COmparative Molecular Field Analysis (COMFA)

By Gijs Schaftenaar

Overview of this tutorial

Introduction

COMFA (COmparative Molecular Field Analysis) is a special case of QSAR.

QSAR (Quantitative Structure Activity Relationships) is linear regression technique based on data from known active molecules. QSAR can be applied when the 3D structure of the receptor is unknown. To apply QSAR, all that is needed are the activities, the 2D and/or 3D structures and properties/descriptors of the molecules. Of course, activities have to be measured, but 3D structures can be determined either by measurement (crystal X-ray analysis) or by calculation from the 2D diagram and (optionally) subsequent optimization.

COMFA is a QSAR approach where the descriptors consist of the Molecular Fields at each point in a 3D grid around the aligned set of 3D ligand molecules.

The aim of QSAR/COMFA is to derive a correlation between the biological activity of a set of molecules and their properties/Molecular Fields. Where some properties used with QSAR can be calculated when only the 2D structures of the molecules are available, other properties require a 3D structure. The Molecular Field used as descriptors with COMFA requires a set of 3D ligand structures, complete with hydrogens attached, aligned at their 'biologically active' conformations. Herein lies the problem with the COMFA method: since it is a ligand based method, it is preferably used at the beginning of a drug design project when no crystal structures of the target are available. Hence the 'biologically active' conformations of the ligands can only be guessed at. However what we can safely assume is that these conformations lie not too far above the optimal conformation in energy (at most 5 Kcal/mol) and that the conformations of the active ligands which have similar conformations are more likely to be valid.

How does COMFA work?

- COMFA uses a partial least-squares (PLS) analysis to predict activity from linear combinations of properties/descriptors.

What is needed before doing COMFA?

- Molecules with activities spanning about three log units of KI or IC50 values are required.
- The basic COMFA assumption is that similar molecules have similar activities, hence COMFA works best with a series of closely related molecules. As a consequence, the further the molecule, who's activity you are trying to predict with COMFA, is from the training set molecules, the worse your predicted activity will be.

In this tutorial, you will create a COMFA model and study its application.

Setup the working environment

From the Unix shell (command prompt):

- Change directory to Practicals/07_COMFA/ by typing
  ```
  cd Practicals/07_COMFA/
  ```
- and call gmolden by typing
  ```
  gmolden
  ```

The Molden Comfa interface

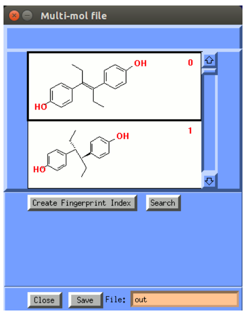
Molden provides an interface to the open-source package Open3DQSAR. Open3DQSAR uses molecular interaction fields (MIF) to calculate descriptors on each point on a 3D grid surrounding a set of pre-aligned molecules, supplied in the form of an **.sdf** file.

A series of 30 compounds with moderate to high activity for the estrogen receptor (ER-α) (*Endocrinology* **1997**, *138(9)*, 4022-4025) have been constructed and stored in the file **est+act.sdf**. This set, along with the measured activity data, will be used for training in a COMFA model.

First, we need to write out the activities stored in the **.sdf** file into a simple **.txt** file.

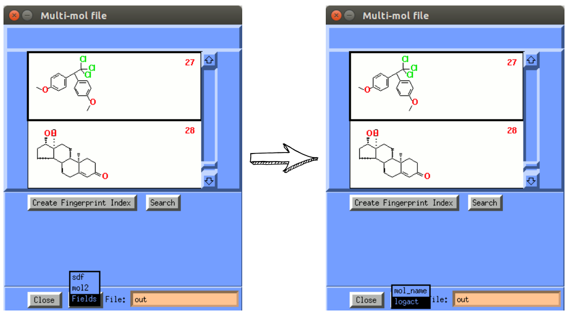
- In the molden control window, select **Read**. The *Molden File Select* window will pop up.
- Activate the *text entry field* below the **Filter:** string by clicking it. Now enter the filter **sdf** and hit the **Enter** key.
- Click the `est+act.sdf` file. The *Multi-mol file* window will pop up.

