AuPosSOM 2.1 online

Tutorial

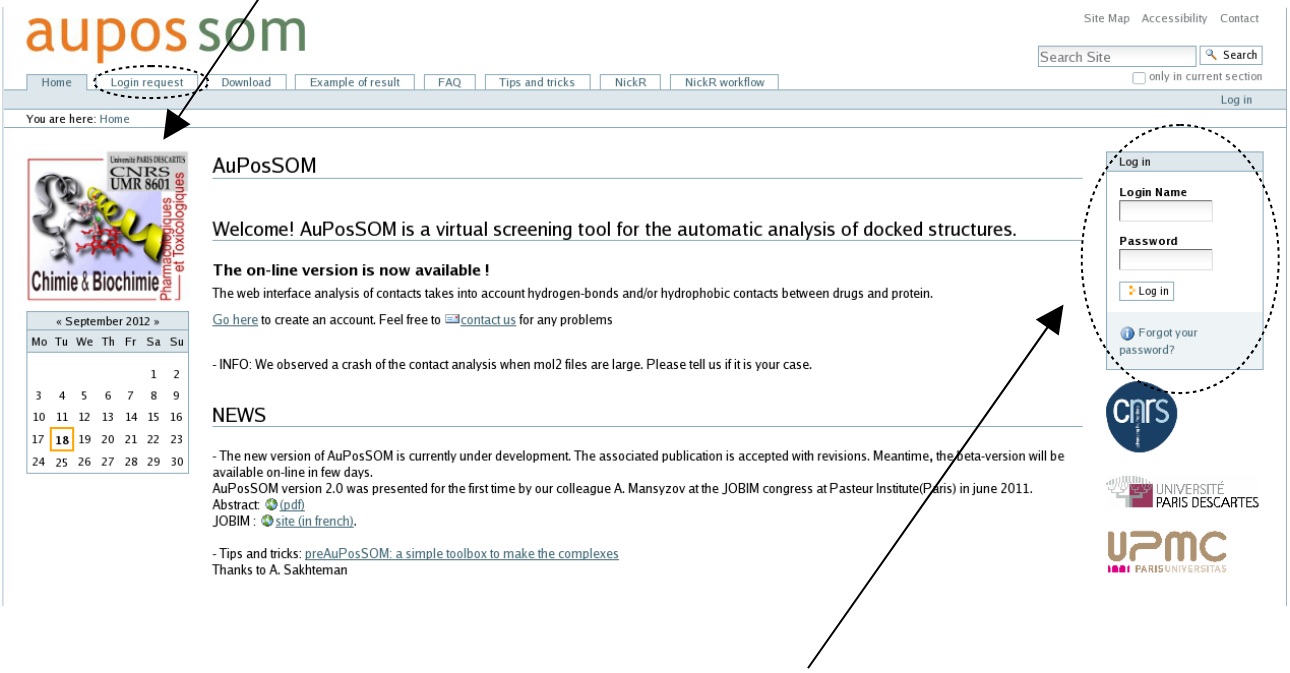
Welcome to the **AuPosSOM tutorial** which will guide you through the procedure to run the Automatic analysis of Poses using Self-Organizing Map (AuPosSOM)**.**

AuPosSOM is able to classify ligands docked to a protein according to their contact footprint. This tool is helpful to identify active compounds inside a dataset of molecules.

Table of contents :

1 Create an account

Click on login request and fill in the form



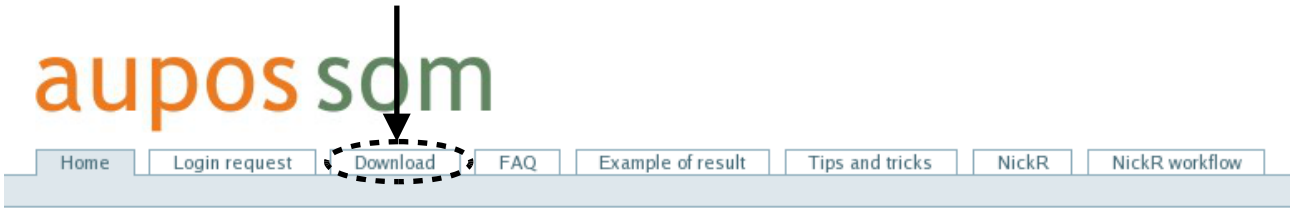
Once activated you can access to the web interface by filling your login name and password

You will be informed by mail when your account will be created.
It should take less than 24 hours to activate the account
during the weekdays upon request to the administrator.

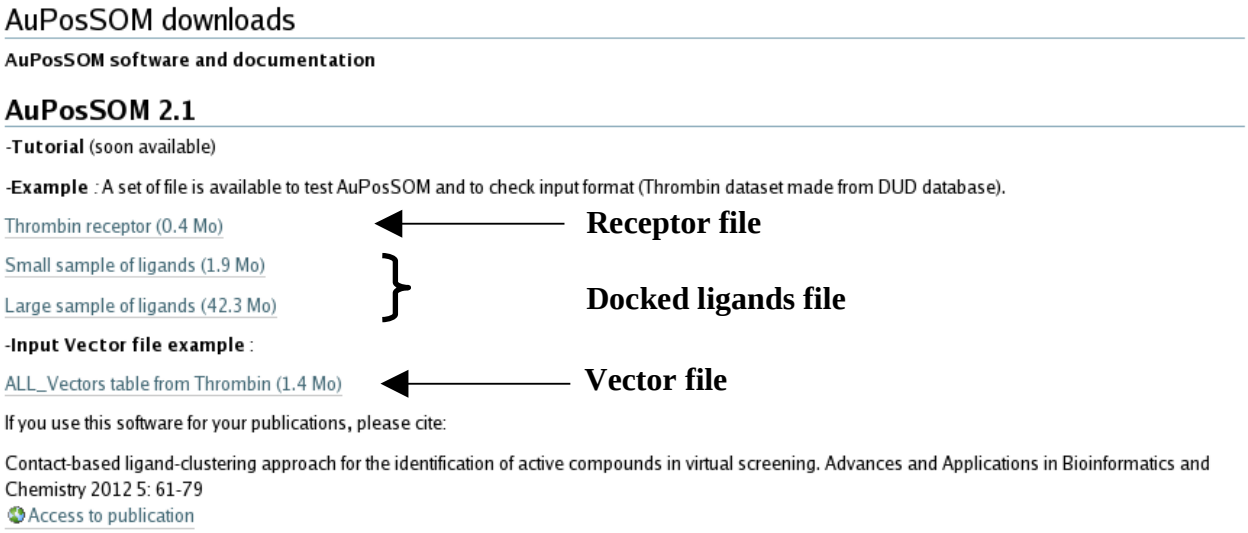
Prior to run a calculation, it is essential to know the proper input files format that we will use. This is true for any program. Let's have a look at these files in the next section.

