#29

Data Mining on Structure-Activity/Property Relationships Models

Sorana Daniela BOLBOACĂ, Lorentz JÄNTSCHI

"Iuliu Hațieganu" Medicine and Pharmacy University, Technical University, Cluj-Napoca, Romania

Keywords: Knowledge-Discovery in Database (KDD), Cluster analysis, Structure-Activity/Property Relationships (SAR/SPR), Molecular Descriptors Family (MDF)

Abstract

#29 + Intro Material Method Results Discussion Conclusion Ref

Molecular descriptors family on structure-activity/property relationships studies were carried out in order to identify the link between compounds structure and their activity/property. A number of fifty-five classes of properties or activities of different compounds sets were investigated. Single and multi-varied linear regression models using molecular descriptors as variables were identified. The models with estimation and prediction abilities and associated characteristics were stored into a database. A data mining analysis using classification and clustering were applied on the obtained database for searching and extracting useful information. The methodology applied in searching and extracting for information and the obtained results are presented.

Intro

#29 Abstract + Material Method Results Discussion Conclusion Ref

Data mining (DM), also called Knowledge-Discovery in Databases (KDD) or Knowledge-Discovery and Data Mining, is the process of automatically searching large volumes of data for patterns using tools such as classification, association rule mining, and/or clustering. The term has been defined as the nontrivial extraction of implicit, previously unknown, and potentially useful information from data [1], being considered as the science of extracting useful information from large data sets or databases [2].

Data mining techniques are use in search of consistent patterns and/or systematic relationships between variables in business [3], evaluation of web-based educational programs [4], computer science [5], chemistry [6], engineering [7], medicine [8], and in all domains where a large amount of date must be analyzed.

A new method of quantitative structure-activity/property relationships called MDF SAR/SPR (molecular descriptors family on the structure-activity/property relationships) has been introduced by Jäntschi in 2004 [9] and reviewed in 2005 [10]. Since then, samples of compounds with different properties or activities have been investigated and analyzed. Some results on different properties (retention chromatography index [9], relative response factor [11], molar refraction [12], octanol/water partition coefficient [13-15]) or activities (insecticidal activity [16], herbicidal activity [17], antioxidant efficacy [18], inhibition activity [19-21], toxicity [22,23], antituberculotic activity [24], and antimalarial activity [25]) have been reported. In addition, the overall results from the use of molecular descriptors family on structure property/activity relationships has also been published [26].

The best performing models in terms of correlation coefficients and cross-validation scores were collected into a database. On this amount of information, data mining techniques have been applied in order to identify consistent patterns and/or relationships between variables of MDF SAR/SPR models.

Material

#29 Abstract Intro + Method Results Discussion Conclusion Ref

A number of fifty-five sets of compounds were included into analysis. The set abbreviation, activity or property of interest and class of compounds are presented in Table 1.

Table 1. Characteristics of the sets included into analysis

No	Abbreviation	Activity /Property	Compounds

1	DevMTOp00	LC50/EC50 - fertilization of sea urchin	ordnance
2	DevMTOp01	LC50/EC50 - embryological development of sea urchin	
3	DevMTOp02	LC50/EC50 - germination of sea urchin	
4	DevMTOp03	LC50/EC50 - zoospore germination of green macroalgae	
5	DevMTOp04	LC50/EC50 - germling length of green macroalgae	
6	DevMTOp05	LC50/EC50 - germling cell number of green macroalgae	
7	DevMTOp06	LC50/EC50 - survival and reproductive success of polychaete	
8	DevMTOp07	LC50/EC50 - redfish larvae survival	
9	DevMTOp08	LC50/EC50 - juveniles survival of opossum shrimp	
10	DevMTOp09	NOEC - fertilization of sea urchin	
11	DevMTOp10	NOEC - embryological development of sea urchin	
12	DevMTOp11	NOEC - germination of sea urchin	
13	DevMTOp12	NOEC - germling length and cell number of green macroalgae	
14	DevMTOp14	NOEC - survival and reproductive success of green macroalgae	
15	DevMTOp15	NOEC - survival and reproductive success of polychaete	
16	DevMTOp16	NOEC - redfish larvae survival	
17	DevMTOp17	NOEC - juveniles survival of opossum shrimp	
18	DevMTOp18	LOEC - fertilization of sea urchin	
19	DevMTOp19	LOEC - embryological development of sea urchin	
20	DevMTOp20	LOEC - germination of sea urchin	
21	DevMTOp21	LOEC - germling length and cell number of green macroalgae	
22	DevMTOp22	LOEC - survival and reproductive success of green macroalgae	
23	DevMTOp23	LOEC - survival and reproductive success of polychaete	
24	DevMTOp24	LOEC - redfish larvae survival	
25	DevMTOp25	LOEC - juveniles survival of opossum shrimp	
26	DHFR	inhibition activity	2,4-Diamino-5-(substituted-benzyl)pyrimides
27	Dipeptides		dipeptides
28	RRC433_lbr	toxicity	para substituted phenols
29	RRC433_pka	relative toxicity	
30	Ta395	cytotoxicity	quinolines
31	Tox395	mutagenicity	
32	19654	antiallergic activity	substituted N 4-methoxyphenyl benzamides