You should now have a *multi-mol file* with 30 molecules:



Take a look at the collection of molecules; You can copy a molecule from the *multi-mol file* to the molecular viewing area by single-clicking the 2D representation. Rotate the molecule by using the mouse (keep left mouse button pressed down). The majority of the molecules are steroids; do you recognize the steroid skeleton? Notice that there are also a number of non-steroidal molecules in the training set.
Write down the numbers of the non-steroidal compounds. You will need them later on in the tutorial.

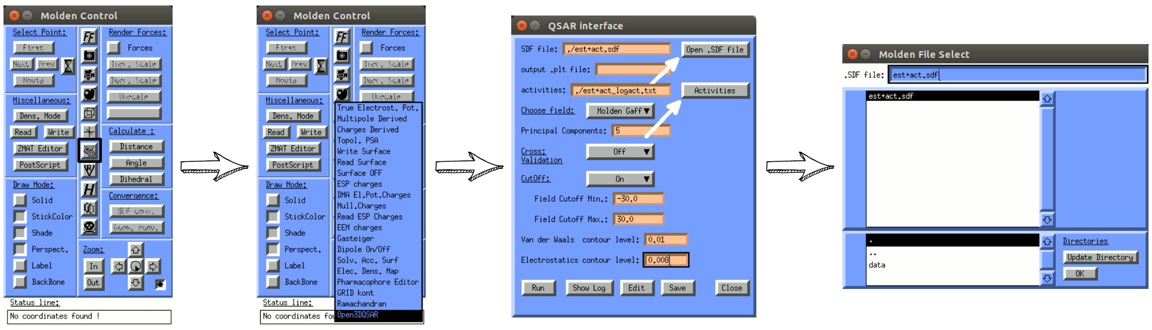
Now the activity data are exported from the *multi-mol file* to a text file, by clicking the **Save** button.:



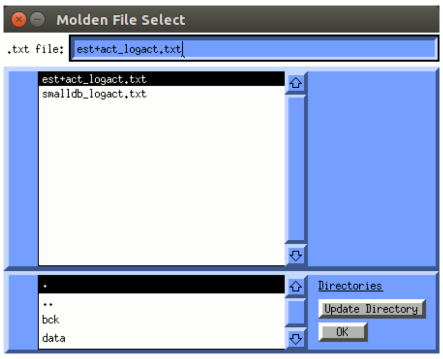
The **est+act_logact.txt** text file has as first line *Affinity* and for each structure a line with the logarithm of the activity.

The activities are relative binding affinities, expressed as (nanomolar!) concentration values. In COMFA/QSAR the logarithms of concentrations are used, similar to using pH for the acidity of a solution. So the values we'll be using are -log(concentration*10^{-9}).

You should now have a **est+act.sdf** database with 30 molecules and a corresponding file **est+act_logact.txt** with activities.
Now let us start the molden *Open3DQSAR* interface:


SDF file: ./est+act.sdf
output .plt file:
activities: ./est+act_logact.txt
Choose field: Molden Gaff
Principal Components: 5
Cross: Validation: Off
CutOff: On
Field Cutoff Min.: -30.0
Field Cutoff Max.: 30.0
Van der Waals contour level: 0.01
Electrostatics contour level: 0.008

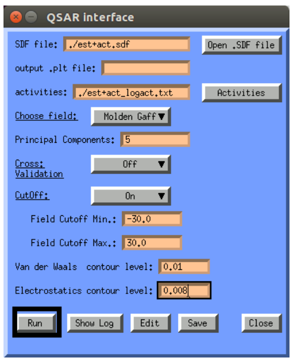
.SDF file: est+act.sdf



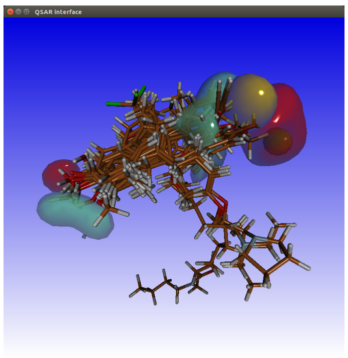
.txt file: est+act_logact.txt

```
est+act_logact.txt
smalldb_logact.txt
```

```
.
..
bck
data
```



