Quantitative Structure Activity Relationships (QSAR)

By Gijs Schaftenaar

Overview of this tutorial

Introduction

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.

The aim of QSAR is to derive a correlation between the biological activity of a set of molecules and their properties. Some properties can be calculated when only the 2D structure of the molecules is available, while other require a 3D structure. An additional requirement can be the presence of attached hydrogen atoms.

How does QSAR work?

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

What is needed before doing QSAR?

- Molecules with activities spanning about three log units of KI or IC50 values are required.
- The basic QSAR assumption is that similar molecules have similar activities, hence QSAR 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 QSAR, is from the training set molecules, the worse your predicted activity will be.

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

Setup the working environment

From the Unix shell (command prompt):

- Change directory to Practicals/07_QSAR/ by typing
  ```
  cd Practicals/07_QSAR/
  ```
- and call MOE by typing
  ```
  moe
  ```

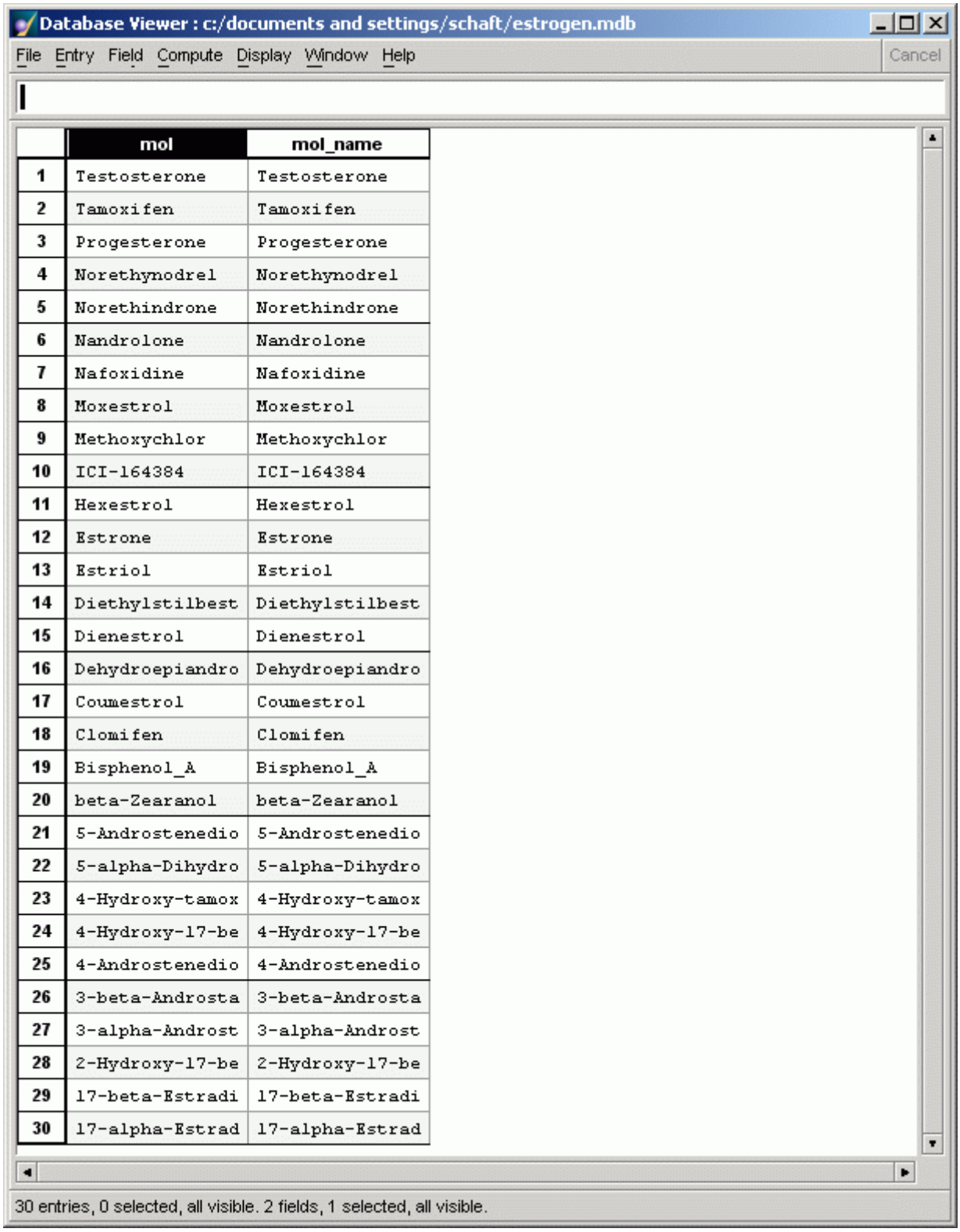
Create a molecular database

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. This set, along with the measured activity data, will be used for training in a QSAR model.

