MATCH
Communications in Mathematical
and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **63** (2010) 585-599

ISSN 0340 - 6253

Optimal Partition of Datasets of QSPR Studies: A Sampling Problem

Alan Talevi[1],, Carolina L. Bellera[1], Eduardo A. Castro[2], Luis E. Bruno-Blanch[1].*

1 Medicinal Chemistry, Department of Biological Sciences, Faculty of Exact Sciences, National University of La Plata (UNLP), 47 y 115, La Plata (B1900AVV), Buenos Aires, Argentina. Phone: 542214235333 ext 41, E-mail: atalevi@biol.unlp.edu.ar

[2] Instituto de Investigaciones Fisicoquímicas Teóricas y Aplicadas (INIFTA), Department of chemistry, Faculty of Exact Sciences, National University of La Plata (UNLP), CCT La Plata CONICET

(Received December 23, 2009)

ABSTRACT: Starting from different partitions of a 160-compounds dataset into training and test sets, we developed discriminant functions to classify drugs into different categories of human intestinal absorption rate. For each partition of the dataset, models that included up to ten Dragon descriptors were built, and the performance of each discriminant function in the classification of the training and test sets was assessed. The classification ability of the model on both the training and test sets of each partition was assessed and explored graphically through divergence diagrams. Results suggest that external validation tends to underestimate the predictive capability of QSAR models and that the more reliable results from external validation are obtained with even partitions of small and medium size datasets.

Introduction

The motivation behind any modeling effort is to infer a predictive model from a limited sample, in order to apply it later in the prediction of a property or behavior in a wider population of objects or individuals. A critical question that the modeler shall answer is: what population is the training sample of the model representative of? There are two critical processes that should be taken into account to answer this issue: assessing the applicability domain of the model[1], and avoiding overfitting[2]. Estimating the applicability domain of the model is equivalent to define which cases or individuals (external to those in the training sample) can be reliably predicted by the model. Avoiding overfitting means to assure that parsimony principle has been obeyed and that excessive fitting of the sample data has not occurred at the expense of generalizability, that is, the predictive capability on the general population on which the model is supposed to be applied. Validation procedures are the tools of choice to evaluate if overfitting has occurred and to assess the model general predictive capability.

The two most used validation procedures are external validation and internal cross-validation (e.g. leave-one-out and leave-group-out cross validation). External validation (holding out a fraction of the dataset as an independent test set) is generally considered the most important step to measure the robustness and predictive capability of a QSAR model[3]. However, it is also been pointed out that external validation is only reliable when using large hold-out samples and that reserving a fraction of the dataset for external validation may be a waste of useful information in the context of QSAR modeling, where often only small or medium-sized datasets are available[4,5]. Internal cross-validation is usually indicated as a good, trustworthy alternative to assess model predictive capability[4,5]. Nevertheless, several systematic studies demonstrate that cross-validated r^2 (i.e. q^2) tends to be an overoptimistic measure of the general predictive ability of the model[6-8]. According to those reports, a high q^2 is a necessary but not sufficient condition for a model to have high predictive power and external validation might be mandatory, or else new definitions of q^2 should be applied.

In this short-communication we study different partitions of a dataset of 160 compounds used to derive classificatory models of human intestinal absorption rate (%HIA), finding new evidence of the inability of small hold-out samples to reliably assess the predictive power of a QSAR model. According to our findings, external validation might tend to underestimate the general predictive capability of a model when using scarce test sets. We also propose the use of ‘divergence diagrams’ for ease, visual evaluation of the risk of overfitting in a QSAR modeling campaign.

Methods

Dataset. As a part of our ongoing research to model physicochemical properties related to drug oral bioavailability[9,10], we have designed a 160-compounds dataset gathered from previous reported models of drug intestinal absorption[11-14]. This study, however, is not focused in the development of reliable QSAR models to predict %HIA: the purpose of this communication is to study if external validation is indeed a trustworthy validation tool to assess the generalizability of a model and how shall a medium size dataset be partitioned into training and test sets in order to reduce the chances of overfitting.