2 Prepare your input files

In order to check the format for **input files**, let's go to the download section



The following page contains samples for the different input files you will need to run calculations with AuPosSOM.



AuPosSOM calculation can be run in three different modes. The following input files are required :

* A **receptor file** and a corresponding **docked ligands file.**
  → AuPosSOM will perform « contact detection » and « ligand clustering ».

  or

* A **vectors file** which already contains contact information.
  → AuPosSOM will execute only « ligand clustering ».

  or

* Two **vectors files** containing the same ligand set which can be merged to combine two different contact information (e.g. hbond and lipophilic).
  → AuPosSOM will execute « vector fusion » and « ligand clustering ».

A common problem which occurs in bio-informatics regards the use of input files. For example, PDB or mol2 format files may exist in several sub-formats regarding number of columns, position, decimals in float number.

Please note that AuPosSOM2.1 no longer accepts the pdb format file. However, there are some programs (fconv, chimera...) which allow to convert your files from pdb to mol2 format. If you cannot convert your files into mol2 format, the « AuPosSOM web interface » (old version) still allows the use of pdb files.

AuPosSOM was written in order to accept most of the mol2 «sub-formats». However, we recommend the use of a sub-format similar as displayed below. Check also there is **no blank line** in the area between `@<TRIPOS>ATOM` and `@<TRIPOS>BOND`.

Here is an overview of the different format files supported by AuPosSOM2.1 web interface :

2.1 Protein mol2 file

```
@<TRIPOS>MOLECULE
rec_cys.pdb
4060 4113 252 0 0
PROTEIN
AMBER ff99SB


@<TRIPOS>ATOM
      1 N           5.0820   -9.3130   18.3130 N.4       1 ILE1        0.0311
      2 CA          4.2660   -9.0560   19.5150 C.3       1 ILE1        0.0257
      3 C           3.2250  -10.1280   19.7230 C.2       1 ILE1        0.6123
      4 O           2.4080  -10.4370   18.8370 O.2       1 ILE1       -0.5713
      5 CB          3.5700   -7.6220   19.4670 C.3       1 ILE1        0.1885
      6 CG1         4.5630   -6.4780   19.1510 C.3       1 ILE1       -0.0387
      7 CG2         2.7170   -7.3670   20.7420 C.3       1 ILE1       -0.3720
      8 CD1         5.6650   -6.2030   20.1910 C.3       1 ILE1       -0.0908
      9 N           3.1140  -10.5720   20.9790 N.am      2 VAL2       -0.4157
     10 CA          2.1910  -11.5930   21.4000 C.3       2 VAL2       -0.0875
     11 C           1.1430  -10.9650   22.3180 C.2       2 VAL2        0.5973
     12 O           1.4480  -10.3480   23.3340 O.2       2 VAL2       -0.5679
     13 CB          2.9440  -12.8000   22.0460 C.3       2 VAL2        0.2985
     14 CG1         1.9940  -13.8860   22.5780 C.3       2 VAL2       -0.3192
     15 CG2         3.9620  -13.4450   21.1200 C.3       2 VAL2       -0.3192
     16 N          -0.1030  -11.2150   21.9150 N.am      3 GLU3       -0.5163
     17 CA         -1.2850  -10.7460   22.6270 C.3       3 GLU3        0.0397
     18 C          -1.3910   -9.2250   22.6160 C.2       3 GLU3        0.5366
     19 O          -1.9090   -8.6340   23.5790 O.2       3 GLU3       -0.5819
     20 CB         -1.2660  -11.2770   24.0520 C.3       3 GLU3        0.0560
     ----------------------------------------------------------------------------
   4055 HE1        14.1870    6.6596   31.0651 H       215 MET215      0.0684
   4056 HE2        12.8084    6.0798   32.0299 H       215 MET215      0.0684
   4057 HE3        13.7580    4.9321   31.0556 H       215 MET215      0.0684
   4058 HE2         8.1709   15.8006   17.3844 H       116 HIS116      0.3339
   4059 HE2        28.8682    1.4348   26.6615 H        87 HIS87       0.3339
   4060 HE2        17.3534    1.5784   34.6785 H       235 HIS235      0.3339
@<TRIPOS>BOND
```

Important remarks :

-Please note that contact detection in the lipophilic mode requires the presence of hydrogen atoms.
-Charge information is contained in the last column at the right side. In the present example, it was generated using an AMBER force field. You may adjust the module of the partial charge for Coulomb contact analysis (polar contacts) to fit with your own charges.
-Sometimes, the residue number maybe be shifted with respect to the full sequence residue number (e.g. `47 TYR150`). The residue number used in AuPosSOM calculation is the one placed after the three-letters residue name (e.g. 150 in «`47 TYR150`»).