First, create a molecular database (MDB) and fill it with molecules:

- Select **File** >> **New** >> **Database**.
- Type **`estrog_act.mdb`** for the database name.
  The *Database Viewer* will pop up.
- In the *Database Viewer window*: **File** >> **Import**.
  The *Database Import window* will pop up.
- In the *Database Import window*: **Add** >> from the directory **data** select the file: **estrogen_act.mol2** >> **Add** >> **OK**.
- Back in the Database Viewer window: select **estrogen_act.mol2** by clicking on it >> **OK**.
- In the Database Viewer window: select the **mol** field by clicking on it >> **Compute** >> **Molecule** >> **Molecule Names** >> Molecule Names: Outputfield: **mol_name**.

You should now have a database with 30 molecules:



	mol	mol_name
1	Testosterone	Testosterone
2	Tamoxifen	Tamoxifen
3	Progesterone	Progesterone
4	Norethynodrel	Norethynodrel
5	Norethindrone	Norethindrone
6	Nandrolone	Nandrolone
7	Nafoxidine	Nafoxidine
8	Moxestrol	Moxestrol
9	Methoxychlor	Methoxychlor
10	ICI-164384	ICI-164384
11	Hexestrol	Hexestrol
12	Estrone	Estrone
13	Estriol	Estriol
14	Diethylstilbest	Diethylstilbest
15	Dienestrol	Dienestrol
16	Dehydroepiandro	Dehydroepiandro
17	Coumestrol	Coumestrol
18	Clomifen	Clomifen
19	Bisphenol_A	Bisphenol_A
20	beta-Zearanol	beta-Zearanol
21	5-Androstenedio	5-Androstenedio
22	5-alpha-Dihydro	5-alpha-Dihydro
23	4-Hydroxy-tamox	4-Hydroxy-tamox
24	4-Hydroxy-17-be	4-Hydroxy-17-be
25	4-Androstenedio	4-Androstenedio
26	3-beta-Androsta	3-beta-Androsta
27	3-alpha-Androst	3-alpha-Androst
28	2-Hydroxy-17-be	2-Hydroxy-17-be
29	17-beta-Estradi	17-beta-Estradi
30	17-alpha-Estrad	17-alpha-Estrad

Take a look at the collection of molecules; You can copy a molecule from the database to the molecular viewing area by double-clicking the name field and subsequently clicking OK in the window that pops up. Rotate the molecule by using the mouse (keep middle 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.