We considered four categories of %HIA: category one: %HIA ≤ 20%; category two: 20% < %HIA ≤ 50%; category three: 50% < %HIA ≤ 80% and; category four: %HIA > 80%, and we tried to obtain a balanced distribution of the 160 compounds among these categories. The 160 selected compounds are distributed as follows: category one, 46 compounds; category two, 26 compounds; category three, 41 compounds; category four, 47 compounds. This is not an entirely even distribution among the four categories (mainly due to limited experimental data on low permeability compounds) but is not either highly biased towards very permeable compounds, as has been observed in several previous modeling efforts of %HIA[11-23]. The %HIA experimental values are checked in Hazardous Substances Data Bank (HSDB) and Pubchem[24,25]; we have not included in the dataset compounds that according to different sources belong to different of the four categories considered here. The list of compounds that compose the dataset and their correspondent %HIA values can be found in Table 1. Three different partitions of the dataset were considered in order to define the optimal partition to get a reliable external validation: 120:40, 80:80 and 40:120. In the three cases, the compounds that compose the dataset were sorted by category and by alphabetical order and split into training and test sets through systematic random sampling: for the 120:40 partition, one each four alphabetically sorted compounds was extracted to the test set; for the 80:80 partition, one each two compounds was extracted to the test set; for the 40:120 partition, three each four compounds were extracted to the test set. None periodic pattern exists between the name of the compounds used in this study and their chemical structure or pharmacological activity. Average intermolecular Tanimoto similarities (based in atom pairs) between the training and test sets were calculated for each of the three considered partitions. PowerMV software (developed and freely provided by the National Institute of Statistical Sciences) was used for intermolecular distances calculations[26].

Descriptor calculation and discriminant functions modeling

Two descriptor sets calculated through Dragon software[27] were considered. The first descriptor set (set 1) was composed of information indices, topological charge indices, atom-centered fragments and functional group counts; the second set (set 2) was composed of constitutional descriptors, topological descriptors, information indices and topological charge indices (Dragon’s descriptor classification and nomenclature is kept in this report). These descriptors are 0D-2D and thus no conformational optimization or conformational analysis is required.

The General Discriminant Analysis Modeling module of Statistica 7.0[28] was applied to obtain four-categories discriminant functions. Stepwise forward was applied as variable selection technique. The best models obtained through this methodology including between 1 and 10 descriptors were considered; the procedure was stopped when no descriptor with an associated p-value below 0.05 could be incorporated into the model. Descriptors with constant value along all the compounds of the training sets were excluded from the analysis.

Table 1. Name of the compounds that compose the dataset; in the right column the correspondent %HIA value of each compound is presented.

Category 1		Streptomycin	1	Cycloserine	73	Atropine	98
Acamprosate	11	Streptosozin	0	Dipyridamole	58	Benzydamine	87
Acarbose	2	Succinylsulfathiazole	5	Eflornithine	55	Betaxolol	90
Adefovir	12	Ticarcillin	5	Enalapril maleate	66	Bupropion	87
Amygdalin	5	Tobramycin	0	Ethambutol	80	Caffeine	100
Anphotericin b	5	Vancomycin	0	Etodolac	70	Chloramphenicol	90
Arbekacin	0	**Category 2**		Famciclovir	70	Clofibrate	87
Azlocillin	0	AAFC	32	Fenoterol	60	Codeine	95
Aztreonam	1	Amiloride	50	Furosemide	61	Diclofenac	100
Cefodizime	0	Azithromycin	36	Guanabenz	80	Disulfiram	97

Ceftriaxone	1	Benazepril	37	Hydrochlothiazide	65	Felbamate	90
Cefuroxime	5	Bromocriptine	28	Isocarboxazid	70	Fluconazole	96.5
Cidofovir	3	Cefpodoximeproxetyl	50	Ivermectin	60	Hydrocortisone	91
Cromolyn	0.5	Chlorothiazide	23.8	Metformin	53	Ibuprofen	100
Doxorubicin	5	Cymarin	47	Metolazone	63	Indomethacin	100
Edetic acid	5	Dihydroergotamine	35	Mianserin	70	Ketoprofen	92
Foscarnet	17	Famotidine	38	Mibefradil	69	Ketorolac	90
Ganciclovir	3.6	Flucloxacillin	40	Moxisylyte	70	Labetalol	95
Gentamycin	0	Fosfomycin	31	Oxycodone	60	Lamivudine	87
Imipenem	5	Fosmidomycin	30	Oxytetracycline	58	Lansoprazol	85
Iohexol	5	Guanoxan	50	Pimozide	70	Minoxidil	98
Iothalamate	1.9	Lincomycin	27.5	Propylthiouracil	75	Moricizine	88
Iotroxic acid	5	Lisinopril	25	Pyrbuterol	60	Moxonidine	88
Kanamycin	1	Lovastatin	30.5	Quetiapine	73	Naproxen	99
K-strophanthoside	16	Metaproterenol	44	Ramipril	60	Nitrendipine	88
Lactulose	0.6	Methyldopa	41	Ranitidine	52.8	Nordiazepam	99
Lucifer yellow	0	Nadolol	31	Recainam	71	Oxyfedrine	85
Mannitol	16	Pafenolol	29	Reproterol	60	Propanolol	99
Meropenem	0	Pravastatin	34	Terbutaline	62	Rivastigmine	100
Mezlocillin	0	Rimiterol	48	Tolrestat	66	Saccharin	88

