Ministerul Educației şi Cercetării

Universitatea de Medicină şi Farmacie "Iuliu Haţieganu" Cluj-Napoca

Facultatea de Medicină

Catedra de Informatică Medicală şi Biostatistică

Planul Național de Cercetare, Dezvoltare şi Inovare - PN II

Programul:	**IDEI**
Tipul proiectului:	**Proiecte de cercetare exploratorie**
Cod proiect:	**ID_458**
Denumire proiect:	**Biochimie versus Biomatematică în Medicina Moleculară**
Etapa:	**Unică/2010**

- LUCRARE ÎN EXTENSO -

- 2010 -

Cuprins

Obiective planificate şi activităţi prevăzute

Obiective planificate

4.1. Analiza modelelor prin tehnici statistice multivariate

4.2. Realizare librărie virtuală

4.3. Valorificarea rezultatelor

Activităţi prevăzute

Activităţi asociate obiectivului 4.1.

4.1.1. Aplicare metode clusterizare pe cele trei clase de compuşi chimici biologic activi investigate

4.1.2. Analiza factorilor pe baza descriptorilor modelului matematic

4.1.3. Monitorizare - Bucureşti, CNCSIS-UEFISCSU

Activităţi asociate obiectivului 4.2.

4.2.1. Proiectare implementare aplicaţie, integrare modele în baza de date, implementare algoritmi de interogare

4.2.2. Testare mediu virtual

Activităţi asociate obiectivului 4.3.

4.3.1. Documentare, identificare şi selectare compuşi chimici din clasele studiate

4.3.2. Predicţie activitate pe baza structurii prin folosirea modelelor structură-activitate obţinute

4.3.3. Activităţi suport

Activităţile au fost realizate şi obiectivele planificate au fost atinse. Rezultatele estimate au fost obţinute. Scopul cercetării a fost obţinut.

Obiective/Activități/Rezultate

Obiectivul 4.1. Analiza modelelor prin tehnici statistice multivariate

4.1.1. Aplicare metode clusterizare pe clasele de compuşi chimici biologic activi investigate

Analizele de clusterizare au fost aplicate pe activitate/proprietatea măsurată experimental cât şi pe valorile descriptorilor MDFV pentru fiecare clasă de compuşi în parte.

An aliza de clusterizare s-a realizat cu ajutorul programului SPSS 16.0 la un prag de semnificație de 5%.

- Scop: identificarea grupelor de compuşi care sunt similare unele cu celelalte dar în acelaşi timp diferiți față de compuşii din celelalte grupuri.
- Metode: analiza de clusterizare & analiza de discriminare permit clasificarea compuşilor în grupuri. Aplicarea celei de a doua metode necesită cunoaşterea prealabilă a apartenenței la o clasă. În analiza de clusterizare nu se cunoaşte cine sau ce anume cuprinde fiecare grup; cel mai frecvent nu se cunoaşte nici măcar numărul de grupuri.
- Aplicabilitate: nu există asumpții cu privire la distribuția datelor.

Metode (analiza de clusterizare):

1. ***Analiza ierarhică de clusterizare*** (***hiercarchical cluster analysis***): set mic de date.

Există grupuri identificabile în setul de molecule investigate cu caracteristici similare (ex. activitatea/proprietatea măsurată, valori ale descriptorilor moleculari, etc.)?

Tipul variabilelor: calitative, binare sau cantitative.

Ordinea datelor: dacă există distanțe egale (identice) sau similare în datele de input sau apar în timpul alăturării clusterii rezultați pot depinde de ordinea datelor în fişierul analizat. În acest caz se identifică mai multe soluții cu datele sortate după diferite criterii pentru a verifica stabilitatea soluției obținute.

Asumpții: măsurile de similaritate şi/sau distanță utilizate trebuie să fie în concodranță cu datele analizate:

- date de tip interval (alternative posibil de aplicat):
  - distanța Euclidiană (opțiunea implicită) [1]
  - pătratul distanței Euclidiene
  - cosin: valoarea cosinusului unghiului dintre doi vectori ai valorilor
  - coeficientul de corelație Pearson [2]: corelație dintre doi vectori ai valorilor

[1] Black PE, "Euclidean distance", in Dictionary of Algorithms and Data Structures [online], Black PE, ed., U.S. National Institute of Standards and Technology. 17 December 2004. (accessed July 2010) Available from: http://www.nist.gov/dads/HTML/euclidndstnc.html