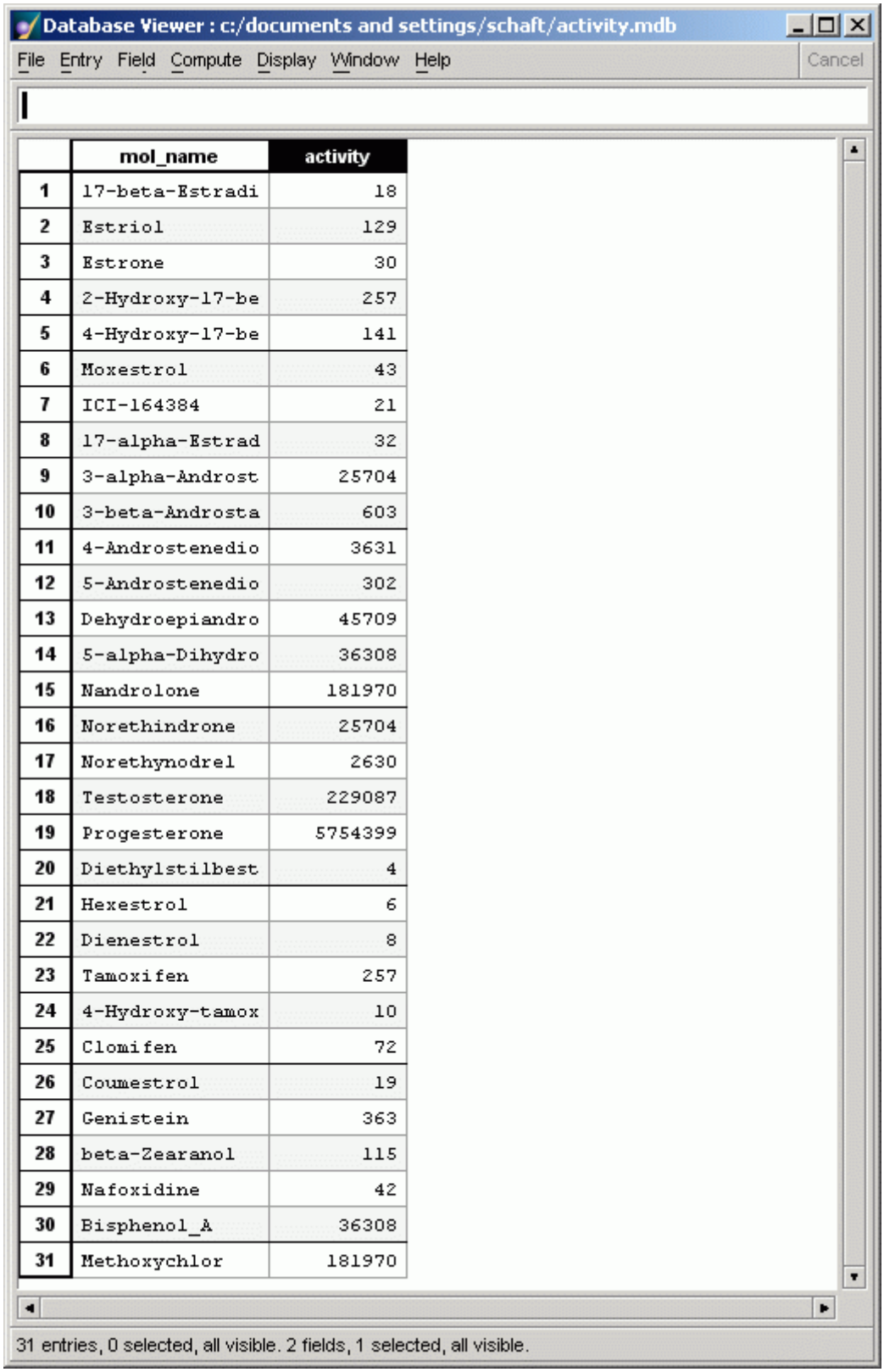
Write down the names of the non-steroidal compounds. You will need them later on in the tutorial.

PS: Alternatively, you can enlarge the size of the **mol** field by clicking on the **mol** field of the first entry and drag the mouse down while holding the mouse button pressed. In this way the contents of the field will be replaced by a 2D picture of the molecules.

Now the activity data are imported from a text file into the spreadsheet. The text file is in a simple, space-delimited format.
However, importing the activities is a bit cumbersome in Moe. To accomplish this we will a use trick. We will first create a second database containing the molecule name and the molecular activities.
Subsequently we will merge these two databases:

- From the main Moe window **File** >> **New** >> **Database**.
- Type **`activity.mdb`** for the database name.
  The *Database Viewer* will pop up.
- In the *Database Viewer window*: **File** >> **Import**.
  The *Database Import window* will pop up.
- In the *Database Import window*: **Add** >> from directory **data** select the file:
  activity.txt >> **Add** >> **OK**
  back the *Database Import window*: Import type: **ascii** >> **Add** >> **OK**.
- Back in the Database Viewer window: select **estrogen_act.txt** by clicking on it >> **OK**.
- In the Database Viewer window: select the **mol** field by clicking on it >> right click (Third mouse button) >> select **Delete** >> from the drop-down menu >> **OK**.
- In the Database Viewer window: click on **Field_1** >> click with 3th mouse button in field >> **Rename** >> **Rename field "Field_1" to:** >> **mol_name**
- In the Database Viewer window: click on **Field_2** >> click with 3th mouse button in field >> **Rename** >> **Rename field "Field_2" to:** >> **activity**