33	22583	anti-HIV-1 potencies	HEPTA and TIBO derivatives
34	26449	antituberculotic activity	polyhydroxyxanthones
35	3300	growth inhibition activity	taxoids
36	41521	insecticidal activity	neonicotinoids
37	52344	antioxidant efficacy	3-indolyl derivates
38	52730	toxicity	alkyl metal compounds
39	23110		benzene derivates
40	23158		mono-substituted nitrobenzenes
41	23167		polychlorinated organic compounds
42	40846_1	inhibition activity on carbonic anhydrase I	substituted 1,3,4-thiadiazole- and 1,3,4-thiadiazoline-disulfonamides
43	40846_2	inhibition activity on carbonic anhydrase II	
44	40846_4	inhibition activity on carbonic anhydrase IV	
45	Triazines	herbicidal activity	substituted triazines
46	23159e	octanol/water partition coefficients	polychlorinated biphenyls
47	33504	boiling point	alkanes
48	36638	water activated carbon adsorption	organic compounds
49	IChr_10	retention chromatography index	organophosphorus herbicides
50	MR_10	molar refraction	cyclic organophosphorus
51	PCB_rrf	relative response factor	polychlorinated biphenyls
52	PCB_lkow	octanol/water partition coefficient	
53	PCB_rrt	relative retention time	
54	RRC433_lkow	octanol/water partition coefficient	para substituted phenols
55	31572		volatile organic compound

LC50 = lethal concentration to 50% of the test organisms
EC50 = effective concentration to 50% of the test organisms
NOEC = no observed effect concentration
LOEC = lowest observed effect concentration

Univariate and multivariate models were obtained by applying the MDF SAR/SPR methodology on the samples of compounds; the models were stored into a database. The molecular descriptors are the variables used by the models. The characters used on molecular descriptors name are presented in Table 2. The significance of each character was previous posted [23].

Table 2. Characters in molecular descriptors name

Position	**Characters**
First	I-i-A-a-L-l

Second	m-M-n-N-S-P-s-A-a-B-b-G-g-F-f-H-h-I-i
Third	m-M-D-P
Fourth	R-r-M-m-D-d
Fifth	D-d-O-o-P-p-Q-q-J-j-K-k-L-l-V-E-W-w-F-f-S-s-T-t
Sixth	C-H-M-E-G-Q
Seventh	g-t

Method

#29 Abstract Intro Material + Results Discussion Conclusion Ref

The MDF SAR/SPR database was interrogated and the interest information was obtained by using a series of PHP programs. The SPSS software was used for data summarizing and analyzing. The 95% confidence intervals were computed by using dedicated software based on binomial distribution hypothesis [27].

Two steps cluster analysis and hierarchical cluster analysis were used as methods in searching the patterns where was appropriate. The two-step cluster analysis was used on searching patterns overall models. This technique was choused because has specific feature: automatic selection of the best number of clusters, and ability to create cluster models simultaneously based on categorical and continuing variables. The hierarchical cluster method has been used for identification of similarities on the best performing MDF SAR/SPR models and was been choused because it is an easy to implement well-documented method and provides as result dendrograms, tree-like structures that illustrate the relationships between the entries.

Results

#29 Abstract Intro Material Method + Discussion Conclusion Ref

Fifty-five sets were included into analysis, cumulating an amount of one-hundred and ninety-five models. One hundred fifty-six models were for activities estimation and prediction (95%CI [144 - 166]) and thirty-eight models for properties estimation and prediction (95%CI [28 - 50]).

Seventy-three models reported estimation and prediction of activity (95%CI [64 - 80]) and nineteen models (95%CI [12 - 27]) estimation and prediction ability of property. The number of MDF SAR models varied from two to eleven (for the set no. 40, Table 1) and for MDF SPR models varied from two to eight (for the set no. 48, table 1). The statistical characteristics of all models, and of the best performing models (in terms of closest squared correlation coefficient and cross-validation score to one) are presented in Table 3.

Table 3. Statistical characteristics of the MDF SAR/SPR models

		nv	Mean [95%CI]	Median	Min	Max	StDev
All models							
Activity	r^2	156	0.9023 [0.8783 - 0.9263]	0.9489	0.0122	1.0000	0.1514
	v		2 [2 - 2]	2	1	5	1.1003
	n_{sample}		28 [24 - 31]	23	5	69	21.468
Property	r^2	38	0.8698 [0.8077 - 0.9319]	0.9772	0.1208	1.0000	0.1889
	v		4 [2 - 6]	2	1	24	6.0663
	n_{sample}		77 [48 - 105]	24	10	209	86.220
Best performing models							
Activity	r^2	45	0.9807 [0.9714 - 0.9900]	0.9992	0.9037	1.0000	0.0310

	v		3 [2 - 3]	2	2	5	1.0288
	n_{sample}		19 [13 - 24]	8	5	69	17.945
Property	r^2	10	0.9572 [0.8993 - 1.0000]	0.9883	0.7368	1.0000	0.0808
	v		3 [2 - 4]	2.5	2	6	1.3703
	n_{sample}		80 [16 - 144]	27	10	209	90.120

r^2 = squared correlation coefficient; v = number of descriptors used in models; n_{sample} = sample size; nv = number of valid samples; 95% CI = 95% confidence interval; Min= minimum; Max = maximum, StDev = standard deviation

