Data Pretreatment Program:

Before performing the actual chemometric analysis in QSAR, the raw data is usually pretreated due to following reasons:

1. To remove those descriptor whose values remains more or less constant for all compounds, since they only add computing time without contributing much to the results.
2. To reduce collinearity among the descriptors thereby preventing data over-fitting and aids in improving the prediction performance of the model.

The purpose of this program is to remove these constant and highly correlated descriptors based on user specified cut-off values:
Variance cut-off value: *Based on this all the descriptor columns with a variance below this cut-off value are removed from the data.*
Correlation cut-off value: *Based on this all the descriptors with higher correlation (higher then cut-off value) among each other are removed, except one with highest correlation with activity among them.*

Input file: Single *.csv file (saved as 'comma delimited' in Microsoft Excel).
Input file format: Columns containing all calculated descriptor and their values for each compound present in dataset set. Header*(e.g. name of respective descriptor, activity title) for each column is must. Observed/actual activity values are required and must be present in last* column. To further clarify file format, sample input files are provided. (* *mandatory for correct calculations*)

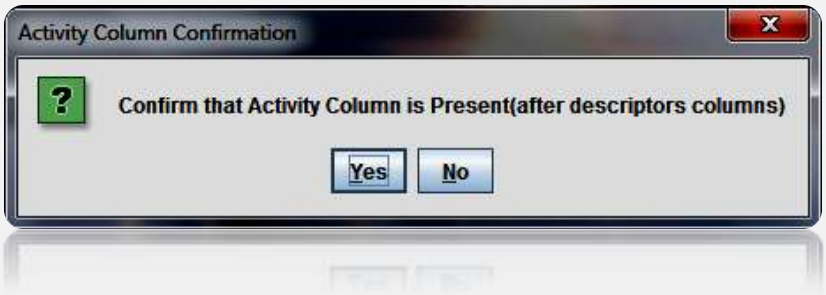
To run the program: Download *dataPreTreat.**jar*** file (it is platform-independent) and just single/double click on it to run the program.

Note: Keep your input file in the same folder where you keep *dataPreTreat.**jar*** file.

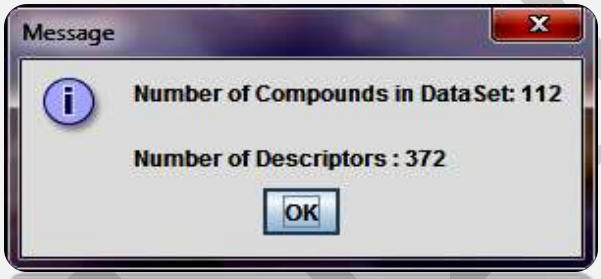
Steps involved:

1. Enter the name of data set input file (name without extension as depicted).

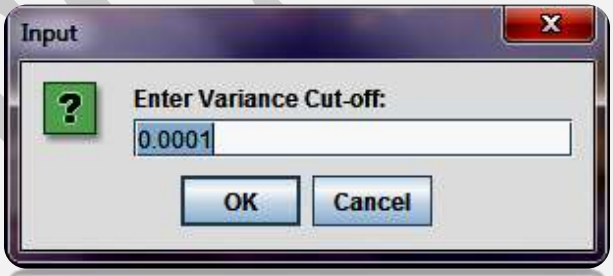
2. This step is to confirm whether observed/actual activity column is present or not (must be last* column in input file). If pressed 'No' program will exit.



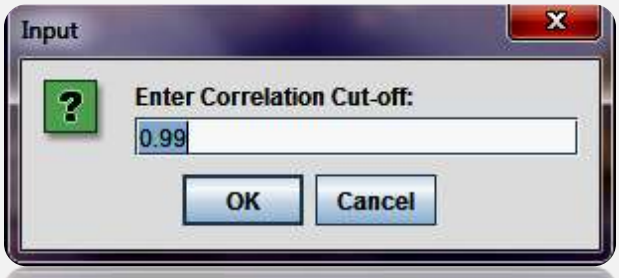
3. This is an important step. Please check both values carefully and see if both are correct. If not, check your input file. If you find your input file correct and error still exist, then create input file freshly once again (by copying all observations and paste it in new excel sheet and save it as csv file). This should eliminate the error.