You should now have a second database with 30 molecule names and corresponding activities:



	mol_name	activity
1	17-beta-Estradi	18
2	Estriol	129
3	Estrone	30
4	2-Hydroxy-17-be	257
5	4-Hydroxy-17-be	141
6	Moxestrol	43
7	ICI-164384	21
8	17-alpha-Estrad	32
9	3-alpha-Androst	25704
10	3-beta-Androsta	603
11	4-Androstenedio	3631
12	5-Androstenedio	302
13	Dehydroepiandro	45709
14	5-alpha-Dihydro	36308
15	Nandrolone	181970
16	Norethindrone	25704
17	Norethynodrel	2630
18	Testosterone	229087
19	Progesterone	5754399
20	Diethylstilbest	4
21	Hexestrol	6
22	Dienestrol	8
23	Tamoxifen	257
24	4-Hydroxy-tamox	10
25	Clomifen	72
26	Coumestrol	19
27	Genistein	363
28	beta-Zearanol	115
29	Nafoxidine	42
30	Bisphenol_A	36308
31	Methoxychlor	181970

Now lets merge these two databases:

- In the Database Viewer of estrogen.mdb: **File** >> **Merge**
- In the Merge Databases window: **Input Database 2:** click **Browse** >> select **activity.mdb** >> **OK** >> deselect: **New Database** >> **Next**

- Key Field Specification: >> select *Database1:,Key1:* **mol_name** and *Database2:,Key1:* **mol_name** >> **Next** >> **Next**
- Merging Options: keep *import shared entries* checked and uncheck the other options >> **Merge** >> **Close**

You should now have a database estrogen.mdb with 30 molecular structures, molecule names and molecular activities

Database Viewer : c:/documents and settings/schaft/estrogen.mdb

File Entry Field Compute Display Window Help Cancel

	mol	mol_name	activity
1	Testosterone	Testosterone	229087
2	Tamoxifen	Tamoxifen	257
3	Progesterone	Progesterone	5754399
4	Norethynodrel	Norethynodrel	2630
5	Norethindrone	Norethindrone	25704
6	Nandrolone	Nandrolone	181970
7	Nafoxidine	Nafoxidine	42
8	Moxestrol	Moxestrol	43
9	Methoxychlor	Methoxychlor	181970
10	ICI-164384	ICI-164384	21
11	Hexestrol	Hexestrol	6
12	Estrone	Estrone	30
13	Estriol	Estriol	129
14	Diethylstilbest	Diethylstilbest	4
15	Dienestrol	Dienestrol	8
16	Dehydroepiandro	Dehydroepiandro	45709
17	Coumestrol	Coumestrol	19
18	Clomifen	Clomifen	72
19	Bisphenol_A	Bisphenol_A	36308
20	beta-Zearanol	beta-Zearanol	115
21	5-Androstenedio	5-Androstenedio	302
22	5-alpha-Dihydro	5-alpha-Dihydro	36308
23	4-Hydroxy-tamox	4-Hydroxy-tamox	10
24	4-Hydroxy-17-be	4-Hydroxy-17-be	141
25	4-Androstenedio	4-Androstenedio	3631
26	3-beta-Androsta	3-beta-Androsta	603
27	3-alpha-Androst	3-alpha-Androst	25704
28	2-Hydroxy-17-be	2-Hydroxy-17-be	257
29	17-beta-Estradi	17-beta-Estradi	18
30	17-alpha-Estrad	17-alpha-Estrad	32

30 entries, 0 selected, all visible. 3 fields, 1 selected, all visible.

The activities are relative binding affinities, expressed as (nanomolar!) concentration values. In 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}).

We will now transform the activity to the logarithm of the activity:

- Select the **activity** by clicking on it >> **Compute** >> **Calculator**
- In the Molecular Database Calculator: click the **log** button >> In *Available Fields* double click the **activity** field >> in *Destination Field* replace result_1 by **logact** >> **Evaluate** >> **Close**
- Now we no longer need the *activity* field, delete it by clicking on it >> click with right mouse button >> **Delete** >> **OK**
- We can sort on activity: by clicking the **logact** field >> click with right mouse button >> **Sort** >> **Descending** >> **OK**

Adding molecular descriptors to the database

Adding molecular descriptors the molecular spreadsheet is quite straightforward:

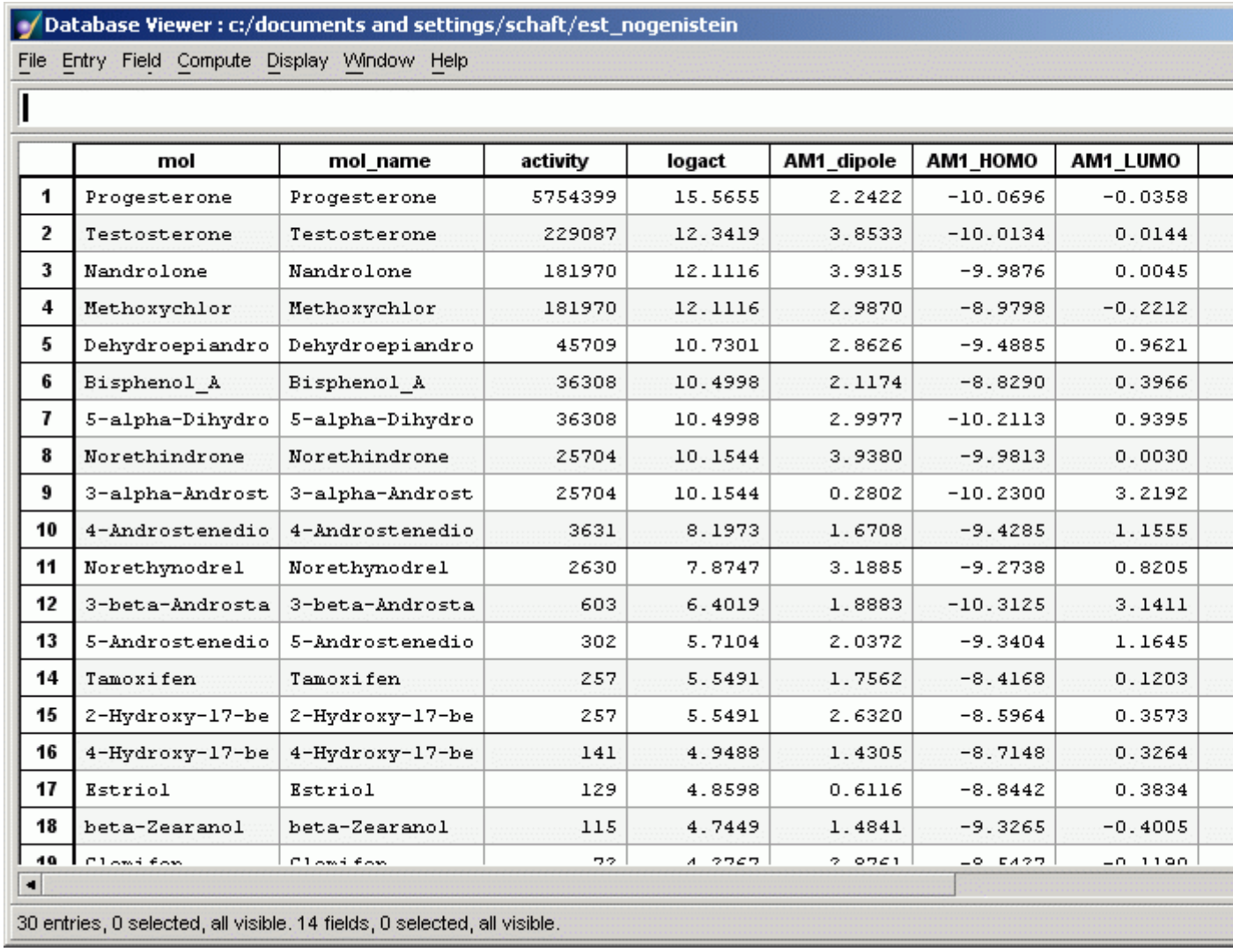
- In the Database Viewer: >> **Compute** >> **Descriptors...** >>
- In QuaSar-Descriptor: select descriptors of the type: **2D** or **i3D**.
- Be sure to include: **AM1_dipole, AM1_HOMO, AM1_LUMO, ASA_P, a_acc, a_don, logP(o/w)** and **logS**.
- **DO NOT CLICK OK** but instead click **Cancel**.

The calculation of particulary the semi-empirical descriptors (AM1_*) will take too long, typically 15 minutes till half an hour.
Instead we will switch to a database with precalculated descriptors.

