MODelability Index 1.0
(MODI)

NANOBRIDGES
-ACollaborative 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).

Last updated: 22.05.2015

Modelability Index 1.0

Background: The predictivity of QSAR models depends upon the dataset characteristics such as chemical diversity, activity cliff, size etc., as well as the modeling protocol followed to build, select and validates the model. But sometimes, even though the use of different algorithm for the development and validation of models does not results in robust, statistically significant and better predictive models. Such attempt are time consuming, when working with the larger dataset (Compounds>1000). Moreover, it is also observed that it is often impossible to develop the good predictive model despite of using complex algorithm and rigorous modeling exercises. The concept of modelability index i.e. a prior estimation of feasibility of developing models would be much helpful to avoid the unnecessary efforts, resources and time involved in the development of QSAR model.

About the algorithm: MODI is defined as an activity class-weighted ratio of the number of nearest-neighbor pairs of compounds with thesame activity class versus the total number of pairs. The steps involved in the calculation of MODI are as follows:

1. Calculation of normalized Euclidean distance

$$d_{ij} = \left\| X_i - X_j \right\| = \sqrt{\sum_{k=1}^{m} (X_{ik} - X_{jk})^2} \qquad \text{(i)}$$

$$\overline{d_i} = \frac{\sum_{j=1}^{n} d_{ij}}{n-1} \qquad \text{(ii)}$$

$$MeanDistance_{(Normalized)} = \frac{(MeanDistance - Min\_MeanDistance)}{(Max\_MeanDistance - Min\_MeanDistance)} \qquad \text{(iii)}$$

$$MODI = \frac{1}{K} \sum_{i=1}^{k} \frac{N_i^{same}}{N_i^{total}} \qquad \text{(iv)}$$

Where,

d_{ij}: the distance score between two compounds

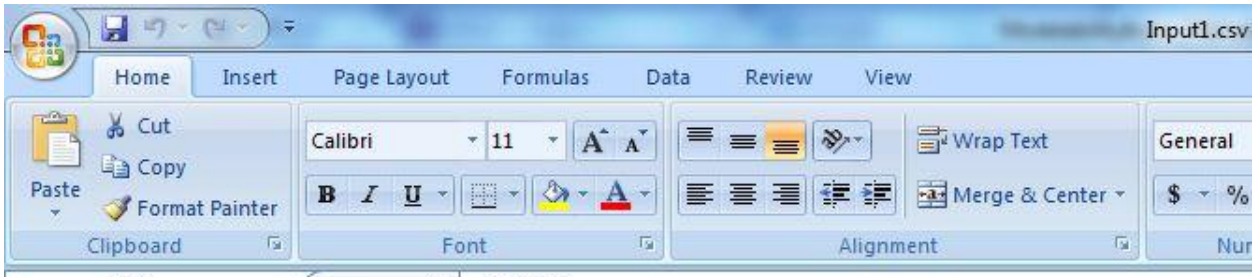
$\overline{d_i}$: Mean distance

K: no. of classes (k=2 for binary dataset)

N_i^{same}: No. of compounds of i-th activity class that have their first nearest neighbors belonging to same activity class

