Method Sheet 110

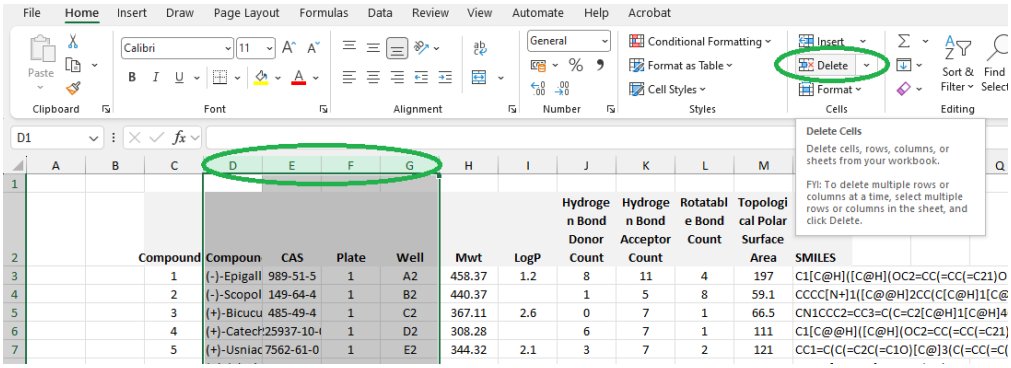
Exploring the structural features of *Puretitre* compounds

Overview

This method sheet explains how to perform correlation analyses to seek evidence of associations between the structural properties of *Puretitre* compounds and their biological activities, from high-throughput screening data. This analysis is a necessary step towards understanding the underlying "structure-activity relationship" (SAR) of lead molecules arising from a drug discovery screen. This SAR insight gives medicinal chemists a useful indication of what modifications they may apply to the molecule to further improve its activity and other properties, during the next steps of drug development.

Aligning the data for correlation analysis

1) Open your analysis Excel file containing the *Puretitre* screening data of interest.
2) In these examples, we will explore the *Puretitre E. coli* growth screening data.
3) Insert a new worksheet and name it 'Correlations'.
4) Go to the 'Compound Info' tab of the original file and copy all of the columns from the original *Puretitre* structural information table, being sure to include the data from all 200 compounds.
5) Open the 'Correlations tab' and paste the data into cell `C2` (i.e. starting from the third column in the worksheet) - this is done to leave space in the first two columns for later data entry.
6) We do not require the data from the 'Compound name', 'CAS', 'Plate', 'Well' or 'SMILES' columns, so delete these columns completely from the 'Correlations' worksheet by highlighting them, then clicking 'Delete cells' in the ribbon.



7) You should now have a table, starting in column `C`, with 7 columns, namely: 'Compound #', 'Mwt', 'LogP', 'Hydrogen Bond Donor Count', 'Hydrogen Bond Acceptor Count', 'Rotatable Bond Count' and 'Topological Polar Surface Area'.
8) Now go to another worksheet in which you have already calculated a data endpoint, such as bacterial growth at 24 hours, or mammalian cell viability, for all 200 *Puretitre* compounds.
9) Arrange this data so that the first column is the 'Compound ID' (numbered from 1 to 200) and the second column is the bioassay data for that compound.

10) Highlight both columns, the use the ‘Sort’ function to order column 1 with values from smallest to largest, so that the compound ID 1 is in the first row, and compound 200 is in the last row.

11) Make sure you highlight both columns and sort them together, as it is critical that the compound ID matches its corresponding measurement.

12) Now copy those two columns and paste the data as values into the two empty columns (`A` and `B`) of the ‘Correlations’ worksheet.

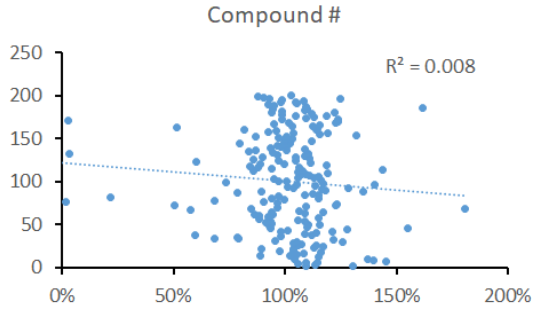
13) Make sure that the compound ID numbers match up exactly in both column 1 and column 3 of the new table - it should now look like this:

	A	B	C	D	E	F	G	H	I
1									
2	ID	Growth	ompound	Mwt	LogP	Hydrogen Bond Donor Count	Hydrogen Bond Acceptor Count	Rotatable Bond Count	Topological Polar Surface Area
3	1	109%	1	458.37	1.2	8	11	4	197
4	2	130%	2	440.37		1	5	8	59.1
5	3	114%	3	367.11	2.6	0	7	1	66.5
6	4	106%	4	308.28		6	7	1	111
7	5	109%	5	344.32	2.1	3	7	2	121

Performing the correlation analysis

1) Highlight all of the data, for all 200 values, in both column B (the measurements from the bioassay) and column C (Compound #), including the top row which includes the heading text for each column.

