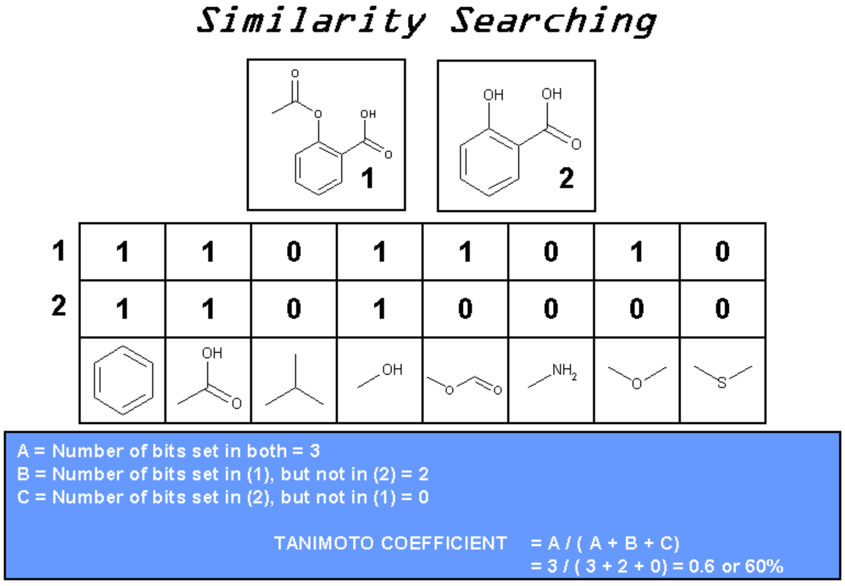
2-D searching Tutorial

By G. Schaftenaar

1. Introduction to 2D-similarity searches

2D-searching is applied to find compounds in a database which are similar in molecular features to a known active molecule(s). Each compound is assigned a 2D-fingerprint. A fingerprint is a set of bits, where each bit indicates the absence or presence of a molecular feature. To determine how similar two compounds are based on their fingerprints, the Tanimoto coefficient is often used. Below you will find an example of two compounds and their fingerprints and the calculation of the Tanimoto coefficient:



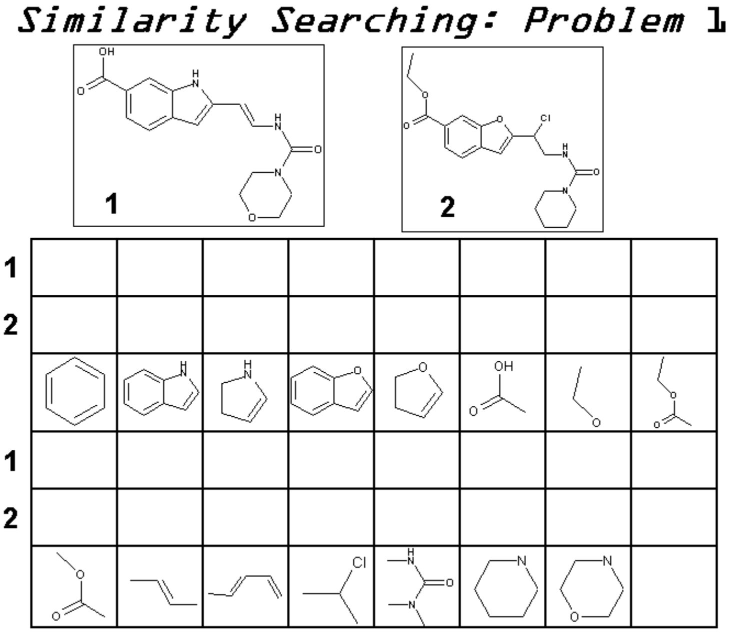
Similarity Searching

1	1	1	0	1	1	0	1	0
2	1	1	0	1	0	0	0	0

A = Number of bits set in both = 3
B = Number of bits set in (1), but not in (2) = 2
C = Number of bits set in (2), but not in (1) = 0

TANIMOTO COEFFICIENT = A / (A + B + C)
= 3 / (3 + 2 + 0) = 0.6 or 60%

Below you will find two new compounds. Calculate the Tanimoto coefficient for this pair of molecules.