- Chebychev [3]: diferența absolută maximă între valorile itemilor
- Blocuri: suma diferențelor absolute ale valorilor unui punct, cunoscută de asemenea ca şi distanța Manhattan
- Minkowski [4]: rădăcina de ordin p a diferențelor absolute la puterea p a între valorile punctelor

- date discrete cantitative:
  - măsuri de tip hi-pătrat [5]: acest indicator eeste bazat pe statistica hi-pătrat de egalitate a două seturi de frecvențe [6, 7]; este opțiunea implicită pentru datele de tip cantitativ discret
  - fi-pătra: această mărime este egală cu mărimea hi-pătrat normalizată de rădăcina pătratică a frecvenței combinate.
- date binare:
  - distanța Euclidiană: calculată pe tabela de contingență de 2×2 ca SQRT(b+c) unde b şi c reprezintă celulele de pe diagonală corespunzătoare prezenței în cazul unui item şi absente pentru celelalte itemuri
  - pătratul distanței Euclidiene: calculat ca numărul de cazuri discordante; ia valori minime de 0 fără a avea o limită superioară
  - diferența mărimii: un indicator al asimetriei; ia valori în intervalul [0, 1]
  - diferența tiparului: măsură a disimilarității ce ia valori în intervalul [0, 1], calculată ca bc/(n*2), unde n = numărul total de observații
  - varianța: calculată ca (b+c)/4n, ia valori în intervalul [0, 1]
  - dispersia: indice de similaritate ce ia valori în intervalul [-1, 1]

---

[2] Pearson K. Mathematical Contributions to the Theory of Evolution. III. Regression, Heredity, and Panmixia, Philosophical Transactions of the Royal Society of London, Series A 1896;187:253-318.

[3] Cantrell CD. Modern Mathematical Methods for Physicists and Engineers. Cambridge University Press, 2000.

[4] Kruskal JB. Multidimensional scaling by optimizing goodness of fit to a non metric hypothesis. Psychometrika 1964;29(1):1-27.

[5] Bolboacă SD, Jäntschi L, Sestraş AF, Sestraş RE, Pamfil DC. Pearson-Fisher Chi-Square Statistic Revisited. Submitted. 2010.

[6] Pearson K. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. Philosophical Magazine 1900;50:157-175.

[7] Fisher RA. On the interpretation of $\chi 2$ from contingency tables, and the calculation of P. Journal of the Royal Statistical Society 1922;85(1):87-94.

- forma: mărime a distanței ce ia valori în intervalul [0, 1] şi care penalizează asismetria nepotrivirilor
- potrivirea simplă: raportul dintre potriviri şi numărul total de valori; pondere egală se aplică atât potrivirilor cât şi nepotrivirilor
- lambda: Goodman and Kruskal's lambda; corespunde reducerii proporționale a erorii utilizând un item pentru a obține predicția celorlalți itemi; ia valori în intervalul [0, 1]
- Anderberg D [8]: reducerea reală a erorii utilizând un item pentru a obține predicția celorlalți itemi – predicție în ambele direcții; ia valori între 0 şi 1
- Hamann [9]: acest indicator este reprezentat de diferența dintre potriviri şi nepotriviri raporta la numărul total de observații; ia valori în intervalul [-1, 1]
- Jaccard: absențele comune nu sunt luate în considerare; se atribuie aceeaşi pondere şi potrivirilor şi nepotrivirilor; este cunoscut şi sub denumirea rația de similaritate
- Kulczynski 1: este raportul dintre prezența asocierilor şi totalitatea nepotrivirilor; limita inferioară este 0 iar cea superiorară ia orice valoare. Este teoretic nedefinit în cazul în care nu există nici o nepotrivire (dar unele programe asignează o valoare de 9999.999 în cazul unei valori nedefinire sau a unei valori mai mare decât 9999.999).
- Kulczynski 2: indicator bazat pe probabilitatea condiționată ca o caracteriztică să fie prezentă pentru un item chiar dacă este prezentă şi la alți itemi
- Lance şi Williams (cunoscut de asemenea ca şi coeficientul non-metric Bray-Curtis) [10]: calculat ca (b+c)/(2a+b+c), unde a reprezintă în tabela de contingență celula corespunzătoare cazurilor preznete în ambii itemi; ia valori în intervalul [0, 1]
- Ochiai [11]: forma binară a măsurii de similaritate cosin; ia valori în intervalul [0, 1]
- Rogers şi Tanimoto [12]: indicator care dă valoare dublă nepotrivirilor
- Russel şi Rao [13]: indicator implicit pentru date binare; ponderi egale sunt date atât potrivirilor cât şi nepotrivirilor

[8] Anderberg MR. Cluster Analysis for Applications, New York: Academic Press, 1973.