2.2 Ligand mol2 file

```
@<TRIPOS>MOLECULE
ZINC01545762
   49    51     0     0     0

SMALL
USER_CHARGES

@<TRIPOS>ATOM
1    C     13.4375  -17.1800   18.6140  C.3   1 <0>         0.0874
2    N     12.5230  -16.7001   19.6128  N.pl3 1 <0>        -0.9521
3    C     12.8686  -15.7257   20.5064  C.ar  1 <0>         0.1789
4    C     12.6636  -15.9118   21.8947  C.ar  1 <0>        -0.1474
5    C     12.9744  -14.8960   22.8114  C.ar  1 <0>        -0.0667
6    C     13.4971  -13.6681   22.3544  C.ar  1 <0>        -0.2041
7    C     13.6903  -13.4493   20.9775  C.ar  1 <0>         0.1354
8    C     13.4070  -14.4949   20.0655  C.ar  1 <0>        -0.1350
9    O     14.1468  -12.2216   20.5672  O.3   1 <0>        -0.3152
10   C     13.1463  -11.3916   19.9252  C.3   1 <0>         0.0403
     ------------------------------------------------------------
40   H      9.7106  -12.0725   24.3715  H     1 <0>         0.1927
41   H     11.3065   -8.5721   25.8759  H     1 <0>         0.1928
42   H     13.0144   -8.5469   24.1608  H     1 <0>         0.1793
43   H     10.6991  -19.9653   19.7745  H     1 <0>         0.1385
44   H     10.1642  -21.7435   18.2303  H     1 <0>         0.1354
45   H      9.6926  -21.2438   15.9032  H     1 <0>         0.1349
46   H      9.7560  -18.9468   15.1166  H     1 <0>         0.1369
47   H     10.2756  -17.1528   16.6464  H     1 <0>         0.1407
48   N     10.4102  -10.3139   25.2133  N.ar  1 <0>        -0.5309
49   H      9.7207  -10.3155   25.8898  H     1 <0>         0.4736
@<TRIPOS>BOND
```

Remarks:

All ligand poses should have exactly the same label written under the flag :
`@<TRIPOS>MOLECULE` (red in this example).

Let's suppose you have five poses and are interested in seeing a unique contact footprint for the molecule, the labels must be identical :

`ZINC01545762, ZINC01545762, ZINC01545762, ZINC01545762, ZINC01545762`

Otherwise, if you want a distinct footprint for each pose and see if some are clustering, you need to label poses with a suffix :

`ZINC01545762_01, ZINC01545762_02, ZINC01545762_03, ZINC01545762_04, ZINC01545762_05`

2.3 Vector file

A vectors file (plain text) can be used as itself to run calculations. The first line contains contact labels (Atom Label_Residue Number_Residue Type_Atom number). The first column contains names for ligands. The matrix contains the number of average contacts for the different ligands.

Original plain text file example:

```
C_194_CYS_1555 O_194_CYS_1556 N_195_GLU_1559 C_195_GLU_1561 (n columns : contact labels)
ZINC03867537 0.05 0.00 0.00 0.45              (n +1 columns : ligand name + n contact values)
ZINC04334201 0.45 0.00 0.00 0.30
ZINC00579389 1.12 0.20 0.33 0.48
ZINC04619253 0.00 0.00 0.00 0.05
```

To display in a table, use a single space as delimiters for columns. In order to align contact labels to their respective values, please insert one cell at the upper left corner to shift all labels by one position to the right. As shown in the example you can add one lines above in order to number the contacts. In this way, it will be possible to assign the different contacts from **fingerprint2d.pdf** output file (see output section 6.1).

Display with Excel or OpenOffice Calc:

Contact number	**0**	**1**	**2**	**3**
→ **insert**	C_194_CYS_1555	O_194_CYS_1556	N_195_GLU_1559	C_195_GLU_1561
ZINC03867537	0.05	0.00	0.00	0.45
ZINC04334201	0.45	0.00	0.00	0.30
ZINC00579389	1.12	0.20	0.33	0.48
ZINC04619253	0.00	0.00	0.00	0.05

Please note that the vector file is either an input format and an output format which is very convenient to extract one leaf from a previous calculation to run a sub-tree for a more accurate analysis of contacts.

2.4 Active ligand list (optional)

If you have some information regarding active compounds (natural ligands, substrates, inhibitors...) you can insert a simple list (one ligand per line) of your active compounds in a *.txt format. Therefore, the program will :

- display active compounds in the contact map
- plot the distribution of active compounds in leaves
- execute ROC plot to evaluate clustering efficiency (see output section 6.4 for details).

This option allows you to compare clustering efficiency between different kinds of contact.

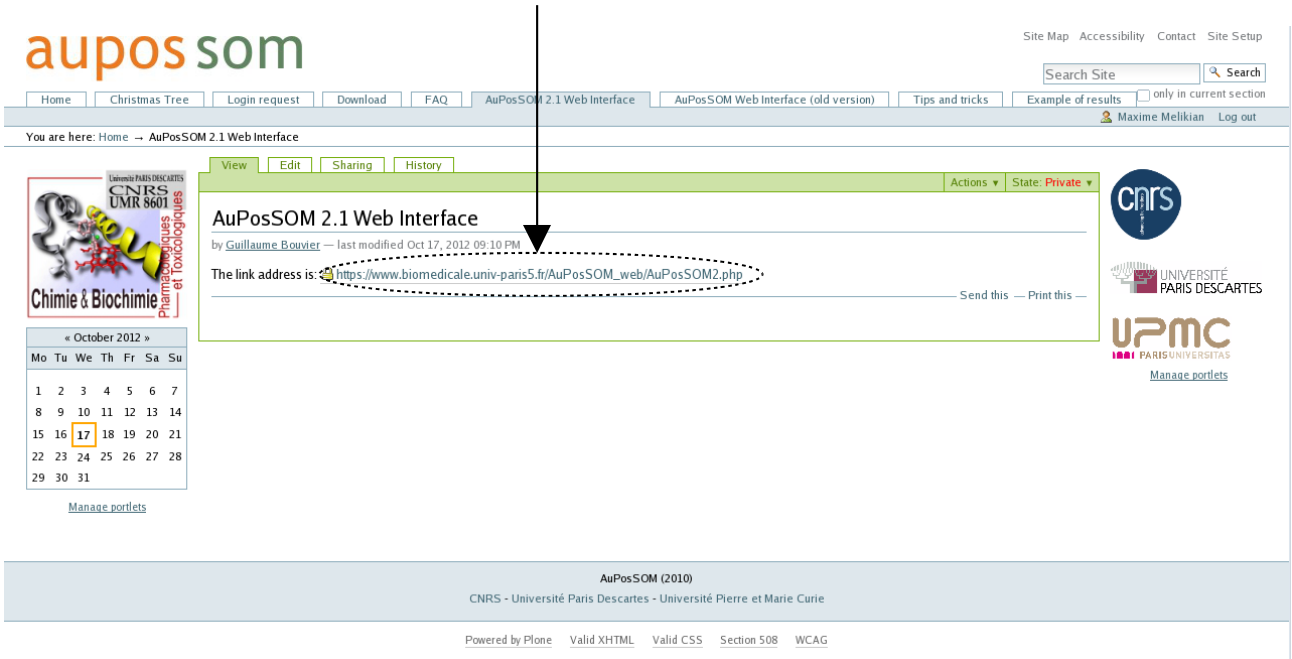
```
ZINC03954972      (ligand name)
ZINC03987397
ZINC03817609
ZINC01909869
...
```

Now, that all our input files are ready, let's access to the web interface.