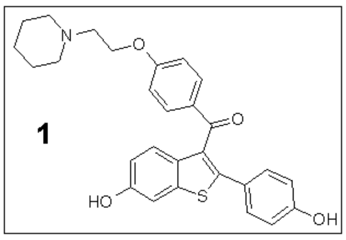
Similarity Searching: Problem 1

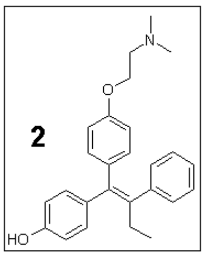
1								
2								

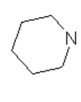
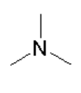
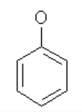
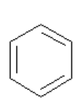
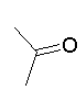
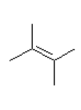
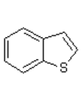
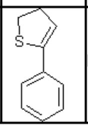
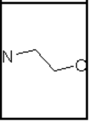
1								
2								

Below you will find two known active compounds for the estrogen receptor; Raloxifene(**1**) and Tamoxifen(**2**). Calculate the Tanimoto coefficient for this pair of molecules. Are these two compounds very similar ?

Similarity Searching: Problem 2


1


2

1									
2									
									

Setup the working environment

From the Unix shell (command prompt):

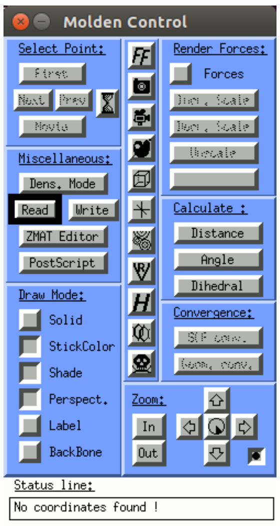
- Change directory by typing
  `cd Practicals/05_2DSearching`
- And call Molden by typing
  `gmolden`

---

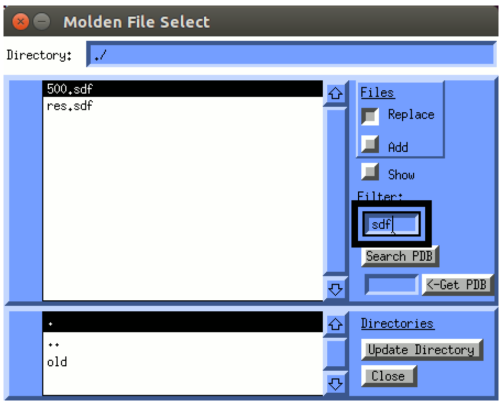
2. Find compounds in a database similar to a known active

How many compounds in the database of 500 compounds (485 randomly selected, 15 SERM (**S**elective **E**strogen **R**eceptor **M**odulator) are similar to tamoxifen, a known active for the ERa receptor.

Read in the database of 500 compounds:



No coordinates found !



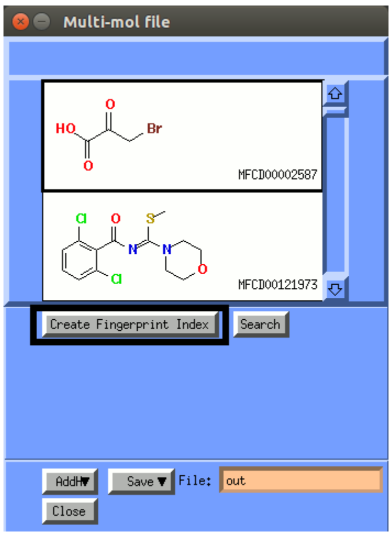
Directory: ./

500.sdf
res.sdf

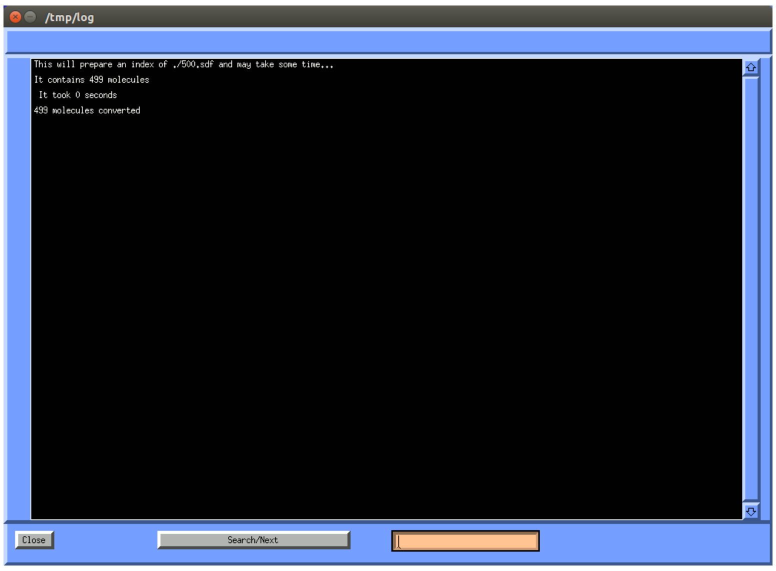
Filter: sdf

.
..
old

Next we have to calculate fingerprints for the molecules in the database:

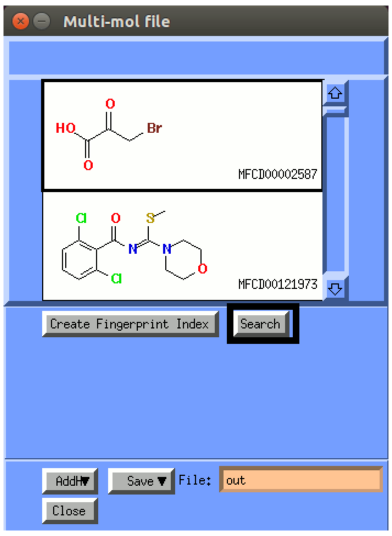


File: out

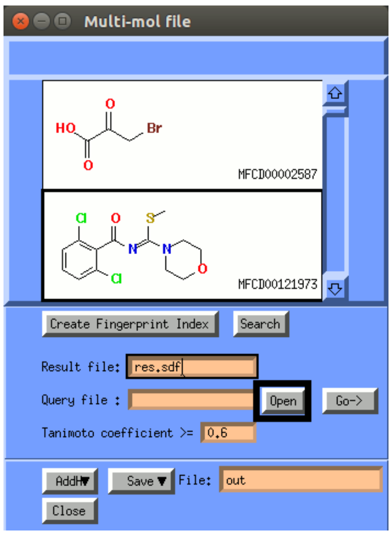


```
This will prepare an index of ./500.sdf and may take some time...
It contains 499 molecules
 It took 0 seconds
499 molecules converted
```

Now close this window
and let's do the 2D similarity search:



File: out



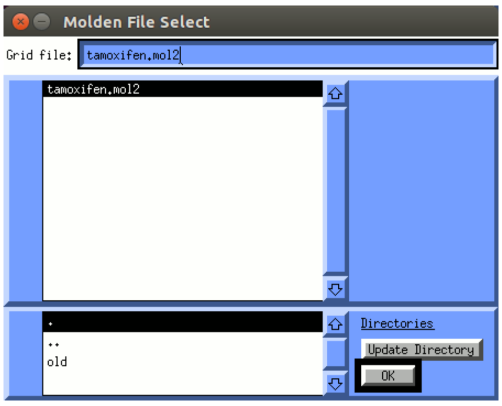
Result file: res.sdf

Query file :

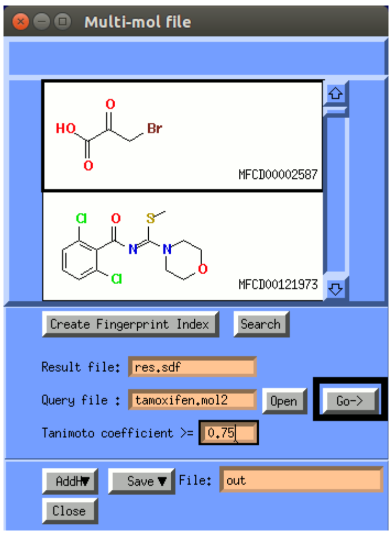
Tanimoto coefficient >= 0.6

File: out

Open the Query file: **tamoxifen.mol2**
and set the overlap (Similarity) to **0.75**:



Grid file: tamoxifen.mol2

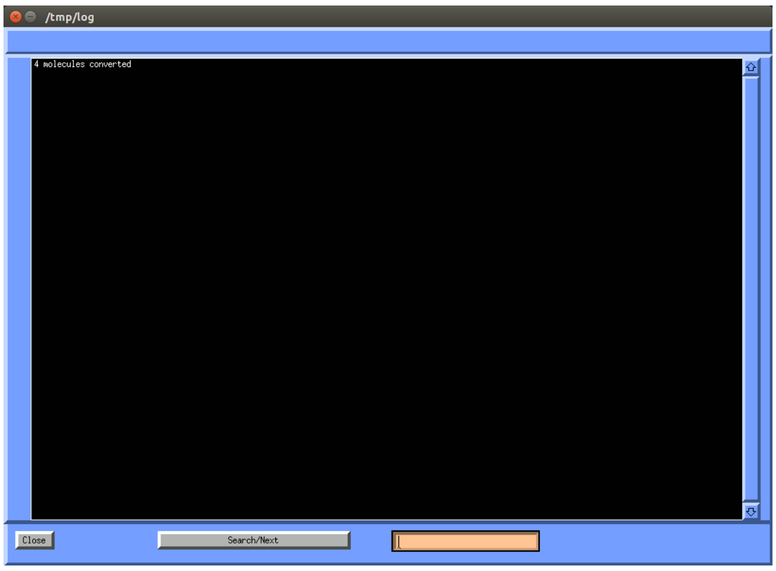


Result file: res.sdf

Query file : tamoxifen.mol2

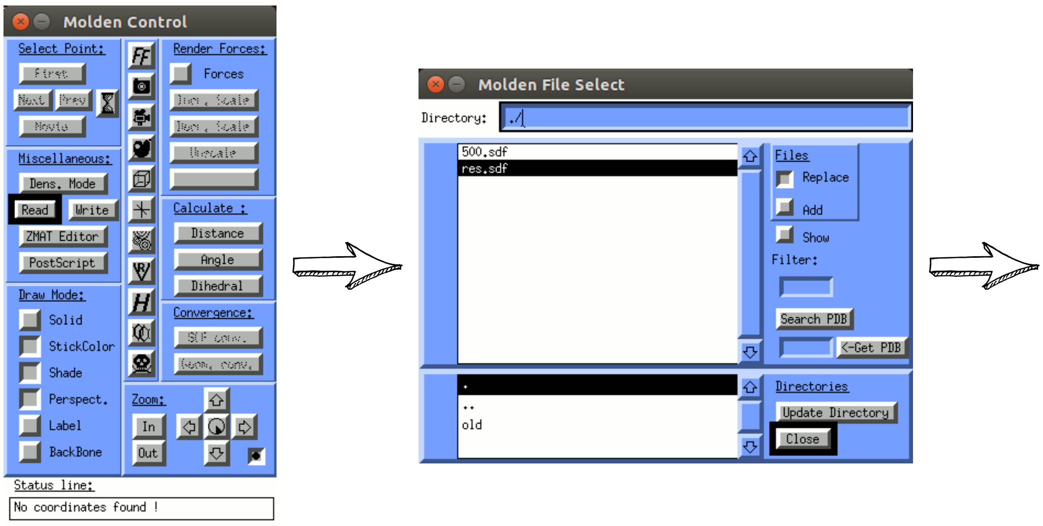
Tanimoto coefficient >= 0.75

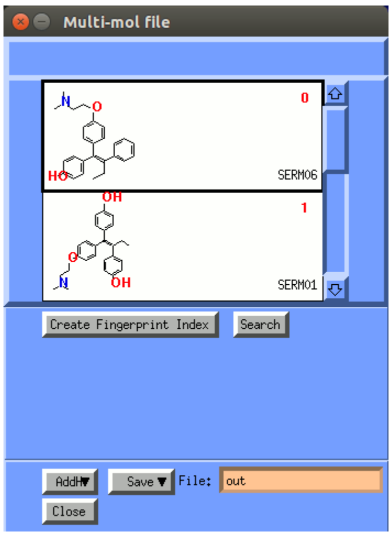
File: out



```
4 molecules converted
```

Now close this search results window.
The results of the search have been written in the **res.sdf** file. Now let us open this file and look at the results.





To navigate through the database use the Up and Down arraow keys.
To view a hit left click the compound icon field and click **OK**. You will 4 serms as hits.

---

2. Optimizing the similarity cutoff

Now repeat the same procedure but change the *Overlap* to **55**. You will now find 7 compounds matched using these similarity criteria.
Have a look at these compounds too (see above). How many of these hits are known drugs and how many of these are likely to be false positives ?
(answer is 7 and 0 respectively.)

If you repeat the same procedure with *Overlap* set to **48**. You will find almost all of the known drugs in the database (14) and twelve false positives.
In this case the *optimal* similarity cutoff is apparently 48.