IS SIMPLE RANDOMIZATION OF COMPOUNDS IN TRAINING AND TEST SET AS GOOD AS OTHER METHODS USED IN QUANTITATIVE STRUCTURE-ACTIVITY EXPERIMENTS?

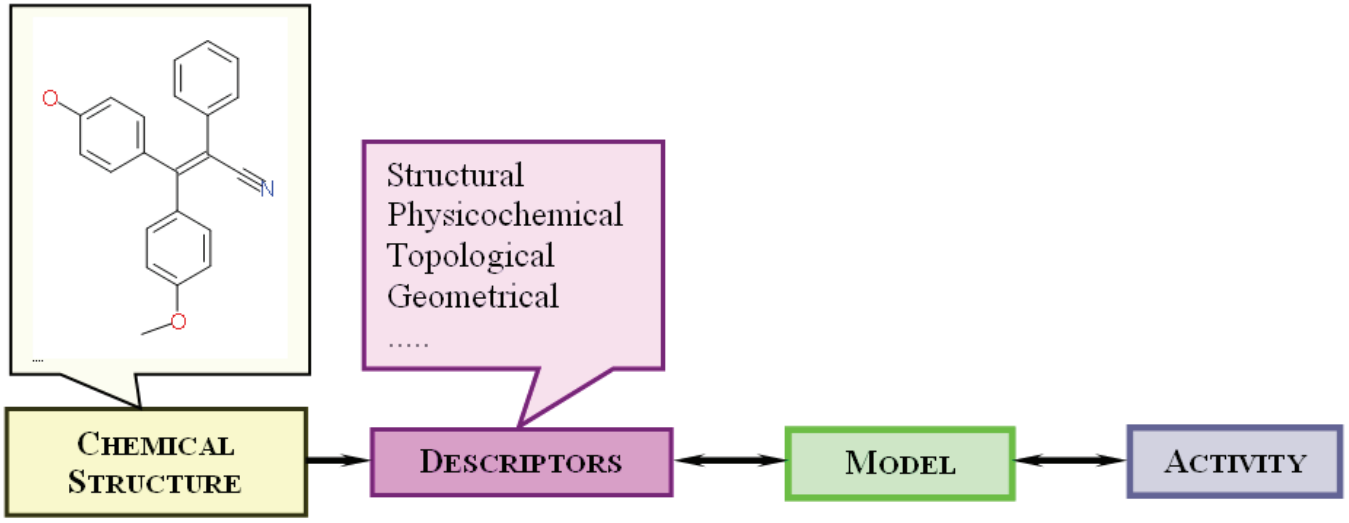
Sorana D. BOLBOACĂ & Lorentz JÄNTSCHI

"Iuliu Haţieganu" University of Medicine and Pharmacy Cluj-Napoca, RO
Technical University of Cluj-Napoca, RO

OUTLINE

- STRUCTURE-ACTIVITY/PROPERTY RELATIONSHIPS
- IS SIMPLE RANDOM SAMPLING A PROPER METHOD FOR SPLITTING THE SET OF COMPOUNDS IN TRAINING AND TEST?
  - MATERIAL & METHOD
  - RESULTS
  - CONCLUSION

STRUCTURE-ACTIVITY/PROPERTY RELATIONSHIPS



- General model:

$$Activity = f(structural\ properties)$$

→ used to predict the biological response of other chemical structures

- Base assumption: similar molecules have similar activities **??? – SAR PARADOX**

STRUCTURE-ACTIVITY/PROPERTY RELATIONSHIPS

- Good model:
  - quality of biological data
  - the choice of descriptors
  - statistical methods

  → accurate and reliable predictions of biological activities of new compounds

- Model validation:
  - internal validation or cross-validation
  - **validation by dividing the data set into training and test compounds**
  - true external validation by application of model on external data
  - data randomization or Y-scrambling

IS SIMPLE RANDOM SAMPLING A PROPER METHOD FOR SPLITTING THE SET OF COMPOUNDS IN TRAINING AND TEST?