The MDF SAR/SPR models stored into database used two hundred and eighty-four molecular descriptors. Almost sixty-nine percent of them were used just by one model (one hundred and ninety-six descriptors, 95%CI [180 - 211]). The distribution of the descriptors used by MDF SAR/SPR models was:

- Two descriptors were used by six models (imDrkQt, and lPMDVQg)
- Four by five models (ASPrVQg, IiMMWHt, IMPrkQg, and iSMMWHg)
- Sixteen descriptors were used by four models (AHMMVQg, aHPMwQt, aIDmjQg, iAMrVQg, iBMmwHg, iHDdFHg, iHMMtHg, IiDrQHg, iIPmWHt, ImmRDCg, imMrFHt, inDmwHg, INPRJQg, inPRlQg, isMdTHg, iSMmEQt)
- Twenty one descriptors were used by three models (ABDmtQg, ASMmVQt, AsPmVQt, aSPRtQg, IADRSHg, IBPMWQt, iGPrfHt, iIMdLGg, iIMdTMg, iImrKHt, InMdTHg, isDRTCg, isDRtHg, ismRSEg, iSPRtQg, lfDdOQg, LHDmjQg, lIDrFEg, lIMdLGg, liMDWHg, LsDMpQg)
- Forty-five descriptors were used by two models (ABmrtQg, AHDmEQg, aHMmjQt, AiMrKQt, AIPmVQt, AiPmVQt, aIPMwQt, anDRJQt, aSMMjQg, iAPmEQg, ibDMFHt, IbMmjHg, IBMrkGg, IBMRQCg, IbPdPHg, iFmRFMt, iFPMECg, IHDRKEg, iHMMTQt, IIDDKGg, IiMMSGg, imDdSCg, ImDmEEt, IMDMtQt, ImDrFEt, iMMMjQg, IMmrKQg, imMrtCg, inMRkQt, InPdJQg, inPRjQt, isDDkGg, IsMRKQg, ISPdlMg, IsPdOQg, lFDMwEt, lfDMWHt, lFMMKQg, LHDROQg, LIDmjQg, lImrKHt, lmMrsGg, lNPmfQt, LSPmEQg, LsPrDQt).

One hundred and forty-seven descriptors have been used in the best performing models. The correspondences between using the descriptors in all models and in best performing models are presented in Table 4.

Table 4. Descriptors in all models versus best performing models

Descriptors		**Total**
all models	**best models**	
1	1	89
1 Total		89
2	1	24
	2	2
	3	1
2 Total		27
3	1	11
	3	1
3 Total		12
4	1	9

	2	4
4 Total		13
5	1	3
	2	1
5 Total		4
6	1	2
6 Total		2
Total		147

The partial squared correlation coefficient (the squared correlation coefficient between each descriptor from the model and property or activity of interest) varied for the all models from 0.0001 to 0.9995 with an average of 0.3645. For the best performing models, the values of the partial squared correlation coefficients varied from 0.0001 to 0.9794 with an average of 0.2959. The average values of partial squared correlation coefficients for all models and for best performing model according with the activity or property of interest are summarized in Table 5. More, the descriptors that obtained greater value of partial squared correlation coefficients are not found in the best performing model.

Table 5. The average contribution of the descriptors to the model

Set abb.	**Avgr2-best**	**Avgr2-all**		**Set abb.**	**Avgr2-best**	**Avgr2-all**
MDF SARs				MDF SPRs		
DevMTOp00	0.8673	0.9113		23159	0.0089	0.1685
DevMTOp01	0.6632	0.7753		31572	0.2274	0.2581
DevMTOp02	0.4144	0.5866		33504	0.5297	0.6416
DevMTOp03	0.0398	0.3232		36638	0.2880	0.3051
DevMTOp04	0.2221	0.4454		IChr10	0.5998	0.4005
DevMTOp05	0.1355	0.3823		MR10	0.8971	0.9075
DevMTOp06	0.3040	0.5251		PCB_lkow	0.2268	0.3327
DevMTOp07	0.4579	0.6160		PCB_rrf	0.2712	0.2843
DevMTOp08	0.3384	0.5284		PCB_rrt	0.4687	0.7021
DevMTOp09	0.4035	0.5883		RRC433_lkow	0.2308	0.3011
DevMTOp10	0.4169	0.5941		*Min*	0.0089	0.1685
DevMTOp11	0.1692	0.4368		*Max*	0.8971	0.9075
DevMTOp12	0.0214	0.3060		*Average*	0.3748	0.4302
DevMTOp14	0.1092	0.3786				
DevMTOp15	0.1100	0.3905				
DevMTOp16	0.2451	0.4669				