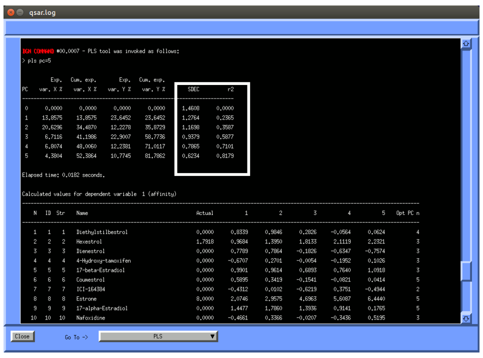
SDF file: ./est+act.sdf
output .plt file:
activities: ./est+act_logact.txt
Choose field: Molden Gaff
Principal Components: 5
Cross: Validation: Off
CutOff: On
Field Cutoff Min.: -30.0
Field Cutoff Max.: 30.0
Van der Waals contour level: 0.01
Electrostatics contour level: 0.008



For clarity's sake make the surfaces transparent by typing a **t** character in the display screen.

You will have to adjust the default values for the van der Waals and Electrostatics contour levels to **0.01** and **0.008** respectively. The higher these values are the more compacted the surfaces become. Higher contour values also mean greater correlation to the activity. The light blue and red surfaces show the correlation of the electrostatics part of the molecular field and the activity. These surfaces concentrate on the two OH groups which are essential for activity. Green contours indicate regions where a relatively bulkier substitution would increase the activity, whereas the yellow contours indicate areas where a bulkier substituent would decrease the activity.
The regions where increasing the positive charges would increase the activity are in blue and regions where increasing the negative charges increases the activity are in red.

Hit the **Show Log** button to view the contents of the *qsar.log* file. This will open the *qsar.log* window. In this window select the **PLS** option to set the viewport to the PLS part of the file.



```
BGN COMMAND #00.0007 - PLS tool was invoked as follows:
> pls pc=5


        Exp.    Cum. exp.       Exp.    Cum. exp.
PC    var. X %   var. X %    var. Y %   var. Y %       SDEC         r2
-------------------------------------------------------------------------
 0      0.0000     0.0000      0.0000     0.0000      1.4608      0.0000
 1     13.8575    13.8575     23.6452    23.6452      1.2764      0.2365
 2     20.6296    34.4870     12.2278    35.8729      1.1698      0.3587
 3      6.7116    41.1986     22.9007    58.7736      0.9379      0.5877
 4      6.8074    48.0060     12.2381    71.0117      0.7865      0.7101
 5      4.3804    52.3864     10.7745    81.7862      0.6234      0.8179

Elapsed time: 0.0182 seconds.


Calculated values for dependent variable  1 (affinity)
------------------------------------------------------------------------------------------------------------------
   N   ID  Str   Name                                Actual          1          2          3          4          5   Opt PC n
------------------------------------------------------------------------------------------------------------------
   1    1    1   Diethylstilbestrol                  0.0000     0.8339     0.9846     0.2826    -0.0564     0.0624          4
   2    2    2   Hexestrol                           1.7918     0.9684     1.3950     1.8133     2.1119     2.2321          3
   3    3    3   Dienestrol                          0.0000     0.7789     0.7864    -0.1826    -0.6347    -0.7574          3
   4    4    4   4-Hydroxy-tamoxifen                 0.0000    -0.6707     0.2701    -0.0054    -0.1952     0.1026          3
   5    5    5   17-beta-Estradiol                   0.0000     0.9901     0.9614     0.6893     0.7640     1.0918          3
   6    6    6   Coumestrol                          0.0000     0.5895     0.3419    -0.1541    -0.0821     0.0414          5
   7    7    7   ICI-164384                          0.0000    -0.4312     0.0102    -0.6219     0.3751    -0.4944          2
   8    8    8   Estrone                             8.0000     2.0746     2.9575     4.6963     5.6087     6.4440          5
   9    9    9   17-alpha-Estradiol                  0.0000     1.4477     1.7860     1.3936     0.9141     0.1765          5
  10   10   10   Nafoxidine                          0.0000    -0.4661     0.3366    -0.0207    -0.3436     0.5195          3
```