MATERIAL

	Class	n	Activity
1	Drug-like compounds (Liu et al., 2004; Narayanan et al., 2005; Rose et al., 2003)	83	Blood-brain barrier permeation
2	Sulfonamide derivatives (Eroğlu et al., 2007)	18	Carbonic anhydrase II Inhibitory activity
3	Taxoids (Morita et al., 1997)	34	Cell growth Inhibitory activity
4	Triphenylacrylonitriles (Mukherjee et al., 2007)	25	Estrogen receptor Affinity

MODELS DEVELOPMENT

- Measured/observed activity (dependent variable): previously reported research
- Structural descriptors calculation (independent variables): **M**OLECULAR **D**ESCRIPTORS **F**AMILY ON **V**ERTICES (Bolboaca and Jantschi, 2009)
- Statistical method: **M**ultiple **L**inear **M**odels
- Training set: used to develop the model (internal validation of the model – leave-one-out analysis)
- Test set: used to test the model (external validation of the model)

MODELS DEVELOPMENT

- Set division:
  - Observed/Measured activity
  - Randomization algorithm
    - Assuring the normal distribution of dependent variable (Fisher, 1922)
  - Training: 75%

- Division reliability:
  - Dependent & Independent variables
  - Generalized cluster analysis - K-means algorithm (Statistica 8 software)
    - Euclidian distance
    - Maximization of the initial distance in regards of cluster center (cross-validation with 10-folds)

RESULTS: RANDOMIZATION

	Class (n)	$n_{Training}$	n_{Test}
1	Drug-like compounds (83)	55	28
2	Sulfonamide derivatives (18)	12	6
3	Taxoids (34)	23	11
4	Triphenylacrylonitriles (25)	19	6

	Set (n)	Training			Test		
		KS	AD	CS	KS	AD	CS
1	Drug-like compounds (83)	No	No	No	No	No	No
2	Sulfonamide derivatives (18)	No	No	No	No	No	n.a.
3	Taxoids (34)	No	No	No	No	No	n.a.
4	Triphenylacrylonitriles (24)	No	No	No	No	No	n.a.

RESULTS: MODELS TRAINING SETS

	1 (n=55, v=4)	2 (n=12, v=3)	3 (n=23, v=3)	4 (n=19, v=6)
R	0.7829	0.9970	0.9804	0.9707
R^2_a	0.5820	0.9918	0.9551	0.9307
s	0.58	0.08	0.25	0.35
F (p)	19 ($8.1 \cdot 10^{-8}$)	444 ($3.1 \cdot 10^{-9}$)	157 ($1.4 \cdot 10^{-13}$)	82 ($1.6 \cdot 10^{-9}$)
Q^2	0.5316	0.9911	0.9497	0.9112

RESULTS: SUMMARY OF GENERALIZED CLUSTER ANALYSIS

	Set (n)	Clust	$No_{comp.}$/clust	Y: F(p)
1	Drug-like compounds (83)	3	27:33:23	28 ($7.52 \cdot 10^{-10}$)
2	Sulfonamide derivatives (18)	5	7:4:2:1:4	211 ($1.12 \cdot 10^{-11}$)
3	Taxoids (34)	5	2:1:21:5:5	89 ($6.66 \cdot 10^{-16}$)
4	Triphenylacrylonitriles (24)	6	6:5:7:4:1:2	18 ($1.22 \cdot 10^{-6}$)

RESULTS:

SUMMARY OF GENERALIZED CLUSTER ANALYSIS

Class	Set|Cluster	1	2	3	4	5	6	∑
Drug-like compounds	training	17	22	16				***55***
	test	10	11	7				***28***
Sulfonamide	training	6	2	1	1	2		***12***
	test	1	2	1	0	2		***6***
Taxoids	training	1	0	16	2	4		***23***
	test	1	1	5	3	1		***11***
Triphenylacrylonitriles	training	5	4	5	3	1	1	***19***
	test	1	1	2	1		1	***6***

RESULTS:

SUMMARY OF GENERALIZED CLUSTER ANALYSIS

- Z test for comparing two-proportion:
  - matrix of p-values

Class|Cluster	1	2	3	4	5	6
Drug-like compounds	0.6601	0.9503	0.6949			
Sulfonamide	0.1904	0.4346	0.6029	0.4775	0.4346	
Taxoids	0.5865	0.1520	0.1852	0.1623	0.5272	
Triphenylacrylonitriles	0.6328	0.8172	0.7419	0.9596	0.5720	0.3784