DevMTOp17	0.1447	0.3694	
DevMTOp18	0.5083	0.6717	
DevMTOp19	0.2888	0.5032	
DevMTOp20	0.1391	0.3846	
DevMTOp21	0.0721	0.3492	
DevMTOp22	0.1946	0.4475	
DevMTOp23	0.1430	0.4033	
DevMTOp24	0.4997	0.6464	
DevMTOp25	0.0441	0.3559	
DHFR	0.1482	0.1680	
Dipeptides	0.5145	0.4603	
RRC433_lbr	0.1612	0.2329	
RRC433_pka	0.2623	0.2144	
Ta395	0.1027	0.1002	
Tox395	0.2053	0.2712	
19654	0.1360	0.3286	
22583	0.2288	0.1908	
26449	0.3874	0.5332	
3300	0.2408	0.2761	
41521	0.2407	0.4365	
52344	0.5083	0.4243	
52730	0.5806	0.7092	
23110	0.1298	0.2106	
23158	0.3011	0.2719	
23167	0.3546	0.3636	
40846_1	0.3264	0.4271	
40846_2	0.1319	0.2170	
40846_4	0.2529	0.2621	
Triazines	0.4323	0.4613	
Min	0.0214	0.1002	
Max	0.8673	0.9113	
Average	0.2800	0.4210	

Avgr2-best = the average of the partial squared correlation coefficient on best performing model; Avgr2-all = the average of the partial squared correlation coefficient on all models

Summarizing the characters that were included into the descriptors name it can be observed that, with a single exception, all characters for first, third, fourth, fifth, sixth and seven descriptor name letters appear in the descriptors names if all MDF SAR/SPR models. The same observation is valid for analysis of the best performing ones. There were identified that three characters out of nineteen from the second descriptor letter (the letters a, g and h, see Table 2) did not appear in any model. In order to applied cluster analysis techniques the frequency of the characters into the models according with the set name were transformed as qualitative variables (yes/no). The summaries of the results obtained by performing the two steps cluster analysis on all models as well as on the best performing models are presented in Table 6 (DescL = the letter in the descriptor name, Ch = character, Best model = the model that obtained the greatest squared correlation coefficient and cross-validation leave-one-out score). There were included into the Table 6 the absolute frequency of appearance of the character into the name of descriptors and the attribute importance into the cluster (‡ = significant importance in cluster at a significance level of 5%).

Table 6. Two steps cluster analysis: results

DescL	**Ch**	**All models**			**Best model**
		Cluster 1(41)	**Cluster 2(14)**	**Total**	
1st letter	I	25	13	38	31
	i	30	14‡	44	38
	A	7	4	11	7
	a	10	3	13	5
	L	13	4	17	8
	l	28	10	38	31
2nd letter	m	10	7	17	9
	M	3	4	7	7
	n	12	7	19	13
	N	7	1	8	5
	S	11	8	19	12
	P	5	1	6	5
	s	19	7	26	18
	A	14	5	19	13
	B	6	7‡	13	9
	b	2	6‡	8	6
	G	7	2	9	8
	F	3	7‡	10	4
	f	2	1	3	2
	H	14	9	23	16

	I	17	8	25	11
	i	3	7‡	10	4
3rd letter	m	13	8	21	10
	M	29	14‡	43	36
	D	31	13	44	34
	P	31	11	42	34
4th letter	R	22	10	32	23
	r	26	13	39	32
	M	11‡	14‡	25	20
	m	28	13	41	25
	D	12	8	20	15
	d	10	10‡	20	14
5th letter	D	7	2	9	4
	d	4	2	6	3
	O	6	0	6	5
	o	3	2	5	2
	P	3	3	6	4
	p	5	2	7	4
	Q	1	3‡	4	3
	q	6	1	7	6
	J	7	6	13	6
	j	9	5	14	6
	K	3	7‡	10	5
	k	10	8‡	18	13
	L	7	2	9	6
	l	4	2	6	5
	V	8	6	14	10
	E	5‡	9‡	14	9
	W	1	4‡	5	5
	w	9	7	16	8
	F	4‡	10‡	14	7
	f	9	2	11	5

	S	7	5	12	8
	s	6	6	12	5
	T	6	6	12	9
	t	10	7	17	8
6th letter	C	10	7	17	6
	H	9‡	14‡	23	20
	M	17	7	24	16
	E	10	5	15	12
	G	12	8	20	11
	Q	40	14	54	44
7th letter	g	40	14	54	41
	t	31	13	44	51

DescL = the letter in the descriptor name
Ch = character
Best model = the model that obtained the greatest squared correlation coefficient and cross-validation leave-one-out score
‡ = significant importance in cluster at a significance level of 5%

The hierarchical cluster technique was applied in order to analyze the best performing models. The Icile plot is presented in Figure 1 and the associated dendrogram in Figure 2.

