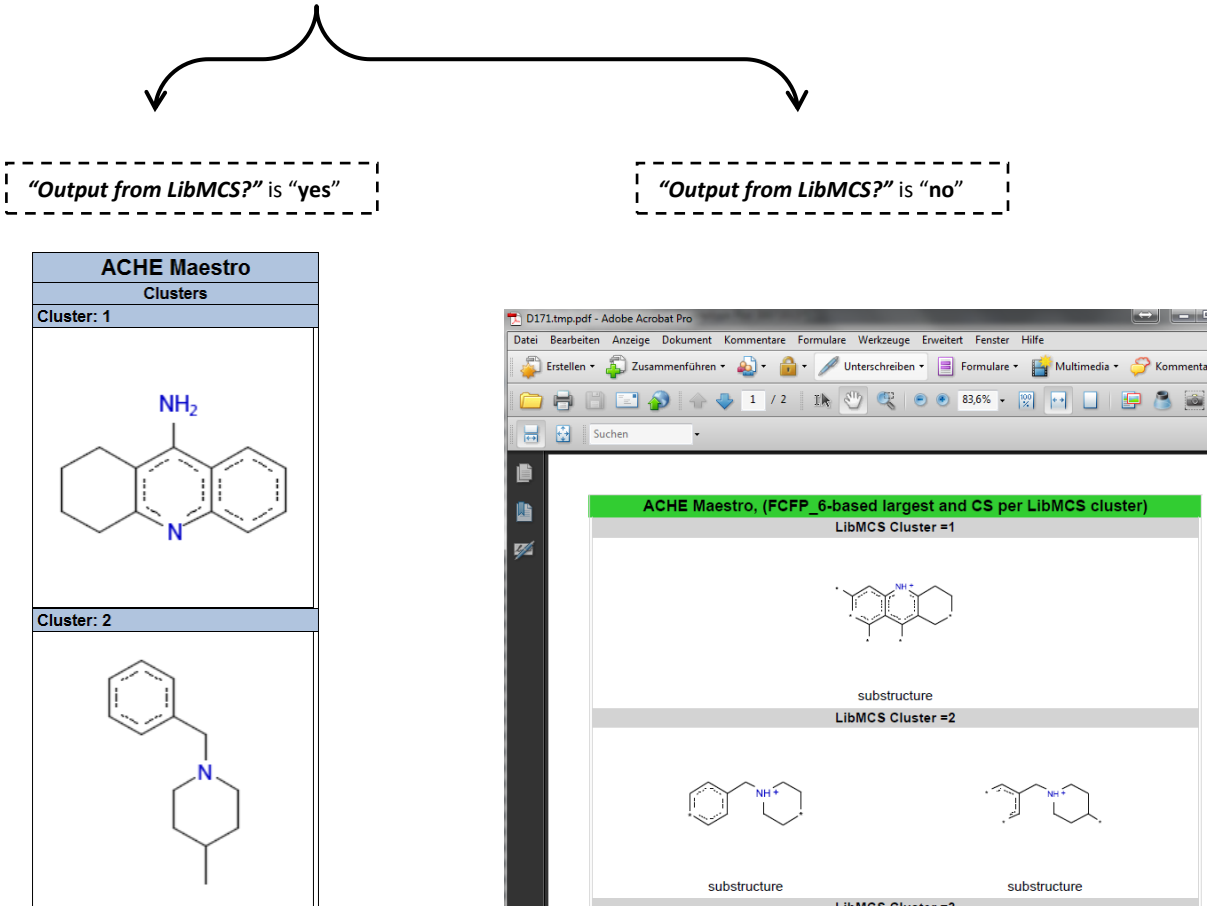

Figure 6. Clustering output. Either a HTML page or a pdf file is produced for "yes" and "no" strings of the parameter "Output from LibMCS", respectively.

2- Docking/benchmarking output:

This includes: pROC-AUC and EF as shown below (Figure 7).

