vWSP

A Data Pre-Treatment Tool

NANOBRIDGES
-A Collaborative Project

UNIWERSYTET GDAŃSKI

Marie Curie Actions
EUROPEAN UNION

THE AUTHORS ARE GRATEFUL FOR THE FINANCIAL SUPPORT FROM THE EUROPEAN COMMISSION THROUGH THE MARIE CURIE **IRSES** PROGRAM, NANOBRIDGES PROJECT (FP7-PEOPLE-2011-IRSES, GRANT AGREEMENT NUMBER 295128).

V-WSP Tool

To remove the constant and highly inter-correlated descriptors based on user specified variance and correlation coefficient cut-off values using ***V-WSP algorithm*** proposed *by Ballabio et. al.* [1]. It is an unsupervised variable reduction method, which is a modification of the recently proposed WSP algorithm for design of experiments (DOE).

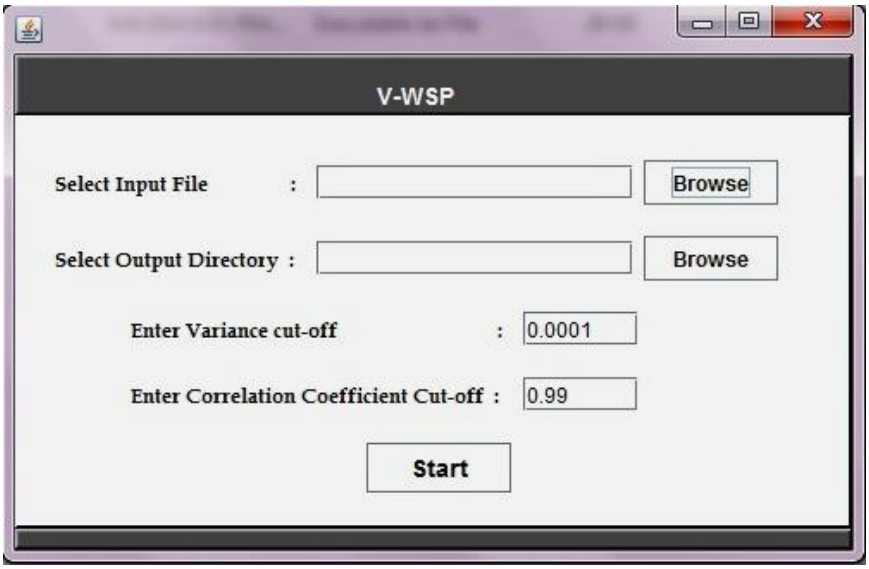
Algorithm

Given a data matrix with *n* rows (*samples*) and *p* columns (*variables*), the algorithm for calculating the V-WSP method is given below:

Steps:

1. Choose an initial variable (seed) *j* and a correlation threshold (*thr*);
2. Calculate the Pearson linear correlation coefficients (*c*) between *j* and all other variables;
3. Eliminate variables *d* such as absolute value of $c_{dj} \geq thr$;
4. Variable *j* is selected and replaced by the variable with the highest absolute correlation value with *j* among the remaining variables;
5. Repeat steps 2, 3 and 4 until there are no more variables to select.