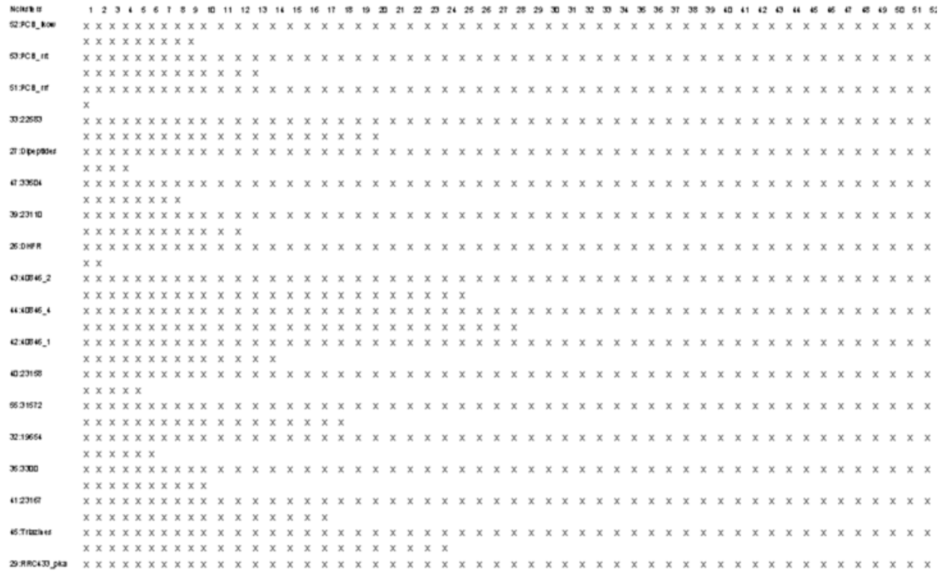

Figure 1. Best performing MDF SAR/SPR models analysis: icile plot

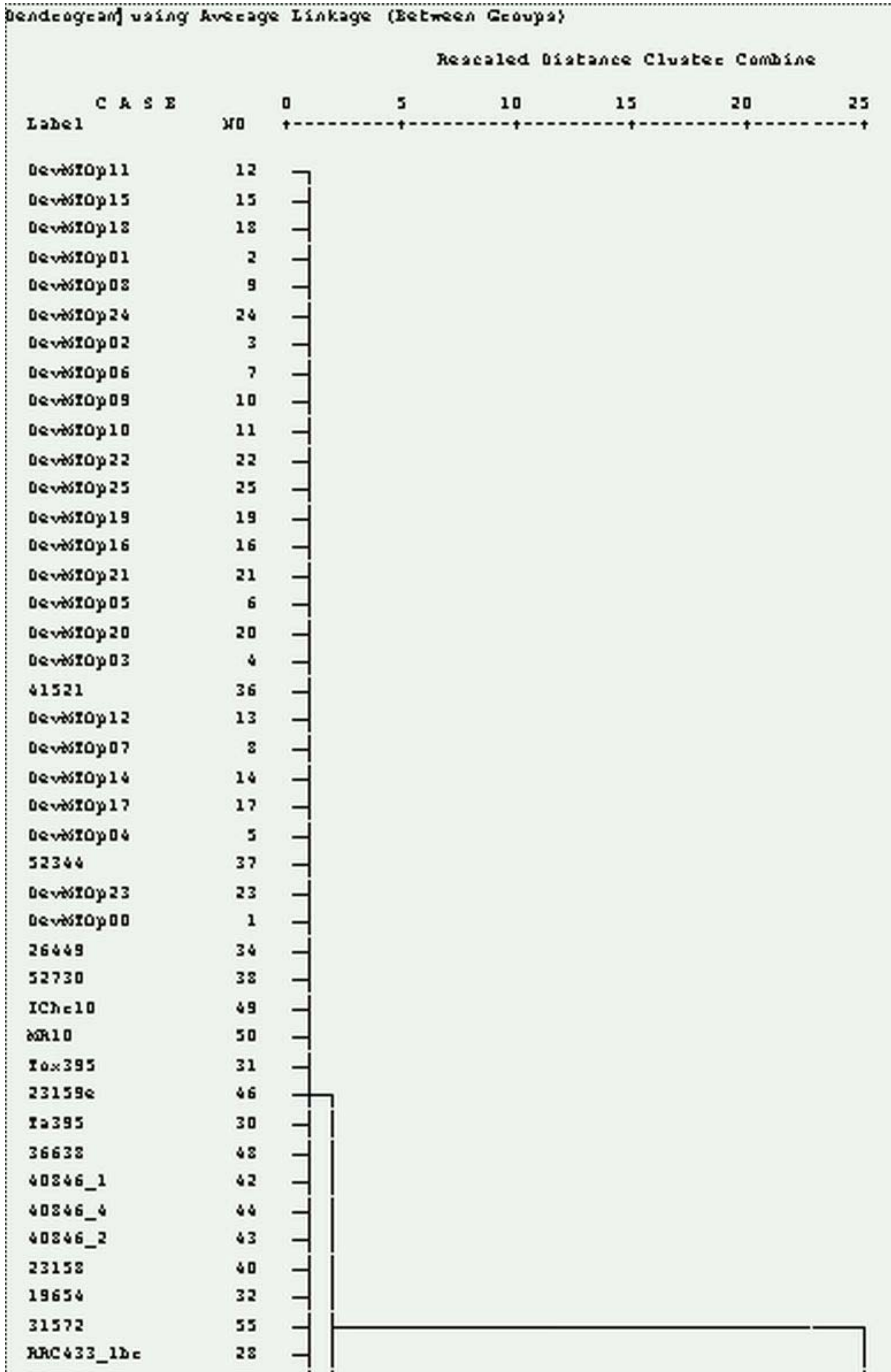
Dendrogram using Average Linkage (Between Groups)
Rescaled Distance Cluster Combine
C A S E
Label
NO
0
5
10
15
20
25
DevWTOp11 12
DevWTOp15 15
DevWTOp18 18
DevWTOp01 2
DevWTOp08 9
DevWTOp24 24
DevWTOp02 3
DevWTOp06 7
DevWTOp09 10
DevWTOp10 11
DevWTOp22 22
DevWTOp25 25
DevWTOp19 19
DevWTOp16 16
DevWTOp21 21
DevWTOp05 6
DevWTOp20 20
DevWTOp03 4
41521 36
DevWTOp12 13
DevWTOp07 8
DevWTOp14 14
DevWTOp17 17
DevWTOp04 5
52344 37
DevWTOp23 23
DevWTOp00 1
26449 34
52730 38
IChc10 49
MR10 50
Tox395 31
23159e 46
Ta395 30
36638 48
40846_1 42
40846_4 44
40846_2 43
23158 40
19654 32
31572 55
RRC433_1bc 28

