Euclidean Program:

This program is to ensure that the structures of the test set are representative of the entire dataset and training set (i.e. whether the test set structures are within the applicability domain or not). It is based on distance scores calculated by the Euclidean distance norm.

Euclidean based applicability domain (AD):

Applicability domain (AD) is the physicochemical, structural or biological space, knowledge or information on which the training set of the model has been developed. The resulting model can be reliably applicable for only those compounds which are inside this domain. This program helps to ensure that the compounds of the test/external set are representative of the training set compounds used in model development. It is based on distance scores calculated by the Euclidean distance norms. At first, normalized mean distance score for training set compounds are calculated and these values ranges from 0 to 1(0=least diverse, 1=most diverse training set compound). Then normalized mean distance score for test set are calculated, and those test compounds with score outside 0 to 1 range are said to be outside the applicability domain. This can also be checked by plotting a 'Scatter plot' (normalized mean distance vs. respective activity/property) including both training and test set. If the test set compounds are inside the domain/area covered by training set compounds that means these compounds are inside the applicability domain otherwise not.

Input File: Two *.csv file, each for training and test Set compounds
Input File Format: Column containing descriptors and their values for each compound present in training/test set in respective csv files. Header*(i.e. name of respective descriptors) for each column is must. The order of descriptors* in training set input file must be identical to that of test set input file. Format is same for both the csv file. To further clarify file format, sample input files are provided. (* *mandatory for correct calculations*)

To run the program: Download ***Euclidean.jar*** file (it is platform-independent) and just single/double click on it to run the program.
Note: Keep your input files in the same folder where you keep ***Euclidean.jar*** file.

Output File:
[Name of training set file]_ [Name of test set file]_Euclidean.csv – It consist of four columns each for training set and test set (observations for test set compounds are present below training set compounds observations): allotted compound number (are in the same order to the compounds present in input training set and test set file respectively), distance score, mean distance score, normalized mean distance score.

*Snapshot: *_Euclidean.csv in Microsoft Excel*

	A	B	C	D	E
1	Training Set:				
2	Alloted CompdNo.	Distance Score	Mean Distance	Normalized Mean Distance	
3	1	1615.7507	4.98688	0.11129	
4	2	1125.14856	3.47268	0.02141	
5	3	1249.37278	3.85609	0.04417	
6	4	1126.14179	3.47575	0.02159	
7	5	1313.76655	4.05484	0.05597	
8	6	1246.97686	3.84869	0.04373	
9	7	1051.94209	3.24673	0.009	
13	Test Set:				
14	Alloted CompdNo.	Distance Score	Mean Distance	Normalized Mean Distance	
15	1	1346.80228	4.1568	0.06202	
16	2	1272.52941	3.92756	0.04841	
17	3	1843.6167	5.69017	0.15304	
18	4	1666.96571	5.14496	0.12068	
19	5	1485.68726	4.58545	0.08746	
20	6	1202.71698	3.71209	0.03562	
21	7	1681.35429	5.18937	0.12331	

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

Disclaimer

For academic purpose only.

The program DTC_Euclidean 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).