EF 1%	Title	logROC-AUC
0	ACHE Maestro	0.716234

Figure 7. Evaluation of docking results. A HTML page is produced containing EF and pROC-AUC information.

3- pROC-Chemotype plot:[1]

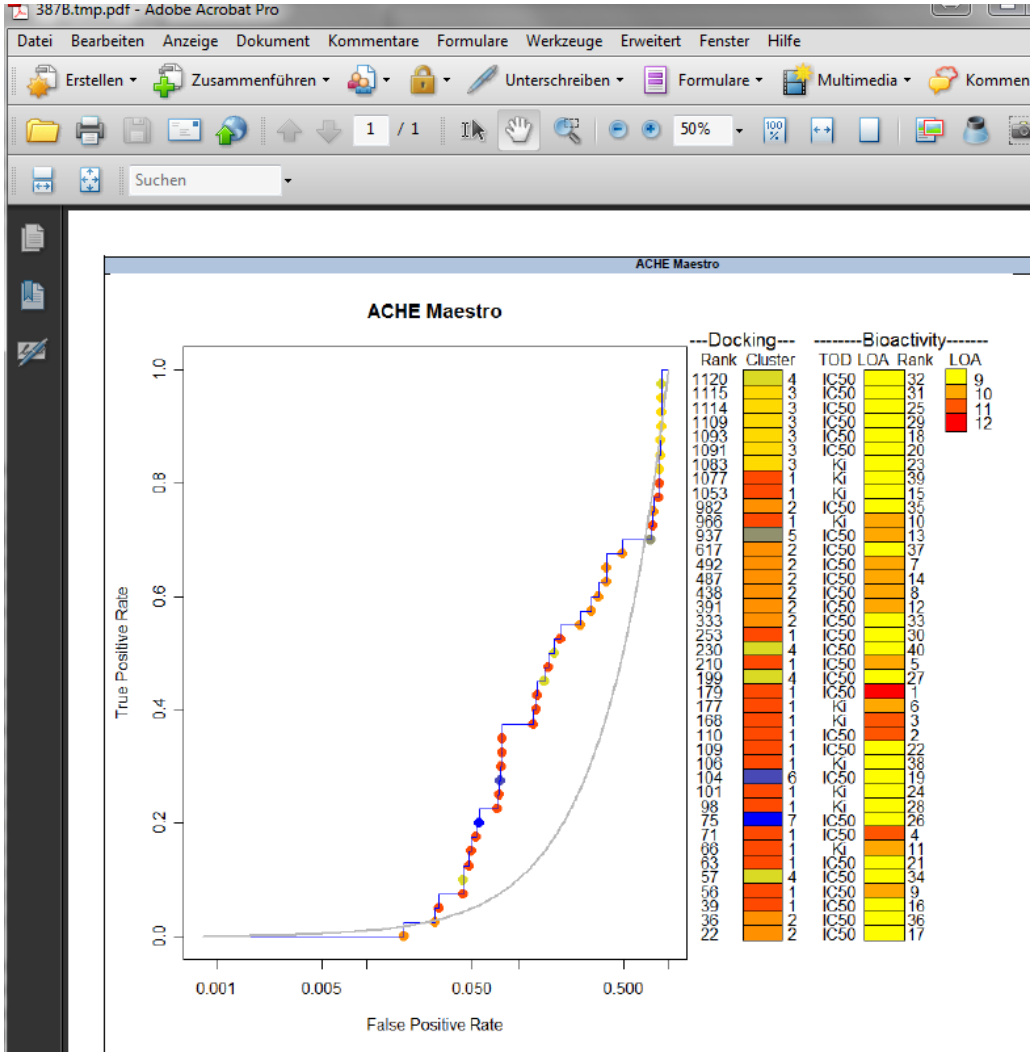

Figure 8. pROC-Chemotype plot. A pdf file is produced displaying the pROC-Chemotype plot.