Figure 2. Best performing MDF SAR/SPR models analysis: dendrogram

Discussion

#29 Abstract Intro Material Method Results + Conclusion Ref

Searching the information regarding the MDF SAR/SPR models for patterns revealed important information for activity/property characterization of compounds classes by applying the molecular descriptors family methodology.

As it can be observed from Table 3, the average of the correlation coefficient obtained by MDF SARs is greater comparing with the value obtained by the MDF SPRs, while the number of variables is less for MDF SARs than for MDF SPRs when all models are considered. When the best performing models are analyzed it can be observed that the squared correlation coefficient average obtained by the MDF SAR models is very closed to the squared correlation coefficient average obtained by MDF SPR models, and the average of the descriptors is the same.

Just forty-five percent of the molecular descriptors that were used in one model on completely sample of models could be found in the best performing models (see Table 4). Sixty percent of the molecular descriptors used by two models on whole samples were found again on the best performing models (see Table 4). Fifty-seven percent of the molecular descriptors used by three models on whole samples were found again on the best performing models; almost eighty-one percent of the molecular descriptors used by four models on whole samples were found again on the best performing models. All molecular descriptors used by five, and respectively six models on whole samples were found as being used on the best performing models too (see Table 4). These observations sustained the stability and consistency of the MDF SAR/SPR method in identification of the molecular descriptors that are able to identify the strongest relationships between compounds structure and associated activity or property.

Analyzing the data presented in Table 4 it can be observed that the average, minimum and maximum values of average contribution of descriptors are smaller values for the best performing models than the values obtained on all models. This observation leads to the conclusion that the best performing models are obtained by combination of descriptors, and the molecular descriptors that had a value of the partial correlation coefficient closest to one are not always found in the best performing model.

Two clusters were obtained by applying the two-step cluster analysis technique on the all models, showing that there exist some similarities between MDF models. One cluster used forty-one sets of compounds while the second cluster used fourteen compounds. Four characters had significant importance into the first cluster obtained on all models (see table 6):

- Character M (the overlapping descriptors interaction on the maximal fragments) from fourth position on descriptors name
- Characters E (interaction descriptor of the second atom property divided to the distance between the atoms) and F (interaction descriptor of the square first atom property divided to the square distance between atoms) as fifth position on descriptors name
- Character H (number of directly bonded hydrogen's as atomic property) from sixth position on descriptors name

In the second cluster, the one that comprise fourteen sets of compounds, fourteen characters revealed to have significant importance in clustering:

- Character i (the inverse linearization procedure applied in global molecular descriptor generation) from the first position on descriptors name
- Characters B (as average mean by atom), b (average mean by bond), F (geometric mean by atom), i (harmonic mean by bond) from the second position on descriptors name (the cumulative method of fragmentation properties)
- Character M (the maximal fragments criteria) from the third position on descriptors name
- Characters M (the overlapping descriptors interaction on the maximal fragments) and d (the overlapping descriptors interaction on threat descriptors as Cartesian vectors) from the fourth position on descriptors name
- Characters Q (the squared product between first and second atoms properties), K (the product between the first and second atoms properties and the distance between them), k (the inverse of K), E (interaction descriptor of the second atom property divided to the distance between the atoms), and W (the square of the first atom property divided to the distance between two atoms) from the fifth position on descriptors name
- Character H (number of directly bonded hydrogen's as atomic property) from the sixth position on descriptors name:

On the sample of best performing MDF SAR/SPR models, the two-step cluster analysis was able to identify two clusters. This could be explained by the absence of similarities of descriptors characters used by the best performing models. The most frequently met characters on the descriptors name on the best performing models were:

- i character for the first position on descriptors name (the inverse linearization procedure applied in global molecular descriptor generation)
- s character for the second position on descriptors name (the product between the first and second atoms properties divided to the distance rice to power three)
- M character for the third position on descriptors name (the maximal fragments criteria)
- r character for the fourth position on descriptors name (the overlapping descriptors interaction obtained by treating descriptors as scalars and computing resultant relative to conventional origin)
- k character for the fifth position on descriptors name (the inverse of the product between the first and second atoms properties and the distance between them)
- Q character for the sixth position on descriptors name (semi-empirical Extended Hückel model, Single Point approach as atomic property)
- t character for the first position on descriptors name (molecular topology)

Taking into account the above information, it can be concluding that there could not be identify similarities or patterns on the MDF SAR/SPR models even if the results of the analysis of all models say something else. Note that in the analysis of the all MDF SAR/SPR models were included for each set of compounds the univariate models that in most of the cases obtained weak performances in terms of estimation and prediction abilities.

The quantitative variables similarities of the best performing models were analyzed with hierarchical cluster technique. Looking at the icile plot (Figure 1) it can be analyzed what happen at each clusterization step. At the start step (the one that is not represented on icicle plot, Figure 1), each set of compounds was a cluster unto itself (the number of clusters at the start point being equal with fifty-five). Starting with the first step, the sets were ordered in the icicle plot according with their combination into clusters. The 15:DevMTOp15 set is linked first with 12:DevMTOp11 set, being follow by the 24:DevMTOp24 set, and so on until all the clusters are formed. From the dendrogram (see Figure 2) it can be observed that at a small distances three clusters are formed: one that comprised forty-seven sets, and other two that comprised five and respectively three sets. The differences between the obtained three clusters are at the level of sample size, and number of descriptors used by model. On the cluster that comprised forty-seven sets the sample sizes varied from five to forty, and the number of molecular descriptors from two to three. On the cluster that comprised five sets the sample seizes varied from fifty-seven to seventy-three and the number of descriptors from two to five, while on the cluster that comprised three sets the number of compounds were of two hundred and nine and the number of variables from two to six. At a short distance, two clusters are linked together (the one that comprised forty-seven and the other that comprised five sets). All the clusters are linked together at the maximum distance as possible.

The research reached its goal of searching the patterns on MDF SAR/SPR models. The results shown that on the studied sets of compounds the MDF SAR/SPR method identified models that are unique for each set do to the complex information obtained from compounds structure. Based on the obtained results the MDF SAR/SPR method will be updated by analyzing of the usefulness of the three characters from the second position descriptor name that were not identified in any model. The development of the MDF SAR/SPR database by analyzing and including of more compounds sets will be done in the future. Data mining techniques applied on larger sets of compounds could revealing important information for characterization of activities or properties of compound based on information obtained from the structure.

Conclusion

#29 Abstract Intro Material Method Results Discussion + Ref

The data mining techniques applied on MDF SAR/SPR models revealed that is not possible any classification of characters used on descriptors name and thus on their construction. This result sustains the ability of MDF SAR/SPR method on identification of those structure characteristics of compounds that are linked with the activity or property of interest.

The hierarchical cluster analysis is a useful technique in identification of similarities of MDF SAR/SPR models regarding the quantitative variables, in our case the squared correlation coefficient, the number of descriptors used by models and the sample sizes.

Data mining techniques applied on larger sets of compounds analyzed with MDF SAR/SPR method could reveal important information for characterization of activities or properties of compound based on information obtained from the structure.