Mitoxanthrone	5	Sulpiride	44	Urapidil	78	Sotalol	95
Moexiprildiacid	5	Trandolapril	50	Valsartan	55	Sultopride	89
Nedocromil	3	Zonavir	28	Ziprasidone	60	Tenidap	89
Neomycin	1	**Category 3**		**Category 4**		Timolol	95
Netilmycin	0	Almotriptan	75	Acebutolol	89.8	Tolbutamide	85
Olsalazine	2.3	Anagrelide	70	Acetaminophen	85	Trapidil	96
Ouabain	1.4	Atenolol	51	Almitrine	90	Trimethoprim	97
Pamidronic acid	5	Benserazide	70	Alprenolol	93.8	Warfarin	98
Pentamidine	0	Benzbromarone	73	Aminopyrine	100	4zalcitabine	85
Phthalylsulfathiazole	5	Bromhexine	70	Amoxicillin	93.8		
Raffinose	0.3	Captopril	68	Antipyrine	100		
Risedronic acid	1	Cefatrizine	76	Aspirin	100		

Results

Table 2 presents the descriptors included into the models derived from both descriptors' sets (set 1 and 2) for the three partitions of the dataset considered (120:40, 80:80, 40:120). Descriptors are presented in the same order as they have been included in the models through the stepwise procedure. Although a maximum of ten steps was allowed in the Stepwise Forward procedure, it can be noted that for set one the models were truncated at 8 descriptors for partitions 120:40 and 80:80 and at 6 descriptors for partition 40:120 (non descriptor from the descriptors pool with p-value below 0.05 could be further added into the model after those steps). For set 2, the models were truncated at 6 descriptors for partition 40:120, 7 descriptors for partition 120:40 and 10 descriptors for partition 80:80.

Table 2. Details on the descriptors added into the discriminant functions derived from both sets of descriptors. The symbol given by Dragon for each descriptor is kept; between parentheses we present the descriptor definition from Dragon. The p-value associated to each descriptor in the final model is showed.

	Descriptors' set 1		**Descriptors' set 2**	
Partition	**Descriptor**	**p-value**	**Descriptor**	**p-value**
120:40	nHAcc (number of acceptor atoms of H-bonds)	0.0000	nO (number of Oxygen atoms)	0.0000
	H-052 (H attached to C0 sp3 with 1X attached to next C)	0.0000	nH (number of Hydrogen atoms)	0.0000
	O-057 (phenol/enol/carboxyl OH)	0.0000	nN (number of Nitrogen atoms)	0.0000
	CIC5 (complementary information content – neighborhood symmetry of 5-order)	0.0019	T(O..O) (sum of topological distances between Oxygen atoms)	0.0002
	BIC3 (bond information content - neighborhood symmetry of 3-order)	0.0205	BIC0	0.0007
	C-002 (count of CH_2R_2)	0.0017	MSD (mean square distance index)	0.0046
	BIC0 (bond information content - neighborhood symmetry of 0-order)	0.0299	D/Dr12 (distance/detour ring index of order 12)	0.0208
	nNO2 (number of nitro groups)	0.0470		
80:80	nHAcc	0.0000	nO	0.0000
	IAC (total information index of atomic composition)	0.0159	nH	0.0000
	C-002	0.0029	JGI5 (mean topological charge index of order 5)	0.0250
	nCOOH (number of aliphatic carboxylic acids)	0.0009	nN	0.0057
	H-047 (H attached to C1 sp3/ C0 sp2)	0.0014	T(O..O)	0.0006
	N-070 (number of Ar-N-Al)	0.0074	BIC1 (bond information content - neighborhood symmetry of 1-order)	0.0088
	H-055 (H attached to C0 sp3 with 4X attached to next C)	0.0079	T(S..Cl) (sum of topological distances between Sulfur and Chloro atoms)	0.0148
	nSO3H (number of sulfonic acids)	0.0297	T(O..Cl) (sum of topological distances between Oxygen and	0.0049