Snapshot 1: V-WSP Tool



V-WSP Program Folder

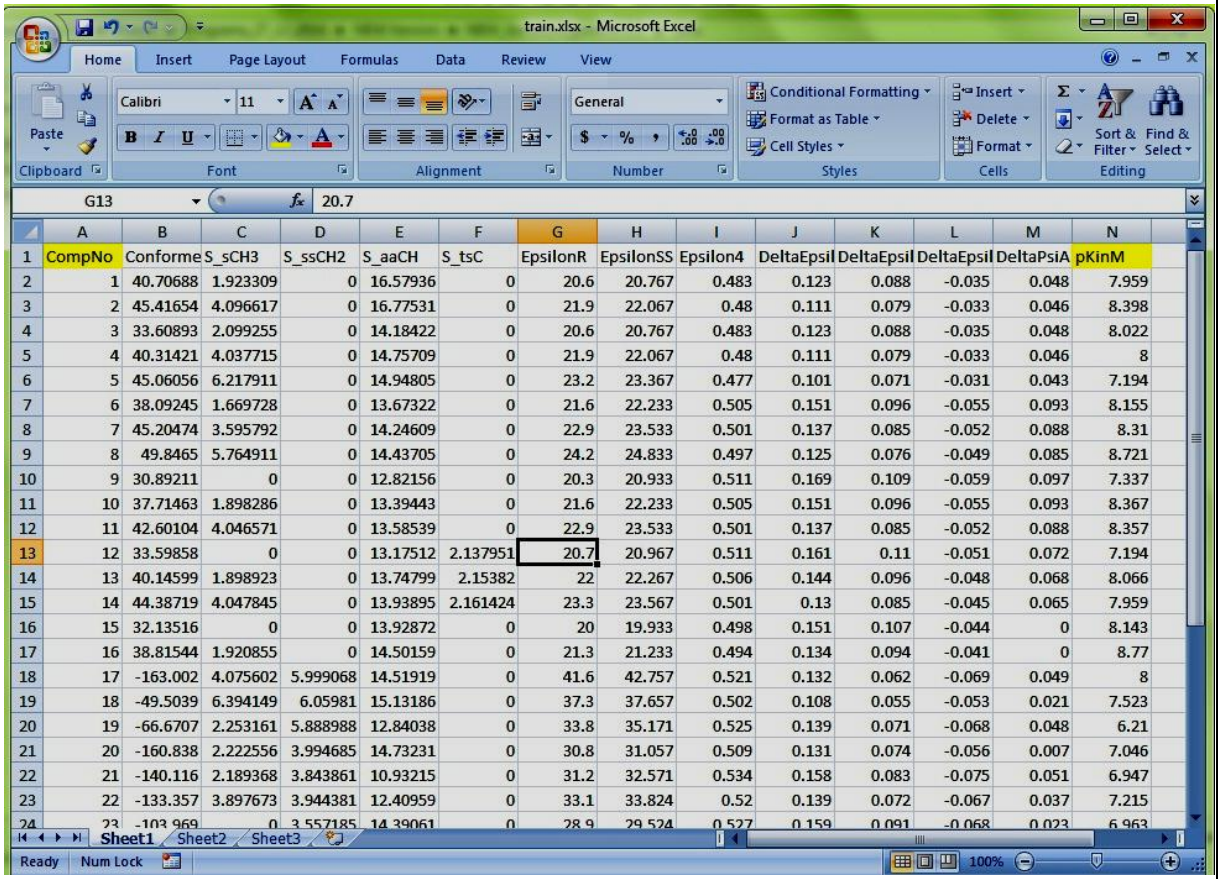
The program folder will consist of three folders "**Data**", "**Lib**" and "**Output**". For convenience, user may keep input file in "**Data**" folder and may save output files in "**Output**" folder, since by default, clicking on the browse button will open these folders.

"**Lib**" folder consists of library files required for running the program. Hence do not move or delete or rename these files.

Input file format

Three different file types are allowed *i.e.* ***xlsx , xls*** and ***csv*** as input file. The input file (*see snapshot 2*) should consist of compound number (*first column*), descriptor values and the endpoint values (*last column*) for each object/compound. The format in which this information should be placed in the file is as follows:

Snapshot 2



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	CompNo	Conforme	S_sCH3	S_ssCH2	S_aaCH	S_tsC	EpsilonR	EpsilonSS	Epsilon4	DeltaEpsil	DeltaEpsil	DeltaEpsil	DeltaPsiA	pKinM
2	1	40.70688	1.923309	0	16.57936	0	20.6	20.767	0.483	0.123	0.088	-0.035	0.048	7.959
3	2	45.41654	4.096617	0	16.77531	0	21.9	22.067	0.48	0.111	0.079	-0.033	0.046	8.398
4	3	33.60893	2.099255	0	14.18422	0	20.6	20.767	0.483	0.123	0.088	-0.035	0.048	8.022
5	4	40.31421	4.037715	0	14.75709	0	21.9	22.067	0.48	0.111	0.079	-0.033	0.046	8
6	5	45.06056	6.217911	0	14.94805	0	23.2	23.367	0.477	0.101	0.071	-0.031	0.043	7.194
7	6	38.09245	1.669728	0	13.67322	0	21.6	22.233	0.505	0.151	0.096	-0.055	0.093	8.155
8	7	45.20474	3.595792	0	14.24609	0	22.9	23.533	0.501	0.137	0.085	-0.052	0.088	8.31
9	8	49.8465	5.764911	0	14.43705	0	24.2	24.833	0.497	0.125	0.076	-0.049	0.085	8.721
10	9	30.89211	0	0	12.82156	0	20.3	20.933	0.511	0.169	0.109	-0.059	0.097	7.337
11	10	37.71463	1.898286	0	13.39443	0	21.6	22.233	0.505	0.151	0.096	-0.055	0.093	8.367
12	11	42.60104	4.046571	0	13.58539	0	22.9	23.533	0.501	0.137	0.085	-0.052	0.088	8.357
13	12	33.59858	0	0	13.17512	2.137951	20.7	20.967	0.511	0.161	0.11	-0.051	0.072	7.194
14	13	40.14599	1.898923	0	13.74799	2.15382	22	22.267	0.506	0.144	0.096	-0.048	0.068	8.066
15	14	44.38719	4.047845	0	13.93895	2.161424	23.3	23.567	0.501	0.13	0.085	-0.045	0.065	7.959
16	15	32.13516	0	0	13.92872	0	20	19.933	0.498	0.151	0.107	-0.044	0	8.143
17	16	38.81544	1.920855	0	14.50159	0	21.3	21.233	0.494	0.134	0.094	-0.041	0	8.77
18	17	-163.002	4.075602	5.999068	14.51919	0	41.6	42.757	0.521	0.132	0.062	-0.069	0.049	8
19	18	-49.5039	6.394149	6.05981	15.13186	0	37.3	37.657	0.502	0.108	0.055	-0.053	0.021	7.523
20	19	-66.6707	2.253161	5.888988	12.84038	0	33.8	35.171	0.525	0.139	0.071	-0.068	0.048	6.21
21	20	-160.838	2.222556	3.994685	14.73231	0	30.8	31.057	0.509	0.131	0.074	-0.056	0.007	7.046
22	21	-140.116	2.189368	3.843861	10.93215	0	31.2	32.571	0.534	0.158	0.083	-0.075	0.051	6.947
23	22	-133.357	3.897673	3.944381	12.40959	0	33.1	33.824	0.52	0.139	0.072	-0.067	0.037	7.215
24	23	-103.969	0	3.557185	14.39061	0	28.9	29.524	0.527	0.159	0.091	-0.068	0.023	6.963

First Row: Header *i.e.* name for each column, for instances, descriptor names, and endpoint name. It can be numerical, alphabet or alphanumerical in nature.

First column: Serial number/Compound number (only numerical values)

Subsequent columns: Property/Independent variables/Descriptor values; each column will consist of each descriptor values for all the nanoparticles. These values should be numerical values and not alphabets or alphanumerical values.

Last column: Endpoint values/Dependent variables

How to run the program

It is simple! Just click/double click on the jar file (vWSP.*jar*) present in the vWSP program folder. A window will open as shown in *Snapshot 1*, with few queries, which a user has to fill before clicking on '**Start**' button to run the program.

"***Select Query Input File***": Click on 'browse' button to select the input file. By default, it will open the "Data" folder present in vWSP program folder. So for convenience, user can keep the input file in the "Data" folder.