Ref

#29 Abstract Intro Material Method Results Discussion Conclusion +

1. W. Frawley and G. Piatetsky-Shapiro and C. Matheus (Fall 1992). "Knowledge Discovery in Databases: An Overview". AI Magazine 1992, pp. 213-228.
2. D. Hand, H. Mannila, P. Smyth. "Principles of Data Mining". MIT Press, Cambridge, MA, 2001.
3. Y.-L. Chen, J.-M. Chen, C.-W. Tung. "A data mining approach for retail knowledge discovery with consideration of the effect of shelf-space adjacency on sales". Decision Support Systems 2007, 42(3), pp. 1503-1520.
4. C. Romero, S. Ventura. "Educational data mining: A survey from 1995 to 2005". Expert Systems with Applications 2007, 33(1), pp. 135-146.
5. A.J.T. Lee, R.-W. Hong, W.-M. Ko, W.-K. Tsao, H.-H. Lin. "Mining spatial association rules in image databases". Information Sciences 2007, 177(7), pp. 1593-1608.
6. U. Maran, S. Sild, I. Kahn, K. Takkis. "Mining of the chemical information in GRID environment". Future Generation Computer Systems 2007, 23(1), pp. 76-83.
7. Q. Yang, J. Yin, C. Ling, R. Pan. "Extracting actionable knowledge from decision trees". IEEE Transactions on Knowledge and Data Engineering 2007, 19(1), pp. 43-55.
8. T. Imamura, S. Matsumoto, Y. Kanagawa, B. Tajima, S. Matsuya, M. Furue, H. Oyama. "A technique for identifying three diagnostic findings using association analysis". Medical and Biological Engineering and Computing 2007, 45(1), pp. 51-59.
9. L. Jäntschi. "MDF - A New QSPR/QSAR Molecular Descriptors Family". Leonardo Journal of Sciences 2004, Issue 4, pp. 68-85.
10. L. Jäntschi. "Molecular Descriptors Family on Structure Activity Relationships 1. Review of the Methodology". Leonardo Electronic Journal of Practices and Technologies 2005, Issue 6, pp. 76-98.
11. L. Jäntschi. "QSPR on Estimating of Polychlorinated Biphenyls Relative Response Factor using Molecular Descriptors Family". Leonardo Electronic Journal of Practices and Technologies 2004, 5, pp. 67-84
12. L. Jäntschi, S. Bolboacă. "Molecular Descriptors Family on Structure Activity Relationships 4. Molar Refraction of Cyclic Organophosphorus Compounds". Leonardo Electronic Journal of Practices and Technologies 2005, 7, pp. 55-102.
13. L. Jäntschi, S. Bolboacă. "Molecular Descriptors Family on Structure Activity Relationships 6. Octanol-Water Partition Coefficient of Polychlorinated Biphenyls". Leonardo Electronic Journal of Practices and Technologies 2006, 8, pp. 71-86.
14. L. Jäntschi. "Delphi Client - Server Implementation of Multiple Linear Regression Findings: a QSAR/QSPR Application". Applied Medical Informatics 2004, 15, pp. 48-55
15. L. Jäntschi, S.D. Bolboacă. "Modeling the Octanol-Water Partition Coefficient of Substituted Phenols by the Use of Structure Information". International Journal of Quantum Chemistry. In Press, Published Online: 3 Jan 2007
16. S. Bolboacă, L. Jäntschi. "Molecular Descriptors Family on Structure Activity Relationships 2. Insecticidal Activity of Neonicotinoid Compounds". Leonardo Journal of Sciences 2005, 6, pp. 78-85.
17. S. Bolboacă, L. Jäntschi. "Molecular Descriptors Family on Structure-Activity Relationships: Modeling Herbicidal Activity of Substituted Triazines Class". Bulletin of University of Agricultural Sciences and Veterinary Medicine - Agriculture 2006, 62, pp. 35-40.
18. S. Bolboacă, C. Filip, S. Tigan, L. Jäntschi, "Antioxidant Efficacy of 3-Indolyl Derivates by Complex Information Integration". Clujul Medical 2006, Issue LXXIX(2), p. 204-209.
19. L. Jäntschi, M.L. Unguresan, S.D. Bolboacă. "Integration of Complex Structural Information in Modeling of Inhibition Activity on Carbonic Anhydrase II of Substituted Disulfonamides". Applied Medical Informatics 2005, 17(3, 4), pp. 12-21.
20. L. Jäntschi, S. Bolboacă. "Modelling the Inhibitory Activity on Carbonic Anhydrase IV of Substituted Thiadiazole- and Thiadiazoline- Disulfonamides: Integration of Structure Information". Electronic Journal of Biomedicine 2006, 2, p. 22-33.
21. S. Bolboacă, S. Tigan, L. Jäntschi. "Molecular Descriptors Family on Structure-Activity Relationships on anti-HIV-1 Potencies of HEPTA and TIBO Derivatives". Proceedings of the European Federation for Medical Informatics Special Topic Conference, April 6-8, 2006, pp. 222-226.
22. S.D. Bolboacă, L. Jäntschi. "Modeling of Structure-Toxicity Relationship of Alkyl Metal Compounds by Integration of Complex Structural Information". Terapeutics, Pharmacology and Clinical Toxicology 2006, X(1), pp. 110-114.
23. L. Jäntschi, S. Bolboacă. "Molecular Descriptors Family on QSAR Modeling of Quinoline-based Compounds Biological Activities". The 10th Electronic

Computational Chemistry Conference. April 2005, http://eccc.monmouth.edu
24. S. Bolboacă, L. Jäntschi. "Molecular Descriptors Family on Structure Activity Relationships 3. Antituberculotic Activity of some Polyhydroxyxanthones". Leonardo Journal of Sciences 2005, 7, pp. 58-64.
25. L. Jäntschi, S. Bolboacă. "Molecular Descriptors Family on Structure Activity Relationships 5. Antimalarial Activity of 2,4-Diamino-6-Quinazoline Sulfonamide Derivates". Leonardo Journal of Sciences 2006, 8, pp. 77-88.
26. L. Jäntschi, S. Bolboacă. "Results from the Use of Molecular Descriptors Family on Structure Property/Activity Relationships". International Journal of Molecular Sciences 2007, 8, pp. 189-203.
27. Binomial Distribution, © L 2007. Available from: URL: http://l.academicdirect.org/Statistics/binomial_distribution/

Used software: PHP [php.het], FreePascal [freepascal.org], MySQL [mysql.com], SPSS [spss.com].

Online resources: ECCC10#4 [presentation #4 at ECCC10]; L.AcademicDirect [Library from AcademicDirect].

Acknowledgment: MEC/UEFISCSU Romania, Grant ET.46/2006.

Contact: sorana@j.academicdirect.ro, lori@j.academicdirect.org.