In the *Open3DQSAR* log file you will find **SDEC** (**S**tandard **D**eviation of **E**rror of **C**alculation) and $\mathbf{r^2}$

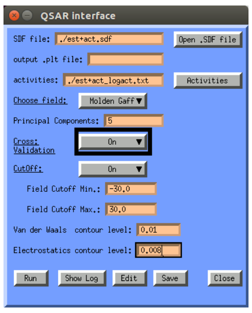
Perform a cross-validated PLS analysis

In a partial least-squares (PLS) analysis, two factors are important: the number of components used in the regression equation (which will be dealt with later) and the (usually squared) correlation coefficient. A *non*-cross-validated PLS analysis gives a squared correlation coefficient usually indicated by r^2. This number, which is also used in (multiple) linear regression, is between zero and one and expresses the quality of the PLS analysis. It indicates the proportion of the variation in the dependent variable (here the activity) that is explained by the regression equation and its value should be as close to one as possible. However, r^2 expresses the quality of the data fit rather than the quality of prediction (which is what we are actually interested in).

To express the predictive power of the analysis, the **cross-validated** $\mathbf{r^2}$, usually indicated by $\mathbf{q^2}$, is used. In cross-validation, one value is left out, a model is derived using the remaining data, and the model is used to predict the value originally left out. This procedure is repeated for all values, yielding q^2. q^2 is normally (much) lower than r^2 and values greater than 0.5 already indicate significant predictive power.

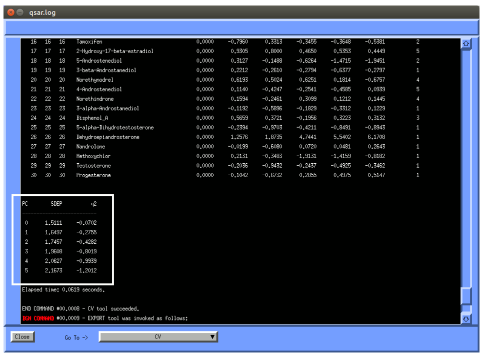
After the cross-validated PLS analysis, you will determine the optimal number of components. Recall that the components are linear combinations of the variables (of which you have very many!), ordered in such a way that the first component will describe most of the variation in the activity, the second most of the remaining variation, etc. The cross-validated PLS analysis will be carried out for different numbers of components. As a rule of thumb, the number of components should not exceed one-third of the number of molecules. More components would lead to a model that is *overtrained* - it has a better fit to the training data but the predictive power is diminished.

To perform a cross-validated PLS, we switch the *Cross Validation* button to **On**.



SDF file: ./est+act.sdf
output .plt file:
activities: ./est+act_logact.txt
Choose field: Molden Gaff
Principal Components: 5
Cross: Validation: On
CutOff: On
Field Cutoff Min.: -30.0
Field Cutoff Max.: 30.0
Van der Waals contour level: 0.01
Electrostatics contour level: 0.008