3 Access to AuPosSOM2.0 web interface

Click on the **« AuPosSOM 2.1 Web interface »** section

Click on the link to access to the AuPosSOM 2.1 Web interface



Let's see how to run contact analysis and ligand clustering.

4 Upload input files and setup parameters

AuPosSOM v2.1 Web Application

Required files and parameters

See download section to check your input file format. You can also use these templates to run a calculation test.

Email address:

Send e-mail? yes

Receptor mol2 file: **Protein file in mol2 format (ex : thrombin_rec_charged.mol2)**

Parcourir… (not required if Vector File submited)

Ligands mol2 file: **Docked ligands file (ex : thrombin_output.mol2)**

Parcourir… (not required if Vector File submited)

(Multi mol2 file with all your docked ligands.
If you have multiple poses all your poses must have the same @<TRIPOS>MOLECULE field.
Poses for the same ligand must be sorted together in the file.)

Files bigger than 20 Mbytes should be submitted in gzipped .mol2.gz format.

Type of contacts for analysis:

polar **Try different kind of interactions for contact analysis**

Threshold for the module of the partial charge (for polar contacts analysis):
0.5 **Setting for contact selection (default value)**

Vectors File ('ALL_vectors_table' or 'leaf_#'):

Parcourir… (not required if receptor and ligands files are submited) **Matrix of contacts : starting point for clustering refinement or to make a sub-tree**

Run AuPosSOM

Click on the box to start calculation !

Optional files and parameters

Second Vectors File for Vectors fusion :

Parcourir… **Add contact information to the first vector file**

Active list file (List of active ligands, text file, one ligand per line):

Parcourir… **Produces ROC plot and mapping of active compounds**

Map size:

X: 5

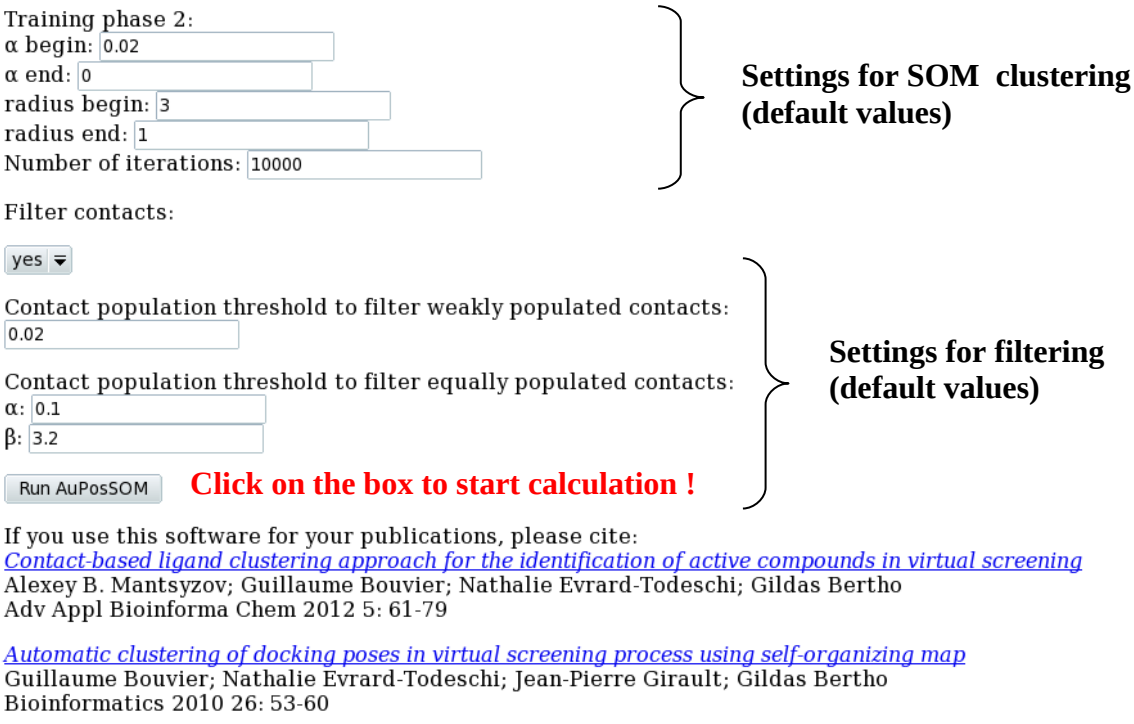
Y: 4 **XY product will determine the maximum number of leaves in the output**

Number of phase:

2

Training phase 1:
α begin: 0.2
α end: 0
radius begin: 6
radius end: 1
Number of iterations: 1000

Settings for SOM clustering (default values)



Note : the output will be named according to the E-mail address you entered.

After calculation has started, if all input files are in a correct format, the program will show you the run progression step by step. If some essential input is missing, the program will end immediately with error.

Let's see what happens when the run finishes successfully and how to retrieve your files.

5 Retrieve your results from calculation

When the run ends successfully, the following message should appear :

Run completed
You can download your files here (click right button)
This tar.gz archive file contains all results.

Description of the archive file content.

Directories:
- **AuPosSOM_analysis_nofilt** : results for the clustering without filtering.
- **AuPosSOM_analysis_filt**: results for the clustering with filtering.