[9] Harman HH. *Modern Factor Analysis*, 3rd ed. Chicago: University of Chicago Press, 1976.

[10] Bray JR, Curtis JT. An ordination of upland forest communities of southern Wisconsin. Ecological Monographs 1957;27:325-349.

[11] Ochiai A. Zoogeographic studies on the soleoid fishes found in Japan and its neighbouring regions. Bill Jpn Soc Sci Fish (Nihon Suisan Gakkaishi) 1957;22:526-530.

[12] Rogers DJ, Tanimoto TT. A Computer Program for Classifying Plants. Science 1960;132:1115-1118.

[13] Rao CR. The utilization of multiple measurements in problems of biological classification. Journal of the Royal Statistical Society, Series B 1948;10:159-193.

- Sokal şi Sneath 1: pondere dublă este dată potrivirilor
- Sokal şi Sneath 2: pondere dublă este dată nepotrivirilor şi absența asocierilor nu se ia în considerare
- Sokal şi Sneath 3: raportul dintre potriviri şi nepotriviri; limită inferioară de 0 şi superioară nedefinită.
- Sokal şi Sneath 4: bazat pe probabilitatea condiționată ca o caracteriztică într-un item să potrivească valorii din alt item. Media valorile separate ale fiecărui item acționând ca şi predictor pentru ceilalți itemi este utilizată pentru a calcula această valoare.
- Sokal şi Sneath 5: media geometrică pătratică a probabilităților condiționate a potrivirilor pozitive şi negative; ia valori în intervalul [0, 1]
- Yule's Y (coefficient of cologation) [14]: funcție a raportului încrucişat în tabela de contingență de 2×2 fiind independentă de totalurile marginale. Ia valori în intervalul [-1, 1]
- Yule's Q: caz special al indicatorului gamma Goodman şi Kruskal; ia valori în intervalul [-1, 1]

- Acest tip de analiză permite gruparea compuşilor investigați în grupuri omogene pe baza unor caracteristici comune.
- Selectarea criteriului de similaritate / distanță între cazuri. Similaritatea este o măsură a cât de similare sunt una față de cealaltă două valori. Distanța este o măsură a cât de departe sunt două valori una față de cealaltă. Pentru valorile care sunt asemănătoare, *distanțele au valori mici* şi *indicatorii de similaritate au valori mari*.
- Statistica:
  - Matricea de distanță / similaritate
  - Aparteneța la un cluster pentru o singură soluție sau pentru mai multe soluții.
  - Reprezentarea grafică: dendrograma sau a graficului de tip țurțure.

Metoda utilizată în clusterizare a fost ***metoda Ward*** pe variabile de tip interval, prin aplicarea ***pătratului distanței Euclidiene***. Metoda Ward utilizează o metodă de analiză a variațelor pentru a evalua distanțele dintre clusteri. În general metoda este cunoscută ca fiind eficientă; apartenența la cluster este evaluată prin calcularea sumei totale a pătratelor deviațiilor de la media clusterului respectiv. Criteriul de fuziune a clusterilor este producerea unei cât mai mici posibile creşteri a sumei pătratelor erorilor.

Se aplică când nu avem nici un fel de informații a priori cu privire la numărul de clusteri.

[14] Yule GU. On the association of attributes in statistics. Philos Trans R Soc A 1900;194:257-319.

2. ***K-means cluster*** [15]: Se aplică atunci când există există o ipoteză în ceea ce priveşte numărul de clusteri asociați variabilelor / cazurilor de interes. Frecvent analiza ierarhică de clusterizare şi clusterizarea cu k-medii se utilizează succesiv. Metoda Ward se utilizează pentru a identificat numărul posibil de clusteri şi modalitatea în care aceştia fuzionează (reprezentarea prin dendograma). Ulterior, se aplică metoda k-means cluster utilizând informația obținută din analiza anterioară în ceea ce priveşte numărul optim de clusteri.