			Chloro atoms)	
			PHI (Kier flexibility index)	0.0183
			D/Dr12	0.0490
40:120	nHAcc	0.0000	Ms (mean electrotopological state)	0.0000
	IAC	0.0005	MW (molecular weight)	0.0000
	N-073 (Ar2NH, Ar3N, Ar2NAl, R..N..R)	0.0046	GGI5 (topological charge index of order 5)	0.0002
	JGI6 (mean topological charge index of order 6)	0.0003	nX (number of halogen atoms)	0.0000
	H-055	0.0049	CENT (centralization)	0.0008
	nCrHR (number of ring tertiary C sp3)	0.0396	JGI2 (mean topological charge index of order 2)	0.0368

Cn refers to a Carbon atom attached to n heteroatoms and presenting the indicated hybridization; R represents any group linked through Carbon; X represents any electronegative atoms (O, N, S, P, Se, halogens); Al and Ar represent aliphatic and aromatic groups, in that order.

Figures 1 and 2 present a comparison of the performance of the discriminant functions derived through the Stepwise Forward procedure in the classification of the training and test sets of the three considered partitions, for descriptors sets 1 and 2, respectively. The performance is assessed as the global percentage of good classifications. For the generation of the models we have considered that the a priori probability for any compound to belong to a given category was 25%. Although this is not true for the dataset employed in this study (we know beforehand that the number of compounds in each category is not the same for all categories), we prefer to adopt this hypothesis thinking of the general application of the models to the general population (nothing indicates that in the general population the number of high permeability compounds exceeds the number of low permeability ones). Adopting different a priori probabilities according to the known distribution of %HIA in the dataset would have allowed us to obtain better performances on the dataset classification, but would have not been adequate in a real application in which the true distribution of the modeled property in the general population is not known for sure. Note that, since four classes of compounds are considered in the discriminant analysis with even a priori probabilities across all the four categories, random classification would have resulted in about 25% of good classifications.

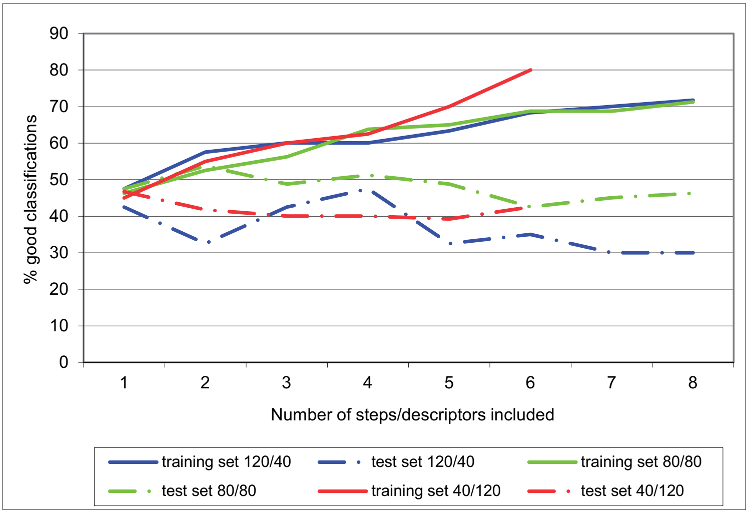

Fig. 1. Comparison of the performance of the models derived from set 1, for the three considered partitions and both the correspondent training and test sets. The percentage of good classifications in the training set is showed with a continuous line while the percentage in the test set is showed with a dashed line (for each of the three partitions).