RESULTS: DRUG-LIKE COMPOUNDS

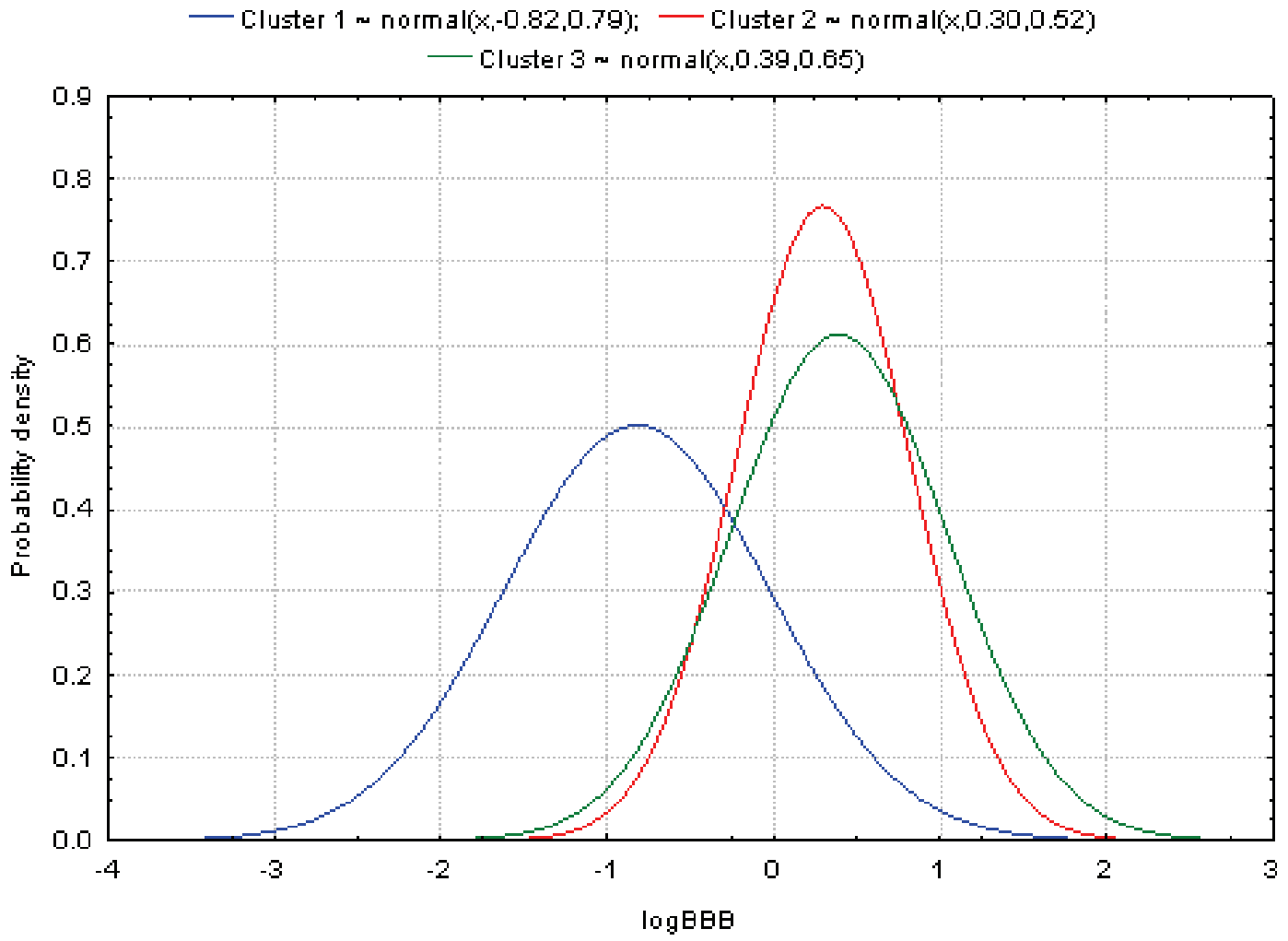
- Statistics for log BBB

	Cluster 1	Cluster 2	Cluster 3	Overall
Minimum	-2.70000	-1.30000	-1.23000	-2.70000
Maximum	0.42000	1.04000	1.44000	1.44000
Mean	-0.82133	0.29779	0.39170	-0.04024
Standard deviation	0.80995	0.52809	0.66701	0.73540

- ANOVA

	Between SS	df	Within SS	df	F	p value
Y	24.71	2	35.7682	80	27.633	7.52E-10
TQXIPADL	175.04	2	31.2604	80	223.980	0.00E-01
TQ5APIDL	128.55	2	179.9356	80	28.577	4.32E-10
GLWACPDR	100.97	2	8.9669	80	450.424	0.00E-01
GLQIIFDL	14413.65	2	186.3928	80	3093.177	0.00E-01