- Tipuri de variabile: cantitative pe scală de tip interval sau rație. Pentru date binare se recomandă utilizarea procedurii ierarhice de clusterizare.
- Statistica:
    a. ***Soluția completă***: valorile centrale inițiale ale clusterilor, Anova
    b. ***Fiecare caz***: informații ale clusterilor şi distanța față de centrul clusterului.
- Calcularea distanțelor: distanța Euclidiană

3. ***Two-step cluster***: volum de eşantion mare (> 1000 cazuri) sau variabile cantitative continue şi calitative. Această tehnică nu a fost aplicată pe seturile de compuşi investigate deoarece nu a fost îndeplinit criteriul

4.1.1.1. Derivați carbochinone - activitate antitumorală

Analiza ierarhică de clusteriza s-a realizat pe datele experimentale prezentate în Tabelul 1.

Rezultatele obținute în investigarea proprietății de interes în termeni de modalitate de aglomerare în clusteri sunt redate în Tabelul 2. Rezultatele din Tabelul 2 pune la dispoziție soluții pentru fiecare număr posibil de clusteri de la 1 la 37 (37 fiind de fapt volumul eşantionului investigat). Analiza coeficienților evidențiază următoarele: coeficientul de aglomerare în cazul unui singur cluster este egal cu 14.472; coeficientul de aglomerare în cazul a 2 clusteri este egal cu 4.865; coeficientul de aglomerare în cazul a 3 clusteri este egal cu 1.605; etc. (citirea se face ce la capătul inferior al coloanei spre cel puperior). Dendrograma asociată analziei este prezentată în Figura 1.

Sumarizarea rezultatelor în termeni de coeficienți de aglomerare este prezentată în Tabelul 3.

Un punct clar de demarcare în ceea ce priveşte diferența este la nivelul 0.9530 (diferență de

[15] MacQueen JB. Some Methods for classification and Analysis of Multivariate Observations. 1. Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. University of California Press. 1967:281-297.

ordin de mărime) → analiza poate să fie reluată pentru un număr fix de 3 clusteri. În urma anlizei s-a obținut apartenența fiecărui compus la un cluster după cum urmează:

- Cluster 1: compușii 1-8 (8 compuși)
- Cluster 2: compușii 9-22 (14 compuși)
- Cluster 3: compușii 23-37 (15 compuși)

Parametrii statistici descriptive asociați fiecărui cluster pentru proprietatea de interes sutn prezentați în Tabelul 4. Aplicăm testul one-way ANOVA pentru a determina dacă există diferențe semnificative statistic între grupuri (Tabelul 5).

Tabelul 1. Date experimentale: derivați de carbochinone

Mol	TEuIFFDL	GLCIicdI	TAkaFcDL	GLbIAcDR	Prop
cqd01	0.3221	0.9851	2.1948	49.8200	4.33
cqd02	0.1903	1.0000	2.2578	49.2500	4.47
cqd03	0.1930	0.9826	2.3021	52.8100	4.63
cqd04	0.1601	1.0000	1.2754	55.9100	4.77
cqd05	0.1675	0.9824	1.9046	49.7600	4.85
cqd06	0.1460	1.0000	1.3150	56.0100	4.92
cqd07	0.1696	0.9824	1.6696	40.7500	5.15
cqd08	0.0806	1.0000	2.3848	17.7280	5.16
cqd09	0.0812	0.9826	1.0246	56.8800	5.46
cqd10	0.0345	1.0000	1.1547	43.1100	5.57
cqd11	0.0503	1.0000	1.0720	33.6700	5.59
cqd12	0.0720	0.9826	1.0749	57.7400	5.6
cqd13	-0.0512	0.9671	2.0179	39.7800	5.63
cqd14	-0.0045	0.9824	0.8108	59.7600	5.66
cqd15	0.0086	0.9826	0.7947	59.0300	5.68
cqd16	0.1216	0.9826	1.0919	42.1800	5.68
cqd17	-0.1179	0.9877	1.6973	41.1500	5.68
cqd18	0.0911	1.0000	1.5281	34.0100	5.69
cqd19	-0.0405	0.9671	1.9086	41.4200	5.76
cqd20	-0.1422	0.9978	1.7685	42.1500	5.78
cqd21	0.0658	0.9826	0.8301	58.3100	5.82
cqd22	0.0345	0.9826	0.6881	58.7500	5.86
cqd23	-0.0244	0.9589	1.7888	42.2200	6.03
cqd24	-0.1048	0.9721	1.8220	39.1000	6.14
cqd25	-0.0704	0.9721	1.7677	36.5000	6.16
cqd26	-0.0795	0.9721	1.3575	41.7600	6.18
cqd27	-0.0613	0.9721	1.4279	37.0900	6.18
cqd28	-0.1709	0.9794	1.4822	42.1400	6.18
cqd29	-0.1614	0.9877	1.1223	42.1600	6.21
cqd30	-0.1384	0.9877	1.2224	41.4000	6.25
cqd31	-0.1777	0.9826	1.0843	48.9500	6.39
cqd32	-0.1159	0.9721	1.3030	41.9500	6.41
cqd33	-0.0918	0.9721	1.6847	37.0900	6.41
cqd34	0.0004	0.9626	0.5827	43.1400	6.45
cqd35	-0.1305	0.9826	1.1679	34.1000	6.54
cqd36	0.0643	0.9625	0.5645	42.7100	6.77
cqd37	-0.0685	0.9824	1.0919	20.6680	6.90