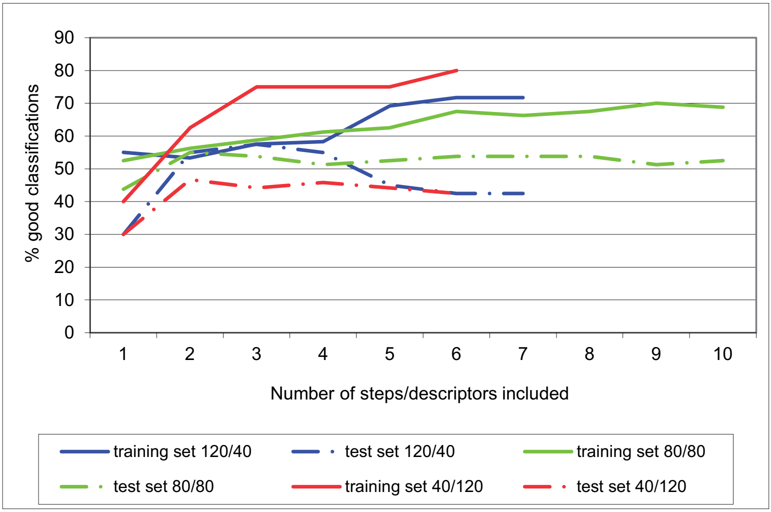

Fig. 2. Comparison of the performance of the models derived from set 2, for the three considered partitions and both the correspondent training and test sets.

Discussion

It can be seen in Figures 1 and 2 that the divergence between the percentages of correct classifications in the training and test sets are accentuated in the case of partition 120:40 and 40:120 of the dataset. In the case of an ideal model of general application (a model for which the ability to characterize the modeled property is identical in the sample that in the general population to which the model is thought to be applied) a perfect superposition between the curves correspondent to the training and test set (each pair of continuous and dashed lines) should be expected. There are three possible explanations for divergence between the pair of curves correspondent to the training and test sets for any given partition: a) the difference between the curves can be due to overfitting; b) the difference could be revealing a tendence of the external validation procedure to underestimate the actual performance of the models or; c) some of the compounds of the test set could be outside the applicability domain of the models, which is defined by the correspondent training set.

For both descriptors' sets and all the three partitions the pair of curves representing the performance on the training and test set converges when a small number of descriptors are included in the model. The divergence between the curves of training and test set is less pronounced in the case of the 80:80 partition (when half the dataset is randomly assigned to the training set and the other half to the test set). The point in which each pair of curves starts diverging corresponds to the possible –but yet uncertain- beginning of the overfitting (e.g., in Figure 2 the risk of overfitting clearly seems to appear when more than 4 descriptors are included in the model for the 120:40 partition).

Let us consider, however, the case of the 120:40 partition. In the worst case (when 8 and 7 descriptors are included in the discriminant functions derived from set 1 and 2, in that order) the ratios between the number of cases (compounds) and the number of predictors are 15.0 and 17.1 for the models derived from sets 1 and 2, which seems to indicate a low chance of overfitting (as Peduzzi et al. have long ago demonstrated, a model tends to be biased towards the true value of the modeled property in the sample when the number of events per independent variable is below 10[28,29]). In models with less descriptors, the cases to predictors ratio in this partition is quite higher than 10 and yet the curves for both the training and test sets are quite divergent (particularly for set 1 - Fig 1), with the proportion of good classifications on the training set being always above the one on the test set. Consistently with our results, Hawkins et al. have shown that for scarce (10 and 20-compounds) test sets the external validation tends to underestimate the actual performance of the models, while an external test set of 50 compounds gives quite reliable results[4]. When the number of compounds in the training set is wide above those in the test set (particularly if the test set is small) it is highly probable to find elements of the training set which are absent in the test set but that could (or could be not) be present in the general population. This explains why the explanatory power of the models on the training set compounds correlate well with the results of the external validation in the case of the 80:80 partition (specially for the second descriptors set): the number of compounds in the test set is quite high for this partition (above the minimum of 50 cases suggested by Hawkins) and the even partition of the dataset into training and test sets reduces the chances of observing high frequent features in the training set that are not reflected with a similar frequency in the test set. How can we explain the results on the 40:120 partition, then? In this case, overfitting is quite probable (since the events to predictors ratio is quite low because of the small number of compounds in the training set) which explains why the performance of the models on the 40-compounds training set is better than in the cases of the 80/80 and 120/40 partitions. The small number of compounds of this partition prevents us from finding multiple-predictor models of general applicability.