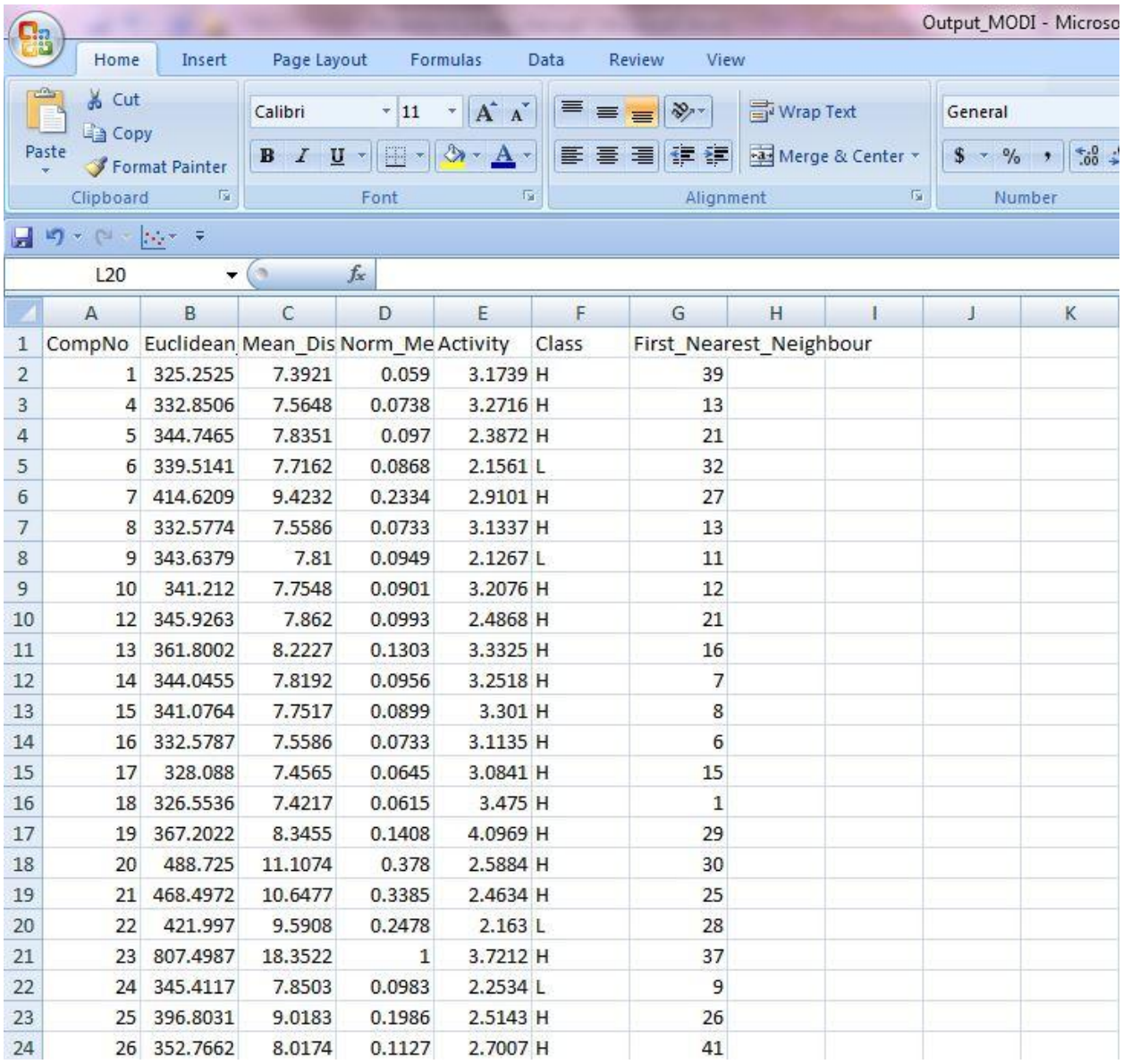
N_i^{Total}:Total no. of compounds belonging to class i

Input file format: This program takes a input file in CSV format.