Tabelul 2. Aglomerarea în clusteri: derivați de carbochinonă

Pas	Cluster combinat		Coef	Momentul în care apare clusterul		Pasul următor
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	32	33	0.000	0	0	11
2	27	28	0.000	0	0	3
3	26	27	0.000	0	2	14
4	16	17	0.000	0	0	5
5	15	16	0.000	0	4	8
6	11	12	0.000	0	0	13
7	7	8	0.000	0	0	31
8	15	18	0.000	5	0	12
9	24	25	0.000	0	0	18
10	19	20	0.001	0	0	21
11	31	32	0.001	0	1	17
12	14	15	0.001	0	8	26
13	10	11	0.002	0	6	16
14	26	29	0.002	3	0	18
15	21	22	0.003	0	0	21
16	10	13	0.005	13	0	26
17	31	34	0.006	11	0	25
18	24	26	0.008	9	14	20
19	5	6	0.011	0	0	23
20	24	30	0.015	18	0	27
21	19	21	0.020	10	15	30
22	36	37	0.029	0	0	32
23	4	5	0.037	0	19	29
24	1	2	0.047	0	0	33
25	31	35	0.060	17	0	32
26	10	14	0.074	16	12	28
27	23	24	0.095	0	20	34
28	9	10	0.125	0	26	30
29	3	4	0.161	0	23	31
30	9	19	0.254	28	21	35
31	3	7	0.429	29	7	33
32	31	36	0.652	25	22	34
33	1	3	1.047	24	31	36
34	23	31	1.605	27	32	35
35	9	23	4.865	30	34	36
36	1	9	14.472	33	35	0

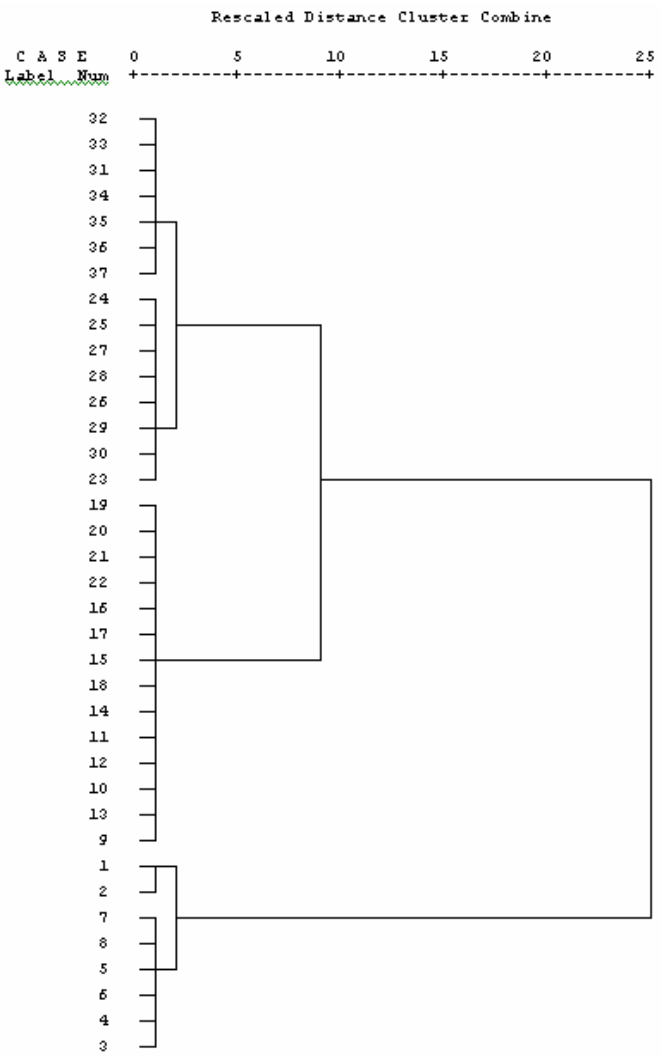


Figura 1. Dendrograma proprietății de interes a derivaților de carbochinonă (Metoda Ward)