Similarity analysis through computation of the average Tanimoto distances (comparing all the possible pairs of compounds between the training and test sets of each partition) showed us that the average similarity between the training and test sets of each partition is quite similar (0.33 for the 120:40 partition; and 0.32 for both the 80:80 and the 40:120 partitions); this, along with the fact that partitions were generated randomly prevent us from thinking that the results could be explained through some bias in the generated training sets towards some specific chemical families of compounds or towards specific structural features. For the test set of the 120:40 partition, the number of neighbors with Tanimoto similarities above 0.7 is 67; for the test set of the 80:80 partition, 44; for the 40:120 partition, 21. This suggest that the lower divergence between the performance of the models on the training and test sets classification in the case of the 80:80 partition could only be explained by an applicability domain issue when comparing the 80:80 to the 40:120 partitions, but not when comparing the 80:80 to the 120:40 partition.

Conclusion

Our results seem to corroborate that external validation is a conservative approach to appraise the general predictive ability of a model, with some tendency to underestimate the true performance of the model on the general population, especially when unbalanced partitions of the datasets into training and test sets are considered for small and medium size datasets. This is to say, we can have certainty that the performance of a model in the general population will be equal or better than that observed in the external validation procedure. Since it has been already demonstrated by Golbraikh et al. that internal cross-validation tends to overestimate the true performance of the model, we may conclude that the true performance of the model may always be somewhere between the internal and external validation results. Therefore, external validation with an adequate test set (above 50 compounds, according to Hawkins) should be always used when possible to assess the worst possible situation.

It is important to underline that the ‘general population’ for any given QSAR model is given by the region of the chemical space defined by the training set, i.e. all the chemical compounds that belong to the applicability domain of the model. Therefore, results of the external validation and predictions on compounds independent from those in the dataset will be more reliable if proper assessment of the applicability domain has been performed.

The ‘divergence graphs’ presented in Figures 1 and 2 may be an useful tool to determine from which step of a stepwise process significant risk (but not certainty) of overfitting arises.

This study should be expanded in the future by considering other datasets and properties and including a more careful assessment of the applicability domain for each generated model.

Acknowledgment. A. Talevi and E.A. Castro thank CONICET. L.E. Bruno-Blanch thanks ANPCyT (PICT BID11985) and Incentivos UNLP.

REFERENCES

1. J. Jaworska, N. Nikolova–Jeliazkova, T. Aldenberg, QSAR applicability domain estimation by projection of the training set in descriptor space: a review, *ATLA – Altern. Lab. Anim.* **33** (2005) 445–459.
2. M. A. Babyak, What you see may not be what you get: a brief, nontechnical introduction to overfitting in regression–type models, *Psychosom. Med.* **66** (2004) 411–421.
3. A. Yasri, D. Hartsough, Toward an optimal procedure for variable selection and QSAR model building, *J. Chem. Inf. Comput. Sci.* **41** (2001) 1218–1227.
4. D. M. Hawkins, S.C. Basak, D. Mills, Assessing model fit by cross–validation, *J. Chem. Inf. Comput. Sci*. **43** (2003) 579–586.
5. D. M. Hawkins, The problem of overfitting, *J. Chem. Inf. Comput. Sci.* **44** (2004) 1–12.
6. A. Golbraikh, A. Tropsha, Beware of q^2!, *J. Mol. Graph. Model.* **20** (2002) 269–276.
7. A. Golbraikh, M. Shen, Z. Xiao, Y. Xiao, K. Lee, A. Tropsha, Rational selection of training and test sets for the development of validated QSAR models, *J. Comput. Aid. Mol. Des.* **17** (2003) 241–253.
8. G. Schüürmann, R. Ebert, J. Chen, B. Wang, R. Kühner, External validation and prediction employing the predictive squared correlation coefficient – test set activity mean vs training set activity mean, *J. Chem. Inf. Model.* **48** (2008) 2140–2145.
9. P.R. Duchowicz, A. Talevi, C. Bellera, L. E. Bruno–Blanch, E. A. Castro, Application of descriptors based on Lipinski´s rules in the QSPR study of aqueous solubilities, *Bioorg. Med. Chem.* **15** (2007) 3711–3719.
10. P.R. Duchowicz, A. Talevi, L. E. Bruno–Blanch, E. A. Castro, New QSPR study for the prediction of aqueous solubility of drug-like compounds, *Bioorg. Med. Chem.* **16** (2008) 7944–7955.
11. E. Deconinck, H. Ates, N. Callebaut, Y. Van Gyseghemb, Y. Vander Heyden, Evaluation of chromatographic descriptors for the prediction of gastro-intestinal absorption of drugs, *J. Cromatograph. A.* **1138** (2007) 190–202.
12. S. Agatonocix–Kustrin, R. Beresford, A. Pauzi, M. Yusof, Theoretically–derived molecular descriptors important in human intestinal absorption, *J. Pharmaceut. Biomed. Anal.* **25** (2001) 227–237.