Hit the **Show Log** button to view the contents of the *qsar.log* file. This will open the *qsar.log* window. In this window select the **CV** option to set the viewport to the cross validation part of the file.



```
  16   16   16   Tamoxifen                           0.0000    -0.7960     0.3313    -0.3455    -0.3648    -0.5381          2
  17   17   17   2-Hydroxy-17-beta-estradiol         0.0000     0.9305     0.8000     0.4650     0.5353     0.4449          5
  18   18   18   5-Androstenediol                    0.0000     0.3127    -0.1488    -0.6264    -1.4715    -1.9451          2
  19   19   19   3-beta-Androstanediol               0.0000     0.2212    -0.2610    -0.2794    -0.6377    -0.2797          1
  20   20   20   Norethynodrel                       0.0000     0.6193     0.5024     0.6251     0.1814    -0.6757          4
  21   21   21   4-Androstenediol                    0.0000     0.1140    -0.4247    -0.2541    -0.4585     0.0939          5
  22   22   22   Norethindrone                       0.0000     0.1594    -0.2461     0.3099     0.1212     0.1445          4
  23   23   23   3-alpha-Androstanediol              0.0000    -0.1192    -0.5896    -0.1829    -0.3312     0.1229          1
  24   24   24   Bisphenol_A                         0.0000     0.5659     0.3721    -0.1956     0.3223     0.3132          3
  25   25   25   5-alpha-Dihydrotestosterone         0.0000    -0.2394    -0.9703    -0.4211    -0.8491    -0.8943          1
  26   26   26   Dehydroepiandrosterone              0.0000     1.2576     1.8735     4.7441     5.5402     6.1708          1
  27   27   27   Nandrolone                          0.0000    -0.0199    -0.6080     0.0720     0.0481     0.2643          1
  28   28   28   Methoxychlor                        0.0000     0.2131    -0.3483    -1.9131    -1.4159    -0.8182          1
  29   29   29   Testosterone                        0.0000    -0.2036    -0.9432    -0.2437    -0.4925    -0.3462          1
  30   30   30   Progesterone                        0.0000    -0.1042    -0.6732     0.2855     0.4975     0.5147          1


PC        SDEP          q2
--------------------------
 0      1.5111     -0.0702
 1      1.6497     -0.2755
 2      1.7457     -0.4282
 3      1.9608     -0.8019
 4      2.0627     -0.9939
 5      2.1673     -1.2012

Elapsed time: 0.0619 seconds.

END COMMAND #00.0008 - CV tool succeeded.
BGN COMMAND #00.0009 - EXPORT tool was invoked as follows:
```

- In the *OPEN3DQSAR* log file you will find the Standard Error of Prediction (**SEP**) and $\mathbf{q^2}$.

Determine the optimal number of components

The two important factors here are the **standard error of prediction** (SEP) and the **cross-validated** $\mathbf{r^2}$ (usually called q^2). SEP should be as low as possible and q^2 should be larger than 0.5 (lower values indicate poor predictive power).

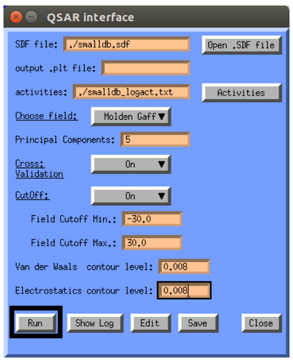
- Find the number of components having the lowest SEP and note the corresponding q^2.
- To accomplish this, in the *QSAR nterface window*, specify the number of components under **Principal Components:**.
- Check whether an additional component increases q^2 with more than 5%.
- If so, the larger number should be used; otherwise, take the lower (original) number of components.

The resulting number of components will be used in the non-cross-validated analysis.

You will find that you can not get the $\mathbf{q^2}$ above 0.5 with this combination of dataset and descriptors. In fact the $\mathbf{q^2}$'s are negative !
We will try to get the $\mathbf{q^2}$ above 0.5 by switching to a more focussed dataset: we will remove the non-steroidal compounds from the dataset.
Remember you wrote down the names of the non-steroidal compounds ? Now remove all non-steroidal compounds from the dataset.

- Go through the list of entries in the **est+act.sdf** file by hiting the *arrow down key* and hit the *Delete key* on any non-steroidal compound. When done deleting, click with the left mouse button, the *File:* text entry field and type **smalldb**.
  Next hit the **Save** button and select the *sdf* option.
  Next, click the **Save** button again, but select the *Fields* option. The **smalldb_logact.txt** file will have been written.

Now repeat the steps of COMFA model creation and cross-validation for the steroidal dataset and determine the optimal number of components. You will have to adjust the default values for the van der Waals and Electrostatics contour levels to **0.008** and **0.008** respectively.