Tabelul 3. Reorganizarea rezultatelor din Tabelul 2

Nr clusteri	CoefAglUltim	CoefAglPrev	Dif
2	14.4720	4.8650	9.6070
3	4.8650	1.6050	3.2600
4	1.6050	0.6520	**0.9530**
5	0.6520	0.4290	0.2230
6	0.4290	0.2540	0.1750
7	0.2540	0.1610	0.0930

CoefAglUltim = coeficientul de aglomerare cu valoarea mare pentru numărul de clusteri de interes;
CoefAglPrevc= coeficientul de aglomerare anterior; Dif = diferența dintre ultim și anterior;

Tabelul 4. Parametrii statistici asociați clusterilor: analiza de clusterizare ierarhică (proprietatea de interes a carbochinonelor)

Cluster	n	Min	Max	Media	StErr
1	8	4.33	5.16	4.7850	0.1058
2	14	5.46	5.86	5.6757	0.0283
3	15	6.03	6.90	6.3467	0.0630
All	37	4.33	6.90	5.7551	0.1042

n = volumul eșantionului; Min = valoarea minimă;
Max = valoarea maximă; Media = media artimetică;
StErr = eroarea standard.

Tabelul 5. ANOVA: proprietarea investigată a derivaților de carbochinonă

	SS	df	MS	F	p
Între clusteri	12.866	2	6.433	136.238	$5.84 \cdot 10^{-17}$
În clusteri	1.605	34	0.047		
Total	14.472	36			

SS = suma pătratelor erorilor; df = grade de libertate;
MS = media pătratelor erorilor; F = statistica Fisher;
p = semnificația statisticii Fisher

Analiza de clusterizare prin utilizarea metodei k-means cluster cu impunerea în căutare a 3 clusteri clasifică 3 compuşi în primul cluster (1-3, valoarea centrală a clusterului = 4.48), 20 de compuşi în cel de-al doiela cluster (4-23, valoarea centrală a custerului = 5.52) şi 14 compuşi în cel de-al treilea cluster (24-37, valoarea centrală a clusterului = 6.37). Parametrii statistici descriptivi pentru asociați analizei sunt redați în Tabelul 6.

Tabelul 6. Parametrii statistici asociați clusterilor: analiza de clusterizare k-medii

Cluster	n	Min	Max	Media	StErr
1	3	4.33	4.63	4.4767	0.0867
2	20	4.77	6.03	5.5170	0.0792
3	14	6.14	6.90	6.3693	0.0631
All	37	4.33	6.90	5.7551	0.1042

n = volumul eşantionului; Min = valoarea minimă;
Max = valoarea maximă; Media = media artimetică;
StErr = eroarea standard.

Distanța față de centrele finale ale clusterilor în analzia clusterilor pe baza mediilor este redată în Tabelul 7. Rezultatele testului ANOVA obținute pentru compararea mediilor celor 3 clusteri sunt redate în Tabelul 8.

Tabelul 7. Matricea distanței între centrele clusterelor: analiza de clusterizare pe baza mediilor

Cluster	1	2	3
1		1.040	1.893
2	1.040		0.852
3	1.893	0.852	

Tabelul 8. ANOVA: analiza de clusterizare pe baza mediilor

	SS	df	MS	F	p
Între clusteri	11.318	2	5.659	61.013	$5.63 \cdot 10^{-12}$
In clusteri	3.154	34	0.093		
Total	14.472	36			

SS = suma pătrateloer erorilor; df = grade de libertate;
MS = media pătratelor erorilor; F = statistica Fisher;
p = semnificația statisticii Fisher

Analiza ierarhică de clasificare s-a aplicat pe proprietatea investigată a derivaților de carbocinone şi cei patru descriptori MDFV identificați ca aparținând celui mai performant model

qSAR [16]. Sumarizarea rezultatelor obținute este prezentată în Tabelul 9. Deoarece variabilele nu aveau aceeaşi unitatea de măsură analiza de clusterizare s-a aplicat ulterior transformării datelor variabilelor ca şi date aparținând intervalului 0-1. Dendrograma obținută în clasificare prin utilizarea atât a proprietății cât şi a descriptorilor MDFV este prezentată în Tabelul 9.