Files:
- **AuPosSOM_Calculation_report.txt** : Information regarding calculation progress.
- **AuPosSOM.conf** : Input and parameters information used for calculation.
- **AuPosSOM.inp** : List of active and inactive compounds. If no active ligands list is provided; all compounds will be listed as unknown.
- **AuPosSOM.log** : Specific information regarding clustering process.
- **ALL_vectors_table**: the contact matrix for all compounds of the dataset.
- **ALL_vectors_1d** : Contact matrix for average contacts in all active vs all inactive compounds.
- **fingerprint2D.pdf** : the heatmap for the clustered contact matrix; numbers of the leaves are pointed out in red.
- **fingerprint2D_tree_0.pdf** and **fingerprint2D_tree_0.png** : heatmap with contact labels shown below and red arrows pointing active compounds (if active list uploaded).
- **leaf_n** : the contact matrix for the compounds of the leaf n.
- **map0.tree**: the Newick tree file which can be opened with PhyloWidget for example.
- **scoring0.txt** and **scoring0.pdf**: the scoring for leaves of the tree.

Additional output (if a list of active compounds has been submitted) :

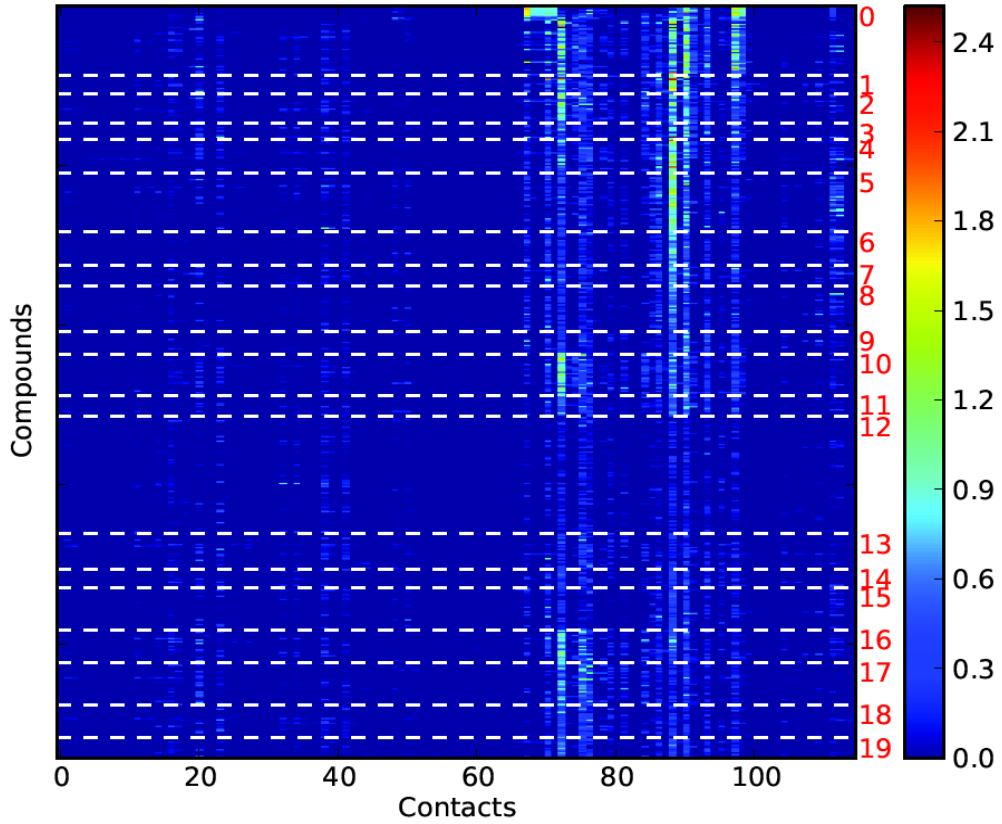
- **active_comp0.pdf** and **active_comp0.png** : Plot representing the percentage of active compounds from all active. Number of active compounds in each leaf divided by total of active compounds.
- **ratio0.pdf** and **ratio0.png** : Plot representing the percentage of active compounds in the leaf. Number of active compounds in each leaf divided by total number of compound in the leaf.
- **roc0.pdf** : ROC plot representing the fraction of true positives (active compounds) out of the positives vs. the fraction of false positives out of the negatives; at various threshold settings (see section 6.4 in tutorial for details).

After the download, the results file needs to be uncompressed. Make sure your output is obtained in a **username@address_0123456789012345.tar.gz** format. If the gz extension is missing, you should rename it manually prior to extract.

Now, let's have a look at the different **output files** generated by AuPosSOM.

6 Output examples

6.1 The contact map



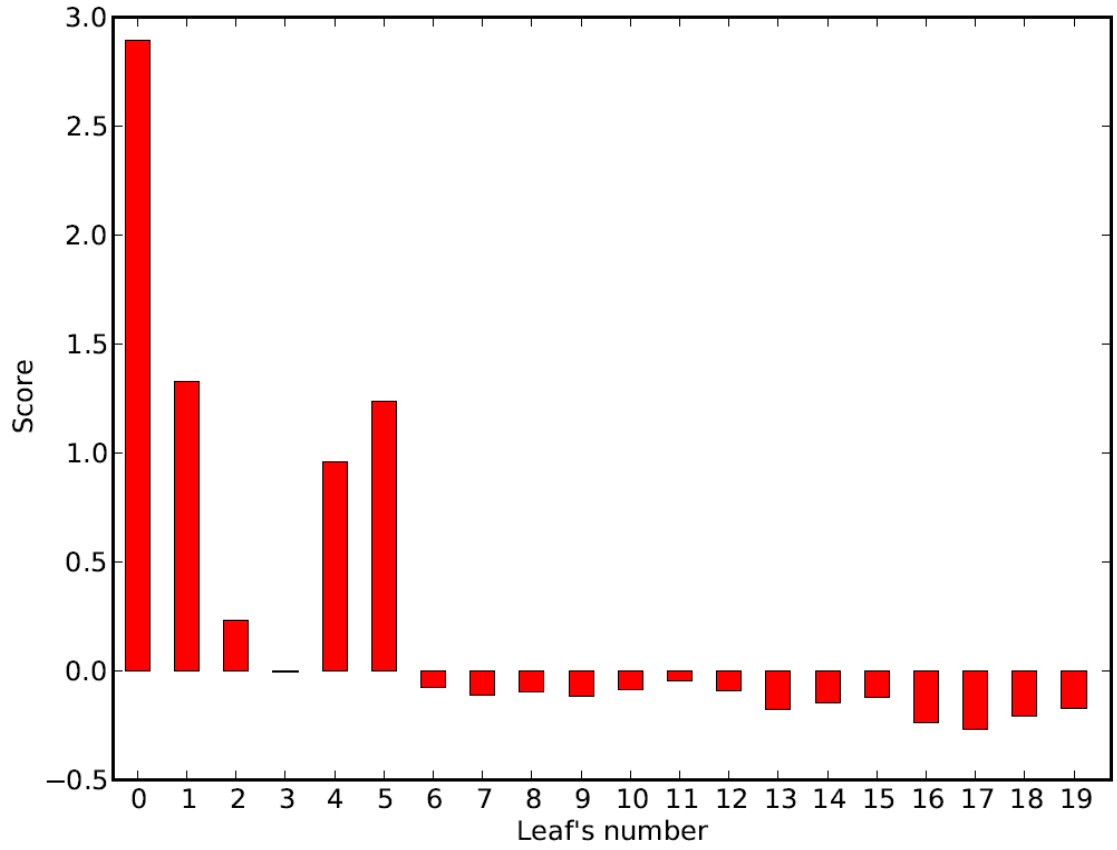
Here is an example of contact map **fingerprint2D.pdf** obtained from calculation using the templates files « *Thrombin receptor* » and the corresponding « *large sample of ligands* » available from the download section. Polar option was selected in the contact option analysis.