In addition to the previous description of the output results, more than one pROC-Chemotype protocol can be concatenated to the same one "output subprotocol" component as shown below (Figure 9):

[1] It is worthy to mention that the parameters and dimensions for pROC-Chemotype plotting (in the R code) are adjusted for the number of the bioactives of the DEKOIS 2.0 data sets. Adjusting such parameters may be needed for other datasets or different numbers of bioactives.

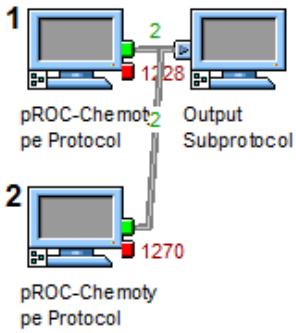

Figure 9. Two p-ROC-Chemotype protocols are concatenated to one Output subprotocol.

This will result into concatenated pROC-AUC and EF results in one HTML page, and two pROC-Chemotype plots in one pdf file. The clustering data will be produced separately into two pdf/HTML files (see Figure 10 and Figure 11).

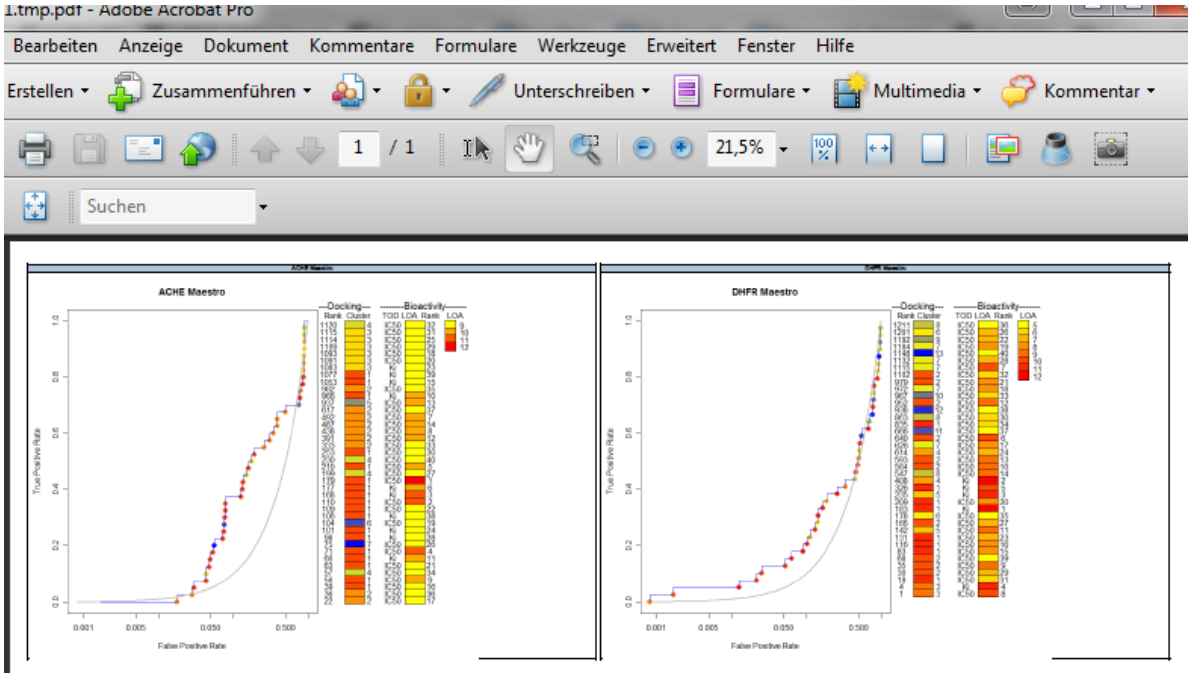

Figure 10. Two pROC-Chemotype plots concatenated into one pdf file.

EF 1%	Title	logROC-AUC
0	ACHE Maestro	0.716234
5.214	DHFR Maestro	0.634052

Figure 11. Two entries of docking evaluation concatenated into one HTML page.