Tabelul 9. Coeficienții asociați analizei ierarhice de clusterizare: proprietate & descriptori MDFV

Nr clusteri	CoefAglomLast	CoefAglPrev	Dif
2	11.94	7.79	4.15
3	7.79	5.87	1.92
4	5.87	4.83	1.04
5	4.83	3.80	1.03
6	3.80	3.01	0.79

CoefAglUltim = coeficientul de aglomerare cu valoarea mare pentru numărul de clusteri de interes;
CoefAglPrevc= coeficientul de aglomerare anterior
Dif = diferența dintre ultim şi anterior

Analiza rezultatelor prezentate în Tabelul 9 pune în evidență că numărul optim de clusteri este 2 (dacă analizăm ordinul de mărime).

Testul ANOVA a fost aplicat pentru a identifica contribuția semnificativă în clasificare pentru un număr fixat de doi clusteri. Parametrii statistici descriptivi asociați variabilelor sunt prezentați în Tabelul 10.

Apartenența compuşilor la cei doi clusteri a fost după cum urmează:

- Cluster 1: cqd01, cqd02, cqd03, cqd04; cqd05; cqd06, cqd07, cqd08, cqd09, cqd10, cqd11, cqd12, cqd14, cqd15, cqd16, cqd18, cqd21 şi cqd22.
- Cluster 2: cqd13, cqd17, cqd19, cqd20, cqd23, cqd24, cqd25, cqd26, cqd27, cqd28, cqd29, cqd30, cqd31, cqd32, cqd33, cqd34, cqd35, cqd36 şi cqd37

[16] Bolboacă SD, Jäntschi L. Comparison of QSAR Performances on Carboquinone Derivatives. TheScientificWorldJOURNAL 2009;9(10):1148-1166.

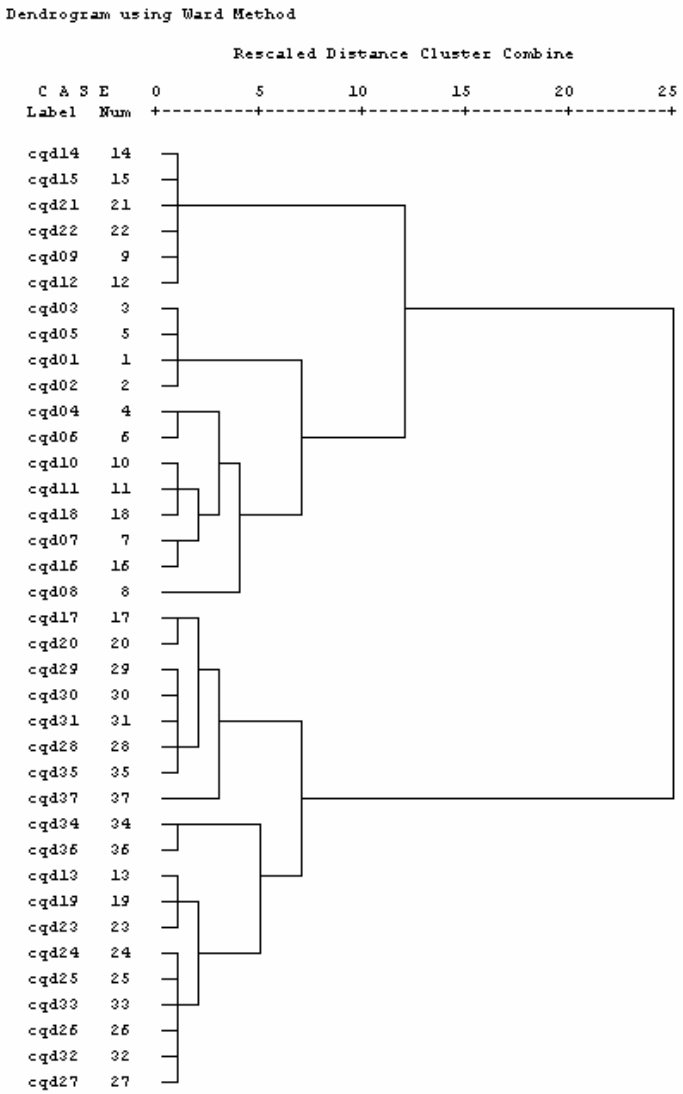


Figura 2. Dendrograma clasificării prin utilizarea proprietății şi a celor 4 descriptori MDFV