SDF file: ./smalldb.sdf

output .plt file:

activities: ./smalldb_logact.txt

Choose field: Molden Gaff

Principal Components: 5

Cross: Validation: On

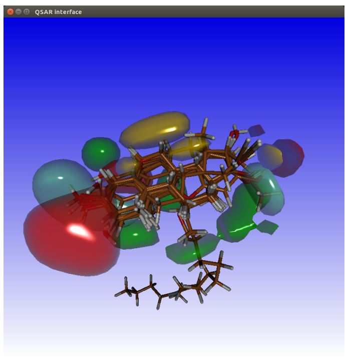
CutOff: On

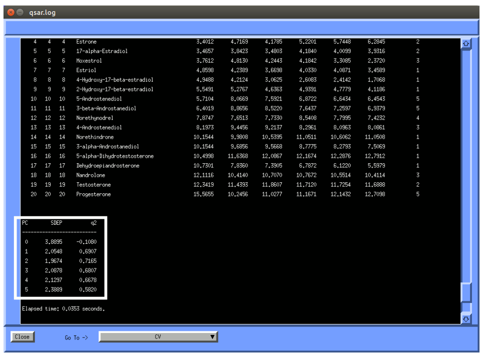
Field Cutoff Min.: -30.0

Field Cutoff Max.: 30.0

Van der Waals contour level: 0.008

Electrostatics contour level: 0.008





```
   4   4   4   Estrone                           3.4012   4.7169   4.1785   5.2201   5.7448   6.2845      2
   5   5   5   17-alpha-Estradiol                3.4657   3.8423   3.4803   4.1840   4.0099   3.9316      2
   6   6   6   Moxestrol                         3.7612   4.8130   4.2443   4.1842   3.3085   2.3720      3
   7   7   7   Estriol                           4.8598   4.2389   3.6698   4.0330   4.0871   3.4589      1
   8   8   8   4-Hydroxy-17-beta-estradiol       4.9488   4.2124   3.0625   2.6083   2.4142   1.7068      1
   9   9   9   2-Hydroxy-17-beta-estradiol       5.5491   5.2767   4.6363   4.9391   4.7779   4.1186      1
  10  10  10   5-Androstenediol                  5.7104   8.0669   7.5921   6.8722   6.6434   6.4543      5
  11  11  11   3-beta-Androstanediol             6.4019   8.8656   8.5220   7.6437   7.2597   6.9379      5
  12  12  12   Norethynodrel                     7.8747   7.6513   7.7330   8.5408   7.7995   7.4232      4
  13  13  13   4-Androstenediol                  8.1973   9.4456   9.2137   8.2961   8.0963   8.0861      3
  14  14  14   Norethindrone                    10.1544   9.9808  10.5395  11.0511  10.6062  11.0508      1
  15  15  15   3-alpha-Androstanediol           10.1544   9.6856   9.5668   8.7775   8.2793   7.5069      1
  16  16  16   5-alpha-Dihydrotestosterone      10.4998  11.6368  12.0867  12.1674  12.2876  12.7912      1
  17  17  17   Dehydroepiandrosterone           10.7301   7.8360   7.3905   6.7872   6.1220   5.5979      1
  18  18  18   Nandrolone                       12.1116  10.4140  10.7070  10.7672  10.5514  10.4114      3
  19  19  19   Testosterone                     12.3419  11.4393  11.8607  11.7120  11.7254  11.6888      2
  20  20  20   Progesterone                     15.5655  10.2456  11.0277  11.1671  12.1432  12.7098      5


PC      SDEP        q2
--------------------------
 0     3.8895    -0.1080
 1     2.0548     0.6907
 2     1.9674     0.7165
 3     2.0878     0.6807
 4     2.1297     0.6678
 5     2.3889     0.5820

Elapsed time: 0.0353 seconds.
```

Determine the optimal descriptors and components

By varying the number of components it is possible to get $\mathbf{q^2}$ as high as 0.717.