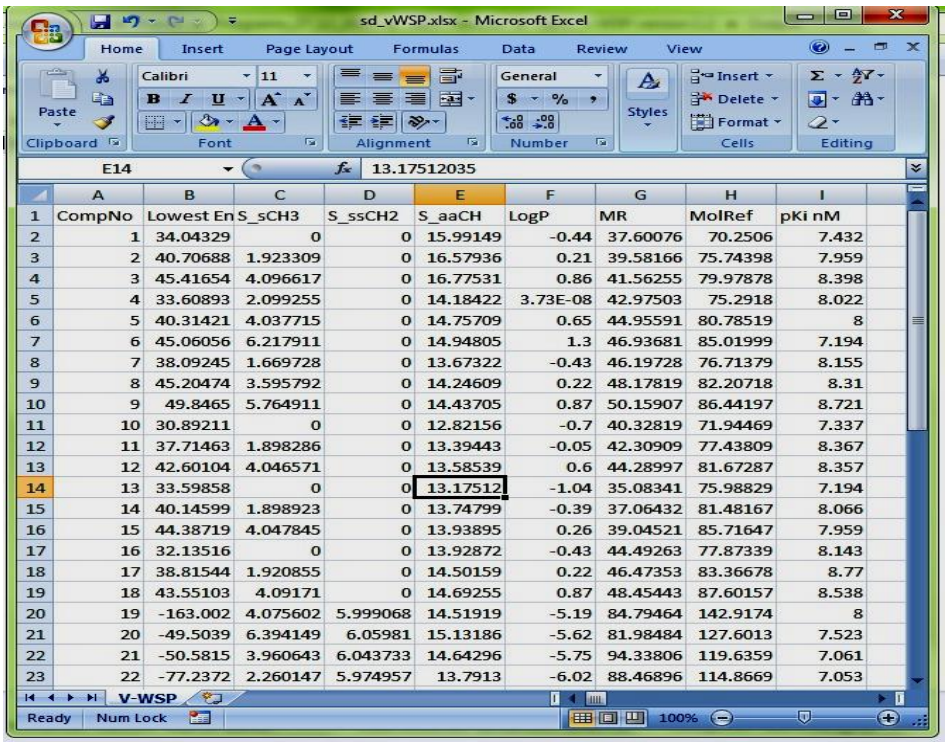
"***Select Output Directory***": Click on 'browse' button to select the destination/output file directory and define output file name. By default, it will open the "Output" folder present in vWSP program folder. So for convenience, user can save the output files in the "Output" folder.

Enter Variance cut-off: Enter the variance cut-off value based on which the constant variables will be removed. By default, the cut-off value is set to *0.0001*.

Enter correlation coefficient cut-off: Enter the inter-correlation coefficient cut-off value based on which the inter-correlated variables will be removed. By default, the cut-off value is set to *0.99*.