In this plot : each line represents a single compound footprint and each column represents a single contact (atom from the protein). Compounds sharing contact similarity are clustered inside leaves separated by white dotted lines. Leaves numbers are shown in red.

The **ALL_vectors_table** file containing the matrix results can be found in AuPosSOM_analysis_nofilt or in AuPosSOM_analysis_filt directory.

The color scale indicates the contact intensity average for all poses from the same ligand. In this scale : Dark blue color corresponds to a low intensity contact close to zero. Green color corresponds to an average intensity of one contact for a given atom from the receptor.

6.2 The scoring plot



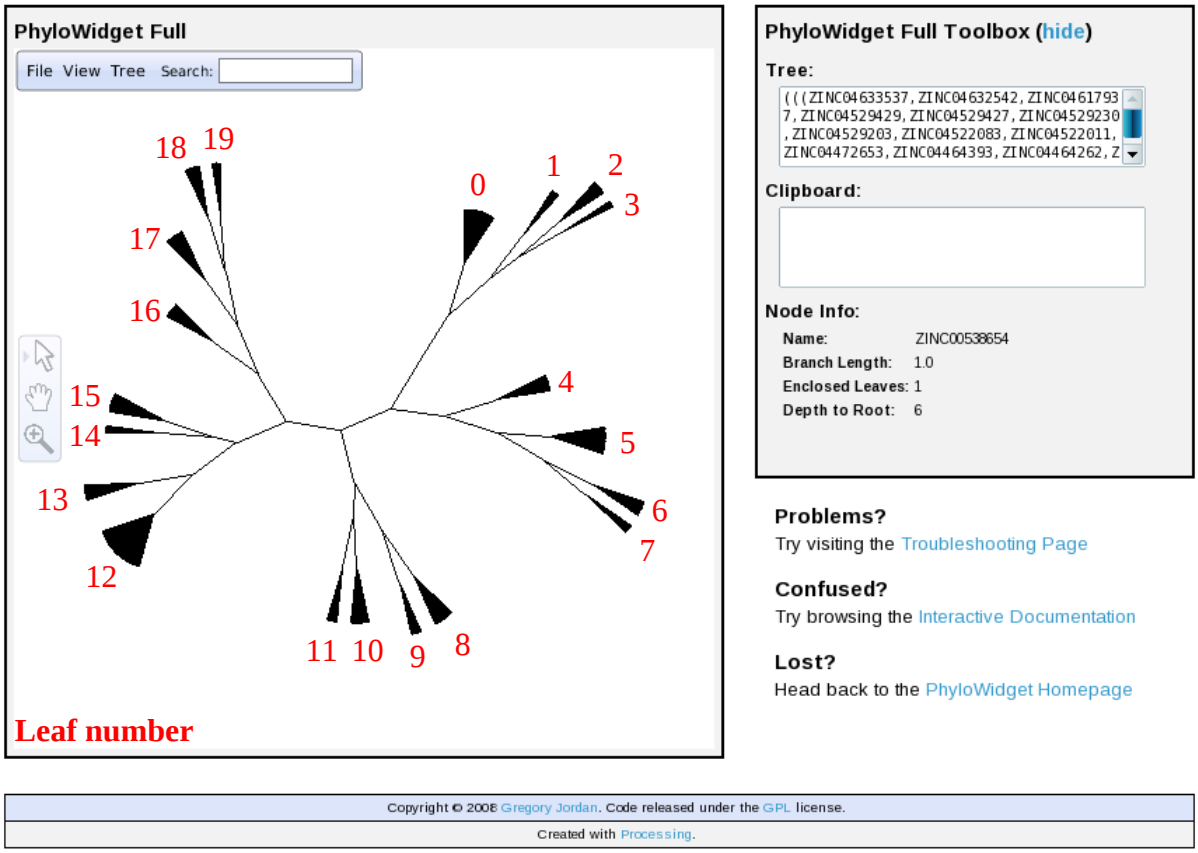
From the file **scoring0.pdf** a complementary analysis can be done.

The score corresponds to a combination between contact specificity and contact intensity for the different leaves. The scoring function was tested on different datasets containing both active compounds and decoys. It managed to successfully identify the leaf containing the highest concentration of active compounds in 7 datasets over 9. (See related publication for details).

This plot highlights the leaves having a higher probability to find active compounds (better leaves should have a higher score in the positive scale). In this example : leaf_0 is evidenced and the leaf_0 file may be submitted in the vector box for a sub-tree analysis to find contacts with increased accuracy. Alternatively, the « *ALL_vectors_table* » may be used to restart a clustering with different SOM and filtering parameters.

Advanced trick : it is possible to build a sub-tree from leaves 0, 1, 4 and 5. In order to perform a subtree, you need to merge the files leaf_0, leaf_1, leaf_4 and leaf_5 in a unique file called « active_leaves », for example (keep only one contact labels line, from the first file).