Close the database viewer of **estrogen.mdb** and open the database with the precalculated descriptors:

- In the database viewer of **estrogen.mdb**: >> **File** >> **Quit**
- In the main Moe window: **File** >> **Open** >> from the directory **data** select **est+descriptors** >> **OK**

Your spreadsheet window should now look like this:


Database Viewer : c:/documents and settings/schaft/est_nogenistein

File Entry Field Compute Display Window Help

	mol	mol_name	activity	logact	AM1_dipole	AM1_HOMO	AM1_LUMO
1	Progesterone	Progesterone	5754399	15.5655	2.2422	-10.0696	-0.0358
2	Testosterone	Testosterone	229087	12.3419	3.8533	-10.0134	0.0144
3	Nandrolone	Nandrolone	181970	12.1116	3.9315	-9.9876	0.0045
4	Methoxychlor	Methoxychlor	181970	12.1116	2.9870	-8.9798	-0.2212
5	Dehydroepiandro	Dehydroepiandro	45709	10.7301	2.8626	-9.4885	0.9621
6	Bisphenol_A	Bisphenol_A	36308	10.4998	2.1174	-8.8290	0.3966
7	5-alpha-Dihydro	5-alpha-Dihydro	36308	10.4998	2.9977	-10.2113	0.9395
8	Norethindrone	Norethindrone	25704	10.1544	3.9380	-9.9813	0.0030
9	3-alpha-Androst	3-alpha-Androst	25704	10.1544	0.2802	-10.2300	3.2192
10	4-Androstenedio	4-Androstenedio	3631	8.1973	1.6708	-9.4285	1.1555
11	Norethynodrel	Norethynodrel	2630	7.8747	3.1885	-9.2738	0.8205
12	3-beta-Androsta	3-beta-Androsta	603	6.4019	1.8883	-10.3125	3.1411
13	5-Androstenedio	5-Androstenedio	302	5.7104	2.0372	-9.3404	1.1645
14	Tamoxifen	Tamoxifen	257	5.5491	1.7562	-8.4168	0.1203
15	2-Hydroxy-17-be	2-Hydroxy-17-be	257	5.5491	2.6320	-8.5964	0.3573
16	4-Hydroxy-17-be	4-Hydroxy-17-be	141	4.9488	1.4305	-8.7148	0.3264
17	Estriol	Estriol	129	4.8598	0.6116	-8.8442	0.3834
18	beta-Zearanol	beta-Zearanol	115	4.7449	1.4841	-9.3265	-0.4005
19	Clomifen	Clomifen	72	4.2767	2.8761	-8.5437	-0.1190

30 entries, 0 selected, all visible. 14 fields, 0 selected, all visible.

Now have a look at the correlation between the separate descriptors and logact:

- In the Database Viewer: **Compute** >> **Analysis** >> **Correlation Plot** >> click the **logact** field (X-axis) >> click one of the other fields for the Y-axis.

The quality of the correlation is reflected in the printed correlation coefficient **R**, the closer to 1.0 the better.

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

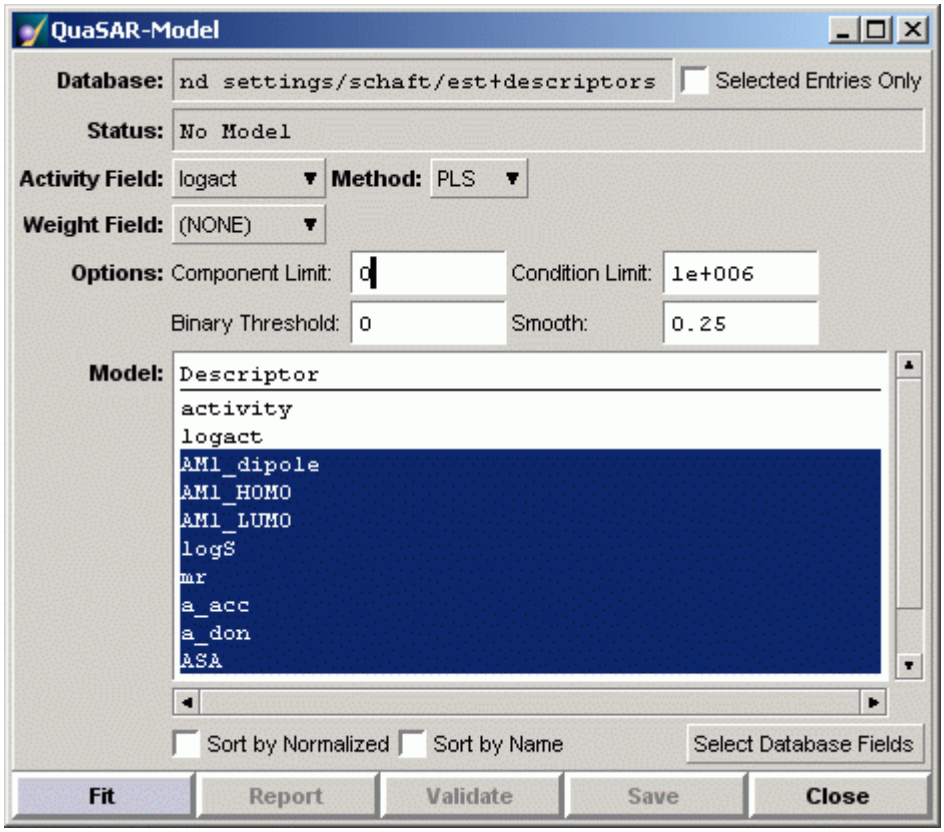
expresses the quality of the data fit rather than the quality of prediction (which is what we are actually interested in).