	A	B	C	D	E	F	G	H	I	J
1	CompNo	pkI(mM)	S_sCH3	S_ssCH2	S_dsCH	S_aaCH	S_dssC	S_aasC	S_aaaC	S_ssssC
2	1	3.173925	0.99786	-0.41755	1.05029	0.39528	0.68812	0.50956	-0.14907	0.17181
3	4	3.271646	-0.50166	-0.41755	1.00002	0.35905	0.68812	0.47428	-0.14907	0.17181
4	5	2.387216	2.11327	-0.41755	0.91396	-0.05687	0.10788	0.45357	-0.14907	0.17181
5	6	2.156145	-0.50166	-0.41755	0.87386	0.082	0.68812	0.31016	-0.14907	0.17181
6	7	2.910095	0.86771	-0.41755	0.93262	-0.34087	0.68812	0.48256	-0.14907	0.17181
7	8	3.133713	-0.50166	-0.41755	1.0653	0.35235	0.68812	0.51415	-0.14907	0.17181
8	9	2.126679	-0.50166	-0.41755	0.99624	0.22631	0.68812	0.3961	-0.14907	0.17181
9	10	3.207608	-0.50166	0.1371	1.08978	0.30244	0.68812	0.47101	-0.14907	0.17181
10	12	2.486782	-0.50166	-0.00186	0.88027	0.17143	0.68812	0.35283	-0.14907	0.17181
11	13	3.332547	-0.50166	1.13381	0.9997	0.37329	0.68812	0.44626	-0.14907	0.17181
12	14	3.251812	-0.50166	1.38075	1.28003	0.89713	0.68812	0.50564	-0.14907	0.17181
13	15	3.30103	-0.50166	1.06575	1.02894	0.21322	0.68812	0.42774	-0.14907	0.17181
14	16	3.113509	-0.50166	0.87495	0.80225	0.23967	0.68812	0.34321	-0.14907	0.17181
15	17	3.084073	-0.50166	0.41601	-1.13155	0.49867	0.68812	0.2697	-0.14907	0.17181
16	18	3.474955	-0.50166	-0.05991	-1.13155	0.66332	0.26145	0.36119	-0.14907	0.17181
17	19	4.09691	-0.50166	0.7498	-1.13155	-0.03814	0.20604	0.50185	-0.14907	0.17181
18	20	2.58838	-0.50166	-0.41755	-1.13155	-1.90135	0.68812	-3.23137	-0.14907	0.17181
19	21	2.463442	-0.50166	-0.41755	-1.13155	-1.90135	-1.52007	-2.88807	-0.14907	0.17181
20	22	2.163043	0.55486	-0.41755	-1.13155	-1.90135	-3.09389	-2.6383	-0.14907	0.17181
21	23	3.721246	-0.50166	-0.41755	-1.13155	-1.90135	0.68812	-0.60833	-0.14907	-6.52063
22	24	2.253366	-0.50166	-0.41755	-1.13155	-0.18185	0.03831	0.36653	-0.14907	0.17181

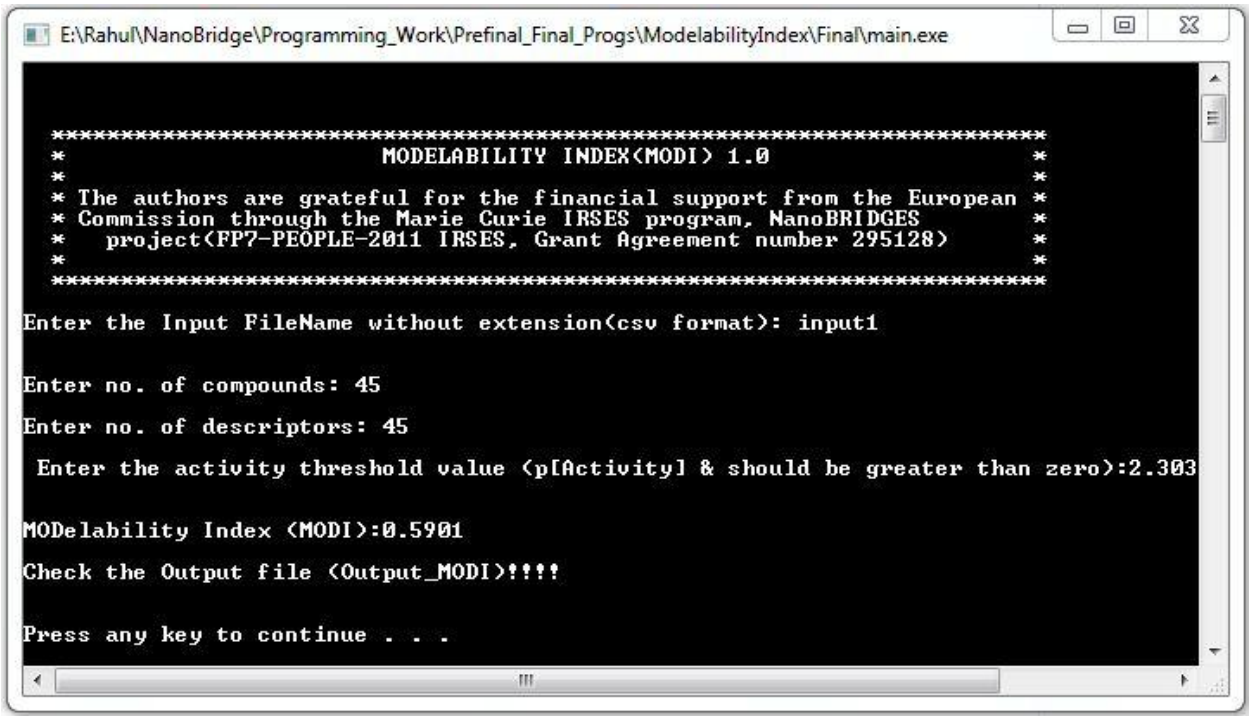
Output file format:



	A	B	C	D	E	F	G	H	I	J	K
1	CompNo	Euclidean	Mean_Dis	Norm_Me	Activity	Class	First_Nearest_Neighbour				
2	1	325.2525	7.3921	0.059	3.1739	H	39				
3	4	332.8506	7.5648	0.0738	3.2716	H	13				
4	5	344.7465	7.8351	0.097	2.3872	H	21				
5	6	339.5141	7.7162	0.0868	2.1561	L	32				
6	7	414.6209	9.4232	0.2334	2.9101	H	27				
7	8	332.5774	7.5586	0.0733	3.1337	H	13				
8	9	343.6379	7.81	0.0949	2.1267	L	11				
9	10	341.212	7.7548	0.0901	3.2076	H	12				
10	12	345.9263	7.862	0.0993	2.4868	H	21				
11	13	361.8002	8.2227	0.1303	3.3325	H	16				
12	14	344.0455	7.8192	0.0956	3.2518	H	7				
13	15	341.0764	7.7517	0.0899	3.301	H	8				
14	16	332.5787	7.5586	0.0733	3.1135	H	6				
15	17	328.088	7.4565	0.0645	3.0841	H	15				
16	18	326.5536	7.4217	0.0615	3.475	H	1				
17	19	367.2022	8.3455	0.1408	4.0969	H	29				
18	20	488.725	11.1074	0.378	2.5884	H	30				
19	21	468.4972	10.6477	0.3385	2.4634	H	25				
20	22	421.997	9.5908	0.2478	2.163	L	28				
21	23	807.4987	18.3522	1	3.7212	H	37				
22	24	345.4117	7.8503	0.0983	2.2534	L	9				
23	25	396.8031	9.0183	0.1986	2.5143	H	26				
24	26	352.7662	8.0174	0.1127	2.7007	H	41				

48	
49	TruePositives: 23.0000 TrueNegatives: 6.0000
50	
51	Total Actives/Toxic (H): 32.0000 Total Less_Actives/LessToxic (L): 13.0000
52	
53	MODelability Index (MODI): 0.5901
54	(Threshold value MODI: 0.65)
55	

How to run a program:



Disclaimer

For academic purpose only.

The program Modelability Indexhas been developed in C++ language and is validated on the known data sets. This program is compatible with both 32- and 64-bit Windows operating system. 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

Rahul Balasaheb Aher,
Phd Research Scholar,
Drug Theoretics and Cheminformatics Lab.,
Dept. of Pharmaceutical Technology,
Jadavpur University,
Kolkata, West Bengal,
INDIA-700032
E-mail Id: rahulba26@gmail.com(*for any queries regarding the program)

References:

1. Golmohammadi, Hassan, Zahra Dashtbozorgi, and William E. Acree Jr. "Quantitative structure–activity relationship prediction of blood-to-brain partitioning behavior using support vector machine." *European Journal of Pharmaceutical Sciences* 47, no. 2 (2012): 421-429.
2. Golbraikh, Alexander, Eugene Muratov, Denis Fourches, and Alexander Tropsha. "Data set modelability by QSAR." *Journal of chemical information and modeling* 54, no. 1 (2014): 1-4.