6.3 The AuPosSOM tree



Results from AuPosSOM calculation may be represented in a tree. Each branch holds several leaves containing the compounds labels with a similar contact footprint (see contact map in section 6.1)

From the online Phylowidget applet java, select :

File > Open > Load Tree > From File... > ***map0.tree*** (in AuPosSOM folders)

View > Rendering > Minimum Text Size 0.0

Select the tree format from the menu by opening :

View > Layout > Unrooted.

To obtain the figure above, tick the following options :

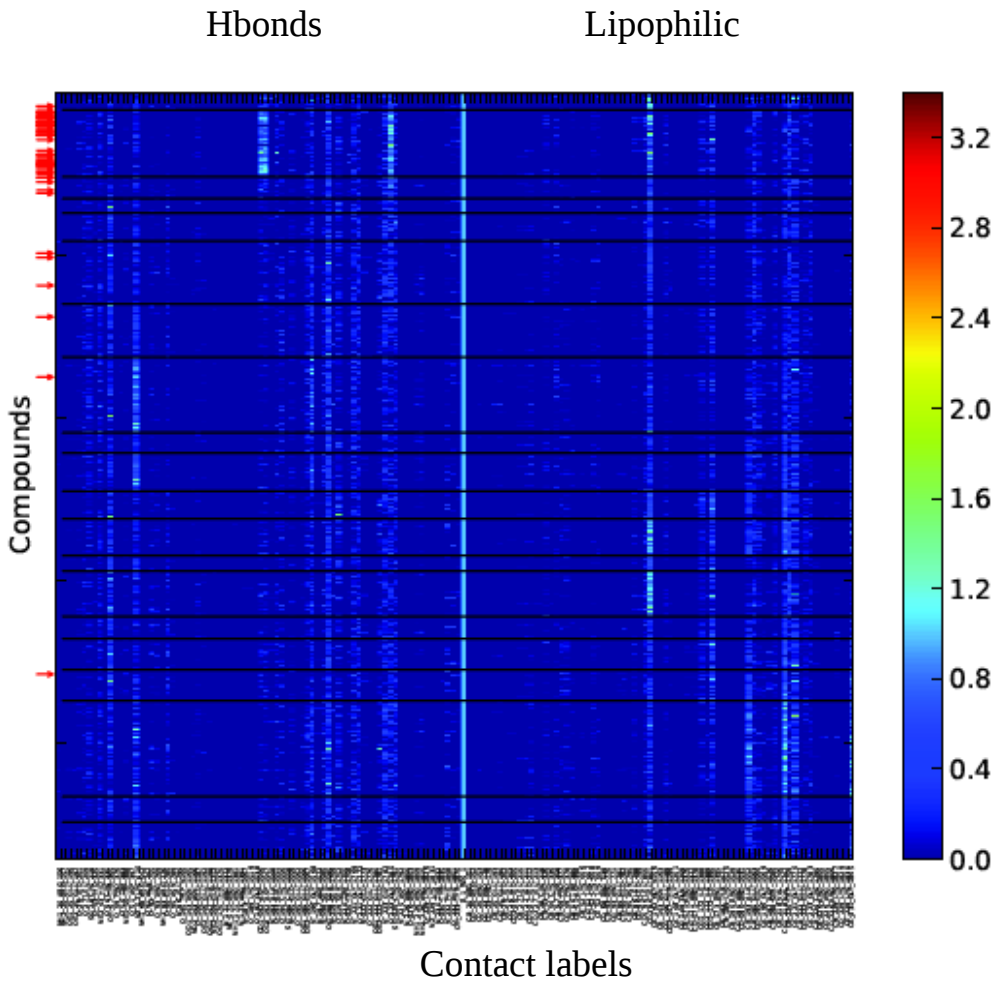
View > Show all ☑

View > Branch Lengths ☑

6.4 Advanced section : active ligand list file and vectors fusion

*Contact map analysis from vectors fusion

From the protein and the docked_ligands files, it is possible to generate different sets of vectors by selecting different contact detection. In this example, a first calculation was run using hbond contact detection and a second one using lipophilic contact detection. The two « ALL_vectors_table » files can be merged in a third calculation with the possibility to upload two vectors files.

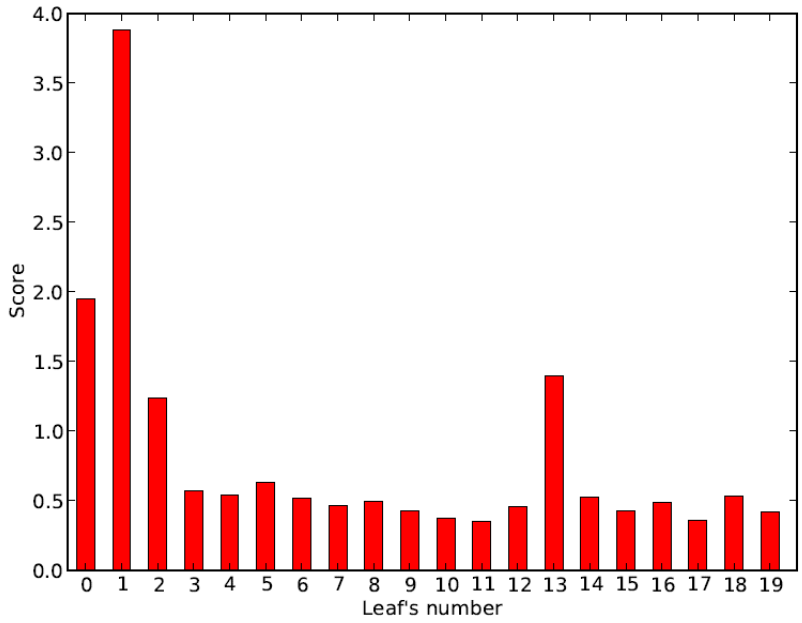


This plot is available from the file **fingerprint2D_tree_0.pdf**; The figure is similar to figure 6.1 except that contact labels are visible instead of numbers, but leaf numbers are not displayed.

The vertical line in the middle of the plot (sky blue line) is an artificial contact frontier to separate the two sets of contacts. By convention, the second vectors file is on the right.