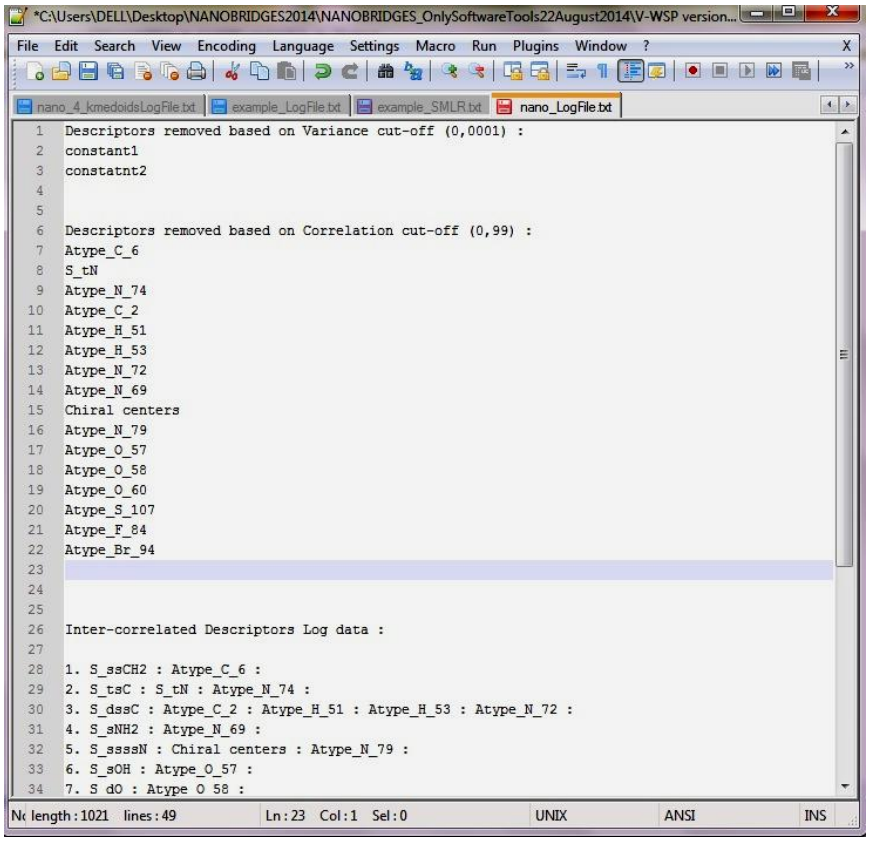
Output

Snapshot 3



	A	B	C	D	E	F	G	H	I
1	CompNo	Lowest En	S_sCH3	S_ssCH2	S_aaCH	LogP	MR	MolRef	pKi nM
2	1	34.04329	0	0	15.99149	-0.44	37.60076	70.2506	7.432
3	2	40.70688	1.923309	0	16.57936	0.21	39.58166	75.74398	7.959
4	3	45.41654	4.096617	0	16.77531	0.86	41.56255	79.97878	8.398
5	4	33.60893	2.099255	0	14.18422	3.73E-08	42.97503	75.2918	8.022
6	5	40.31421	4.037715	0	14.75709	0.65	44.95591	80.78519	8
7	6	45.06056	6.217911	0	14.94805	1.3	46.93681	85.01999	7.194
8	7	38.09245	1.669728	0	13.67322	-0.43	46.19728	76.71379	8.155
9	8	45.20474	3.595792	0	14.24609	0.22	48.17819	82.20718	8.31
10	9	49.8465	5.764911	0	14.43705	0.87	50.15907	86.44197	8.721
11	10	30.89211	0	0	12.82156	-0.7	40.32819	71.94469	7.337
12	11	37.71463	1.898286	0	13.39443	-0.05	42.30909	77.43809	8.367
13	12	42.60104	4.046571	0	13.58539	0.6	44.28997	81.67287	8.357
14	13	33.59858	0	0	13.17512	-1.04	35.08341	75.98829	7.194
15	14	40.14599	1.898923	0	13.74799	-0.39	37.06432	81.48167	8.066
16	15	44.38719	4.047845	0	13.93895	0.26	39.04521	85.71647	7.959
17	16	32.13516	0	0	13.92872	-0.43	44.49263	77.87339	8.143
18	17	38.81544	1.920855	0	14.50159	0.22	46.47353	83.36678	8.77
19	18	43.55103	4.09171	0	14.69255	0.87	48.45443	87.60157	8.538
20	19	-163.002	4.075602	5.999068	14.51919	-5.19	84.79464	142.9174	8
21	20	-49.5039	6.394149	6.05981	15.13186	-5.62	81.98484	127.6013	7.523
22	21	-50.5815	3.960643	6.043733	14.64296	-5.75	94.33806	119.6359	7.061
23	22	-77.2372	2.260147	5.974957	13.7913	-6.02	88.46896	114.8669	7.053

Snapshot 4



1. ***Output _.xlsx file (snapshot 3):*** The generated excel sheet (***.xlsx/xls/csv***) will consist of the compound no./serial no. from input file (*First column*), remaining descriptor columns after removing the constant and inter-correlated descriptors based on user cut-off values (*subsequent columns*). The *last column* will be the endpoint column.
2. ***Log file .txt (snapshot 4):*** This file consists of list of all descriptor names removed based on user specified variance and correlation coefficient cut-off. It also enlists all inter-correlated descriptors information; the one having highest correlation with activity among them is kept and all others are removed.

Reference

1. Ballabio, Davide, *et al.* "A novel variable reduction method adapted from space-filling designs." *Chemometrics and Intelligent Laboratory Systems* (2014), 136, 147-154.

Java External Library Used

Apache POI – the Java API for Microsoft Documents

- Available at http://poi.apache.org/

XMLBeans

- Available at http://xmlbeans.apache.org/

Disclaimer

For academic purpose only.

The program AD-MDI has been developed in Java language and is platform independent. The software is validated on known data sets. Please report for discrepancy of result for any other dataset. Contact us at any of the following addresses:

Dr. Tomasz Puzyn,
NanaBRIDGES Project Coordinator,
Faculty of Chemistry,
University of Gdansk,
Gdansk,
Poland 80-952
Email Id: puzi@qsar.eu.org

Dr. Kunal Roy,
Drug Theoretics and Cheminformatics Lab.,
Dept. of Pharmaceutical Technology,
Jadavpur University,
Kolkata, West Bengal,
INDIA-700032
Email Id: kunalroy_in@yahoo.com

Software Developer details:

Pravin Ambure,
Research Scholar,
Drug Theoretics and Cheminformatics Lab.,
Dept. of Pharmaceutical Technology,
Jadavpur University,
Kolkata, West Bengal,
INDIA-700032
E-mail Id: ambure.pharmait@gmail.com (*for any queries regarding the tool)