13. T. Hou, J. Wang, J. Y. Li, ADME evaluation in drug discovery. 8. The prediction of human intestinal absorption by a support vector machine, *J. Chem. Inf. Model.* **47** (2007) 2408–2415.
14. E. Deconinck, T. Hancock, D. Cooman, D. L. Massart, Y. Vander Heyden, Classification of drugs in absorption classes using the classification and regression trees (CART) methodology, *J. Pharmaceut. Biomed. Anal.* **39** (2005) 91–103.
15. A. Yan, Z. Wang, Z. Cai, Prediction of human intestinal absorption by GA feature selection and support vector machine regression, *Int. J. Mol. Sci.* **9** (2008) 1961–1976.
16. T. E. Yen, S. Agatonovic–Kustrin, A. M. Evans, R. L. Nation, J. Ryand, Prediction of drug absorption based on immobilized artificial membrane (IAM) chromatography separation and calculated molecular descriptors, *J. Pharmaceut. Biomed. Anal.* **38** (2005) 472–478.
17. H. A. Sun, A universal molecular descriptor system for prediction of logP, logS, logBB, and absorption, *J. Chem. Inf. Comput. Sci.* **44** (2004) 748–757.
18. M. A. Cabrera Pérez, M. Bermejo Sanz, L. Ramos Torres, R. Grau Ávalos, M. Pérez González, H. González Díaz, A topological sub–structural approach for predicting human intestinal absorption of drugs, *Eur. J. Med. Chem.* **39** (2004) 905–916.
19. P. R. N. Wolohan, R. D. Clark, Predicting drug pharmacokinetic properties using molecular interaction fields and SIMCA, *J. Comput. Aid. Mol. Des.* **17** (2003) 65–76.
20. T. Niwa, Using general regression and probabilistic neural networks to predict human intestinal absorption with topological descriptors derived from two–dimensional chemical structures, *J. Chem. Inf. Comput. Sci.* **43** (2003) 113–119.
21. G. Klopman, L. R. Stefan, R. D. Saiakhov, ADME evaluation 2. A computer model for the prediction of intestinal absorption in humans,*Eur. J. Pharm. Sci.* **17** (2002) 253–263.
22. M. H. Abraham, Y. H. Zhao, J. Le, A. Hersey, C. N. Luscombe, D. P. Reynolds, G. Beck, B. Sherborne, I. Cooper, On the mechanism of human intestinal absorption, *Eur. J. Med. Chem.* **37** (2002) 595–605.
23. M. D. Wessel, P. C. Jurs, J. W. Tolan, S. M. Muskal, Prediction of human intestinal absorption of drug compounds from molecular structure, *J. Chem. Inf. Comput. Sci.* **38** (1998) 726–735.
24. Hazardous Substances Data Bank (HSDB) – Toxicology Data Network (TOXNET). US National Library of Medicine. http:// toxnet.nlm.nih.gov/
25. The Pubchem Project. US National Library of Medicine. http://pubchem.ncbi.nlm.nih.gov/

26. K. Liu, J. Feng, S. S. Young, PowerMV: a software environment for molecular viewing, descriptor generation, data analysis and hit evaluation, *J. Chem. Inf. Model.* **45** (2005) 511–522.
27. Dragon Academic v. 4.0. Milano Chemometrics (2003).
28. Statistica v. 7.0. Statsoft Inc (2004).
29. P. N. Peduzzi, J. Concato, E. Kemper, T. R. Holford, A. R. Feinstein. A simulation study of the number of events per variable in logistic regression analysis, *J. Clin. Epidemiol.* **49** (1996) 1373–1379.
30. P. N. Peduzzi, J. Concato, A. R. Feinstein, T. R. Holford, Importance of events per independent variable in proportional hazards regression analysis. II. Accuracy and precision of regression estimates, *J. Clin. Epidemiol.* **48** (1995) 1503–1510.