To express the predictive power of the analysis, the **cross-validated r^2**, usually indicated by **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.

First we will create a QSAR model:

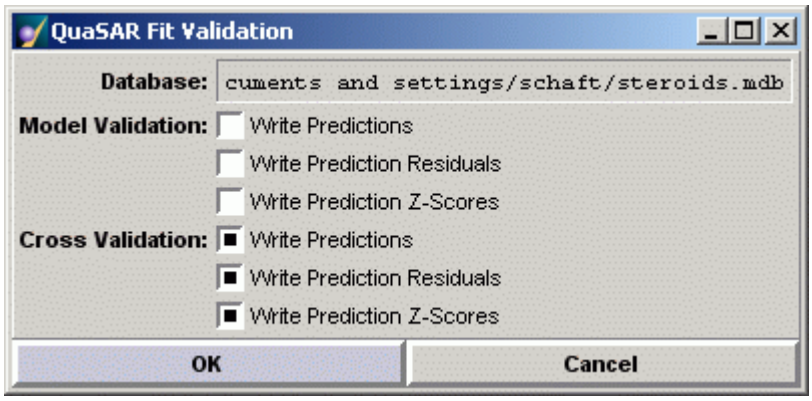
- In the Database Viewer >> **Compute** >> **Model** >> **QuaSAR** >> Activity Field: **logact**
- Select all fields in the Descriptor list box (**Model:**): click the topmost field, scroll down to the bottom field in the list and click it while holding down the SHIFT key. Subsequently deselect the fields activity and logact by clicking them.
- For now we will ignore the number of components and leave it at its default value; *Component Limit:* **0**, which means all components will be used.
- Your window should look like this:



- Click Fit to create to model.
  In the status field of the QuaSAR-Model window you will find **RMSE** and $\mathbf{r^2}$

Secondly, we will cross-validate the created QSAR model:

- In the QuaSAR-Model window >> **Validate**
- Check all **Cross Validation** check boxes and uncheck the **Model Validation** check boxes:



- Now click **OK**.
- In the status field of the QuaSAR-Model window you will find the Standard Error of Prediction (**XRMSE**) and $\mathbf{q^2}$ (XR2).

Determine the optimal number of components

The two important factors here are the **standard error of prediction** (XRMSE in Moe) and the **cross-validated r^2** (usually called q^2, XRE2 in Moe). XRMSE 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 XRMSE and note the corresponding q^2.
- To accomplish this, in the *QuaSAR-Model window*, specify the number of components under **Options:***Component Limit:*.
  Each time followed by a combination of **Fit** and **Validate**.
- 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.
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:

- While holding down the *Control Key* select the non-steroidal compounds by clicking on number field to the left of the **mol** field.
- When done selecting, click with the right (third) mouse button in the last selected field >> **Delete Selected Entries**

Now repeat the steps of QSAR model creation and cross-validation for the steroidal dataset and determine the optimal number of components.

Determine the optimal descriptors and components

By varying the chosen descriptors and the number of components it is possible to get $\mathbf{q^2}$ as high as 0.607.
It is already possible with only four descriptors.
Use the correlation plots to determine the best combination of four descriptors, in terms of $\mathbf{q^2}$.

You can find the answer Here.