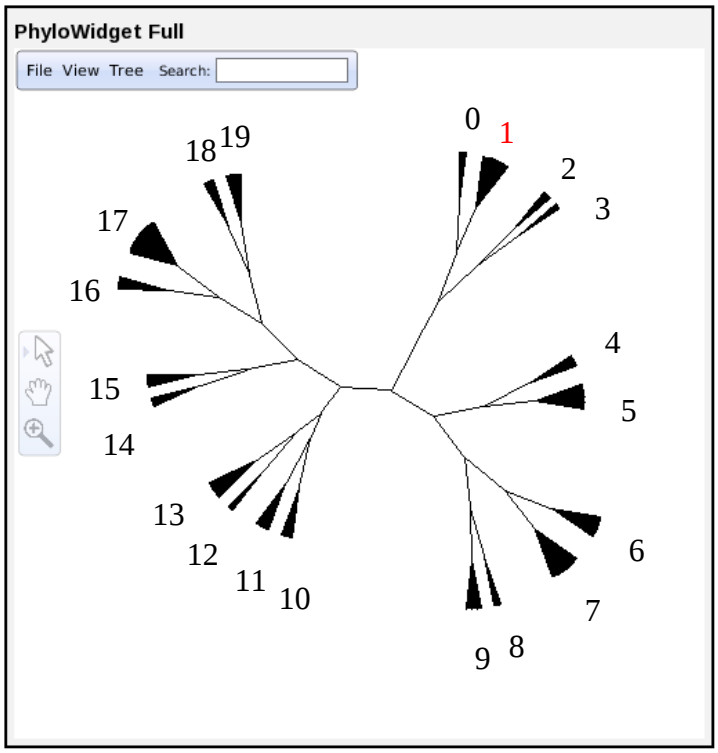
The red arrows in the figure indicate the position of active compounds (from an uploaded **active_list.txt** file) in the map.

*Corresponding scoring plot



From this plot, leaf_1 has obtained the best score about twice the second one in ranking (details figure 6.2). Moreover, leaf_1 contains a majority of active compounds (see contact map).

*The AuPosSOM tree

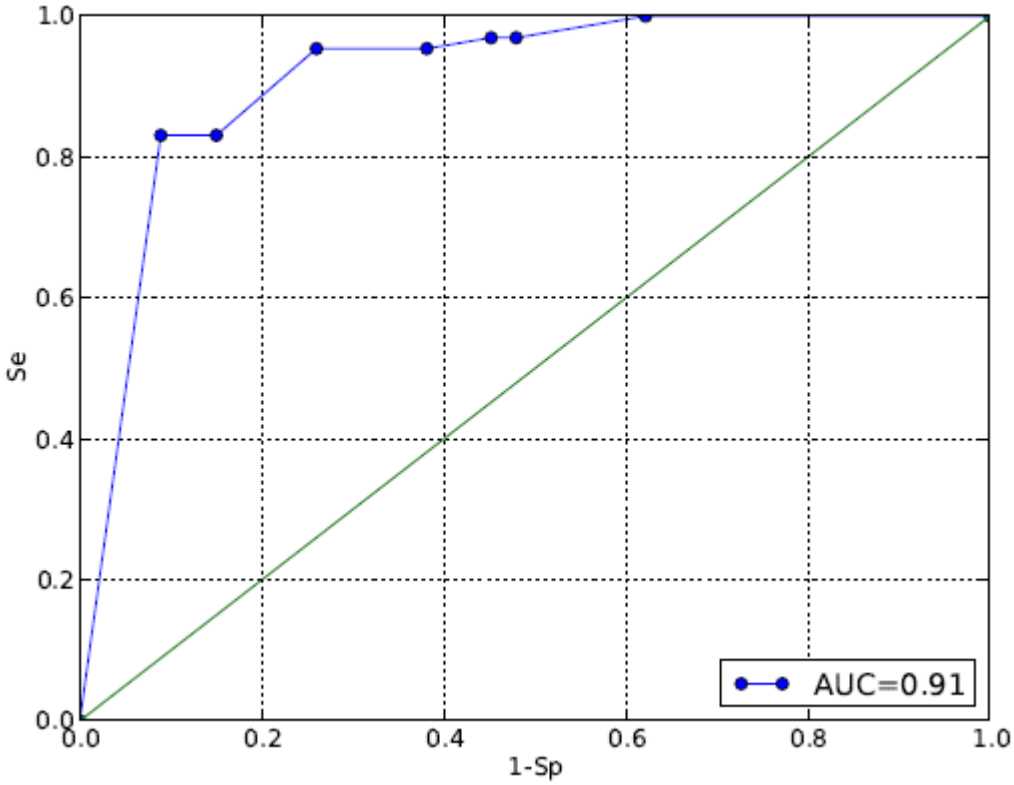


Since leaf_1 contains a high population of active compounds, other ligands in the same leaf whose activity is unknown may be good candidates for biological test.

The leaves 0, 1, 2 and 3 containing active ligands belong to the same branch from the tree. In the same idea, compounds from leaves 0, 2 and 3 have a different mode of interaction from leaf_1. These alternative modes of interaction, close from the known active binding mode, could be interesting. Population of active compounds is smaller, but number of compounds to test is also reduced.

Patterns from other leaves are significantly different. Therefore, the probability to find active ligand inside these leaves is somehow much lower.

*The ROC plot



This ROC plot (**roc0.pdf**) is representative of the clustering quality. It represents the fraction of true positives out of the positives (TPR = true positive rate) vs. the fraction of false positives out of the negatives (FPR = false positive rate), at various threshold settings. TPR is also known as sensitivity, and FPR is one minus the specificity or true negative rate. The green diagonal corresponds to the random selection.

In AuPosSOM, ROC plot is generated by selecting first the leaf with the best score, then the closest leaves in the same branch (0, 2 and 3) and successively the other leaves from other branches according to the distances from the best leaf (selection follows from the closest to the farthest leaves). In this plot, for example, there are more than 80% of total active compounds found in only 10% of the total population.

At this point, you should be able to run AuPosSOM and analyse the docking results obtained. With AuPosSOM, ligand selection can help you finding new active compounds and to progress in your investigation of the interaction of ligands with your target. Essential contacts in a contact activity relationship (CAR) approach can be found.

Nevertheless, if you still have trouble to run the program or to understand results, don't hesitate to contact the team (contact@aupossom.com). We will provide you with all our support and we will be happy to start a new collaboration.

Best regards,

The AuPosSOM Team.