4. Enter variance cut-off (Default value: 0.0001).



5. Enter correlation cut-off (Default value: 0.99).



Output files:

***[Name of dataset input file]_Result.csv* –** This output file will consists of resulting pool of descriptors after data pre-treatment i.e. after removing constant and highly correlated descriptors based on user cut-off values.

*Snapshot:*_Result.csv in Microsoft Excel*

	A	B	C	D	E	F	KU	KV	KW	KX	KY
1	AlogP	AlogP^2	AlogP98	Alogp98^2	LogP	LogP	LS_05	DLS_cons	LLS_01	LLS_02	Activity
2	3.305	10.92303	4.3503	18.92511	4.12	16	0	0.77	0.83	1	3.12
3	3.1516	9.932583	3.9039	15.24044	3.67	13	0	0.77	0.83	1	3.23
4	3.9513	15.61277	5.0815	25.82164	2.67	7	0	0.77	1	1	3.17
5	3.7287	13.90321	5.4502	29.70468	3.345	11.1	0	0.77	0.67	1	3.2
6	1.7935	3.216644	3.3139	10.98193	3.285	10.7	0	0.77	1	1	3.25
7	4.502901	20.27611	5.382602	28.9724	5.21	27	0	0.74	0.67	0.88	3.97
8	6.185401	38.25919	7.502501	56.28751	5.34	28	0	0.74	0.67	0.88	3.93
9	6.185401	38.25919	7.502501	56.28751	5.34	28	0	0.74	0.67	0.88	4.58
10	4.642401	21.55188	5.5881	31.22687	5.41	29	0	0.71	0.67	0.88	4.08
11	4.250201	18.06421	5.366201	28.79612	5.28	27	0	0.77	0.67	1	4.45
12	7.068201	49.95947	8.444801	71.31467	6.41	41	0.5	0.7	0.67	0.75	3.18
13	7.068201	49.95947	8.444801	71.31467	6.41	41	0.5	0.7	0.67	0.75	3.67
14	6.703401	44.93559	8.166902	66.69828	6		0	0.71	0.67	0.88	4.44
15	6.842901	46.82529	8.3724	70.09709	6.140001	37.6	0	0.71	0.67	0.88	4.15
16							0	0.71	0.67	0.88	4.4

***[Name of dataset input file]*_LogFile.txt –** It consist of list of names of all descriptors (*respective header* in input file*) removed based on user specified variance and correlation cut-off. It also enlists all inter-correlated descriptors; the one having highest correlation with activity among them is kept and all others are removed.

*Snapshot: *_LogFile.txt in Wordpad*

```
Descriptors removed based on Variance cut-off :
Jurs-FPSA-3
ETA_Epsilon_3
ETA_dEpsilon_D
ETA_EtaP_B
ETA_EtaP_B_RC
PW4


Descriptors removed based on Correlation cut-off :
Atype_C_39
Atype_C_41
nAB
S_aaN
S_dO
nRCOOR
O-058


Inter-correlated Descriptor Log data :

1. S_aaaC : Atype_C_39 : Atype_C_41 : nAB
2. S_aaN : Atype_N_75
3. S_dO : Atype_O_58 : nRCOOR : O-058
4. S_ssO : SssO
```

Note: Please close all input/output files before running this program.

Disclaimer

For academic purpose only; Not for commercial use.

The program DTC_DataPreTreatment has been developed (in Java) and validated on known data sets by Pravin Ambure (ambure.pharmait@gmail.com) of Drug Theoretics & Cheminformatics (DTC) Laboratory, Jadavpur University (2013).