RESULTS: DRUG-LIKE COMPOUNDS

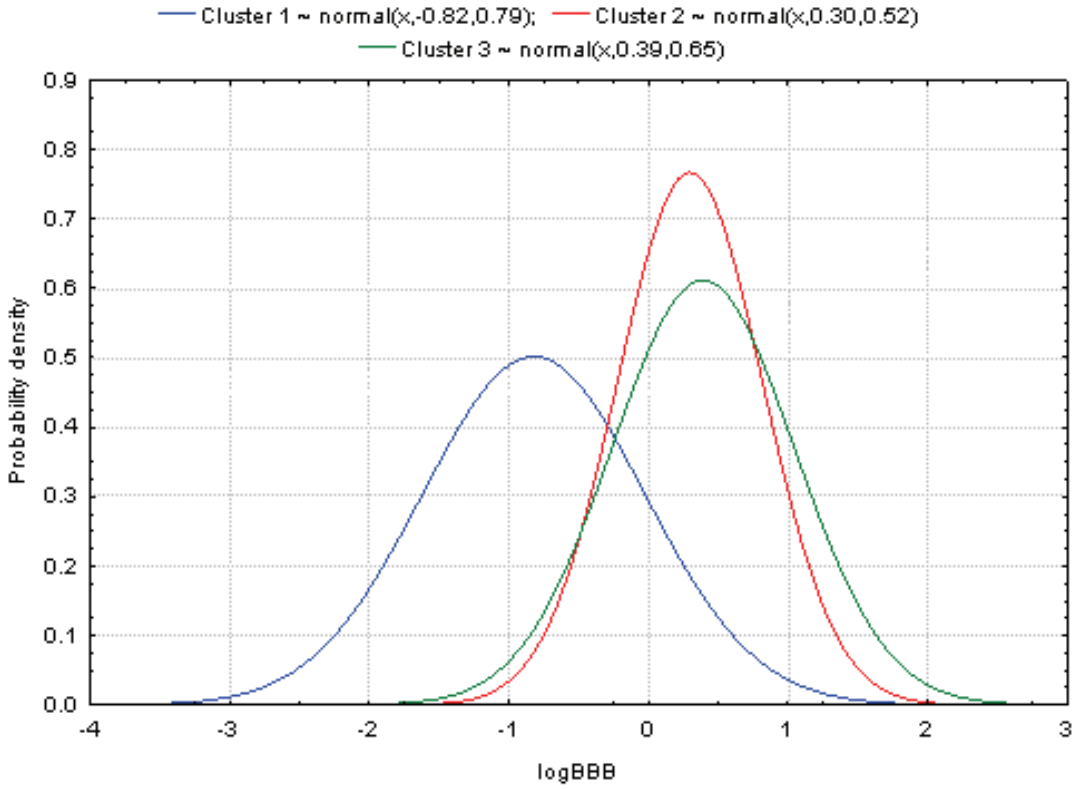


SUMMARY

- Observed/Measured activity:
  - Simple randomization in training and test sets
- QSAR MLR equation with MDFV descriptors
- Testing the randomization with generalized cluster analysis:
  - Observed/Measured activity & MDFV descriptors

CONCLUSION

- **IS SIMPLE RANDOM SAMPLING A PROPER METHOD FOR SPLITTING THE SET OF COMPOUNDS IN TRAINING AND TEST?**



REFERENCES

- Bolboacă SD, Jäntschi L. Comparison of QSAR Performances on Carboquinone Derivatives. TheScientificWorldJOURNAL 2009;9(10):1148-1166.
- Eroğlu E, Türkmen H, Güler S, Palaz S, Oltulu O. A DFT-Based QSARs Study of Acetazolamide/Sulfanilamide Derivatives with Carbonic Anhydrase (CA-II) Isozyme Inhibitory Activity. International Journal of Molecular Sciences 2007; 8(2):145-155.
- Fisher RA. Statistical Methods for Research Workers, 8th ed. Oliver and Boyd, London, 1941.
- Liu X, Tu M, Kelly RS, Chen C, Smith BJ. Development of a Computational Approach to Predict Blood-Brain Barrier Permeability. Drug Metabolism and Disposition 2004;32:132-139.

REFERENCES

- Morita H, Gonda A, Wei L, Takeya K, Itokawa H. 3D QSAR analysis of taxoids from Taxus cuspidate var. nana by comparative molecular field approach. Bioorg Med Chem Lett 1997; 7: 2387-2392.
- Mukherjee, S., Nagar, S., Mullick, S., Mukherjee, A., Saha, A. Pharmacophore mapping of selective binding affinity of estrogen modulators through classical and space modeling approaches: Exploration of bridged-cyclic compounds with diarylethylene linkage. Journal of Chemical Information and Modeling 2007;47(2):475-487.
- Narayanan R , Gunturi SB. In silico ADME modelling: prediction models for blood–brain barrier permeation using a systematic variable selection method. Bioorganic & Medicinal Chemistry 2005;13:3017-3028.
- Rose K, Hall LH, Hall M, Kier LB. Modeling Blood-Brain Barrier Partitioning Using Topological Structure Descriptors. MDL Information Systems. 2003.

Financial support: ERASMUS STAFF MOBILITY