2) Use the ‘Insert Chart’ function to create a scatter plot of the data, then insert a linear trendline on the chart as explained in Method Sheet 109, being sure to select the option to display the R^2 value on the chart - it should look something like this:



3) Take a note of this R^2 value somewhere else in the worksheet, in this example we would have two cells in the row, one with the title ‘Compound #’ and the other with the corresponding R^2 value from the chart.

4) Click away from the chart, then click on it again.

5) You will notice that the cell containing the heading of the column that provides values to plot in the y-axis direction turns red (as shown below).

ID	Growth	ompound	Mwt	LogP	Hydrogen Bond Donor Count	Hydrogen Bond Acceptor Count	Rotatable Bond Count	Topological Polar Surface Area
1	109%	1	458.37	1.2	8	11	4	197
2	130%	2	440.37		1	5	8	59.1
3	114%	3	367.11	2.6	0	7	1	66.5
4	106%	4	308.28		6	7	1	111
5	109%	5	344.32	2.1	3	7	2	121

6) Click on the outer line of this chart, for example where shown by the green circle, then drag the red square one cell to the right.

7) Now it should look like this:

ID	Growth	ompound	Mwt	LogP	Hydrogen Bond Donor Count	Hydrogen Bond Acceptor Count	Rotatable Bond Count	Topological Polar Surface Area
1	109%	1	458.37	1.2	8	11	4	197
2	130%	2	440.37		1	5	8	59.1
3	114%	3	367.11	2.6	0	7	1	66.5

8) The chart will automatically update to plot the new data in the next column.

9) Take a note of the column heading (in this example, Mwt) and the resulting R^2 value as shown on the chart in your table of compound properties and R^2 values.

10) Continue doing this, moving the right hand data selection area for the chart rightwards one more column and noting the R^2 value each time, until every column has been examined.

11) Now insert the formulae shown in Method Sheet 109 to generate a table of all the p-values from the R^2 values for every correlation.

12) The table should look something like this:

Structural property	R2	r	n	t	P-value
Compound ID	0.008	0.089	200	1.26	0.208
Mwt	0.0022	0.047	201	0.66	0.508
LogP	0.0017	0.041	202	0.58	0.560
Hydrogen Bond Donor Count	0.0351	0.187	203	2.70	0.007
Hydrogen Bond Acceptor Count	0.0455	0.213	204	3.10	0.002
Rotatable Bond Count	0.0002	0.014	205	0.20	0.840
Topological Polar Surface Area	0.0382	0.195	206	2.85	0.005

13) This is the table that you will report in your dissertation.

14) Remember, we are looking for any p-values lower than 0.05, which may suggest a significant correlation between the two variables.

15) In this example, the results show that there may be significant correlations between the capacity of compounds to prevent growth of *E. coli*, and the compounds 'Hydrogen Bond Donor Count', 'Hydrogen Bond Acceptor Count' and 'Topological Polar Surface Area'.

16) All of these observations tell us useful things about what types of structure, generally speaking, tend to have more biological activity in our assay.

Correcting for multiple testing

1) In this example, we have conducted seven separate statistical tests on the same overall dataset.
2) This can cause problems with generation of false positive results, so we should apply a correction for multiple testing to be sure the results obtained really are statistically significant.
3) Follow the advice given in Method Sheet 111 for how to do this.

Notes

- 'Mwt' is the molecular weight of the molecule, essentially how large it is - good drugs are often less than 500 Da in mass.
- 'LogP' is a measure of how hydrophobic the molecule is, higher numbers mean more fat-soluble, lower numbers mean more water soluble - good drugs tend to have LogP in the range of 0 to 5.
- 'Hydrogen Bond Donor Count' is a measure of how many -OH and -NH groups are available to interact with water - good drugs tend to have no more than 5 of these.
- 'Hydrogen Bond Acceptor Count' is a measure of how many N or O atoms with lone pairs of electrons are available to hydrogen bond with water - molecules with greater than 10 of these tend to be poorly absorbed from the gut.
- 'Rotatable Bond Count' is a measure of how flexible a molecule is - good drugs tend to have a count of less than 10.
- 'Topological Polar Surface Area' is a measure of how "polarised" the surface of the molecule is - values less than 140 tend to be associated with good gut absorption and cell membrane permeability.
- The first four of these terms are also the cornerstones of **Lipinski's Rule of Five**, which is a general set of rules to ascertain how likely a molecule is to function well as a drug, essentially predicting whether a compound can actually reach its target inside a cell before being broken down or excreted..
- Remember that the data must be correctly aligned with the compound ID at every stage of aligning the two datasets together - any error here will mean the tests will not work at all.
- If the data table were complete with no missing values, it would be quicker and easier to use the `CORREL` function to calculate 'r' values for all of the correlations, but as some of the columns are missing values, it would not work for those columns so the scatter plot and trendline option is the better approach for this data set.