Tabelul 10. Rezultate statistică descriptivă: clasificare pe baza proprietății şi a valorilor descriptorilor MDFV

	Cluster	Efect	n	m	StDev	StErr	Min	Max	BCVar
TEuIFFDL	1		18	0.1102	0.0815	0.0192	-0.0045	0.3221	
	2		19	-0.0885	0.0622	0.0143	-0.1777	0.0643	
	Total		37	0.0082	0.1234	0.0203	-0.1777	0.3221	
	Model	Fix			0.0722	0.0119			
		Random				0.0994			0.019475
GLCIicdI	1		18	0.9895	0.0087	0.0020	0.9824	1.0000	
	2		19	0.9757	0.0103	0.0024	0.9589	0.9978	
	Total		37	0.9824	0.0117	0.0019	0.9589	1.0000	
	Model	Fix			0.0096	0.0016			
		Random				0.0069			8.96E-05
TAkaFcDL	1		18	1.4097	0.5724	0.1349	0.6881	2.3848	
	2		19	1.4138	0.4182	0.0959	0.5645	2.0179	
	Total		37	1.4118	0.4921	0.0809	0.5645	2.3848	
	Model	Fix			0.4991	0.0821			
		Random				0.0821			-0.01347
GLbIAcDR	1		18	48.6377	11.4632	2.7019	17.7280	59.7600	
	2		19	39.7620	5.6066	1.2862	20.6680	48.9500	
	Total		37	44.0799	9.8993	1.6274	17.7280	59.7600	

	Model	Fix			8.9437	1.4703			
		Random				4.4391			35.06175
Prop	1		18	5.2717	0.4948	0.1166	4.3300	5.8600	
	2		19	6.2132	0.3430	0.0787	5.6300	6.9000	
	Total		37	5.7551	0.6340	0.1042	4.3300	6.9000	
	Model	Fix			0.4235	0.0696			
		Random				0.4709			0.433499

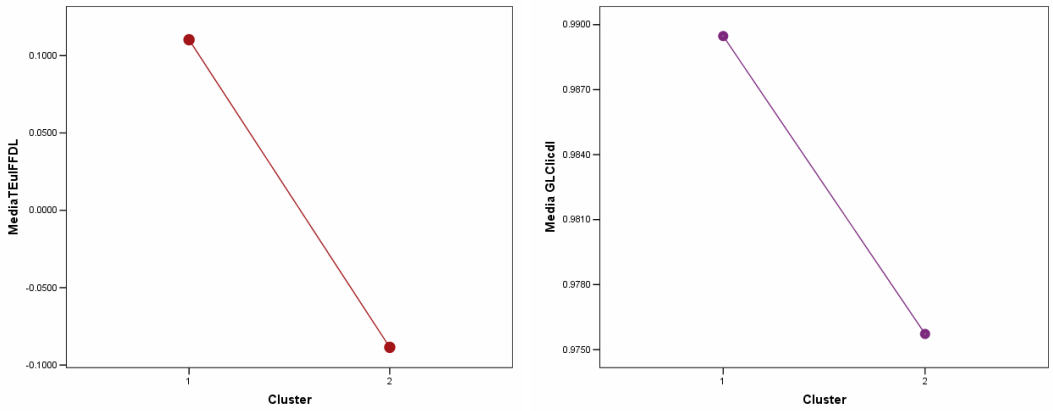
n = volumul eşantionului; m = media aritmetică; StDev = deviația standard;
StErr = eroarea standard; Min = valoarea minimă; Max = valoarea maximă;
BCVar = varianța între componente

Rezultatele testului ANOVA sunt prezentate în Tabelul 11. De remarcat distribuția mediile variabilelor în interiorul clusterilor (Figura 3).

Aşa cum rezultă din Tabelul 11 există un descriptor MDFV care nu are o contribuție semnificativă în clasificare: TAkaFcDL.

Tabelul 11. Testul ANOVA: clasificare în funcție de valorile proprietății și descriptorilor MDFV

Parametru	Clusteri	SS	df	MS	F	p
TEuIFFDL	Între	0.365244	1	0.365244	70.01103	$7.22 \cdot 10^{-10}$
	În	0.182593	35	0.005217		
	Total	0.547837	36			
GLCIicdI	Între	0.001748	1	0.001748	19.0958	0.000106
	În	0.003204	35	9.15E-05		
	Total	0.004951	36			
TAkaFcDL	Între	0.000158	1	0.000158	0.000632	0.980082
	În	8.718812	35	0.249109		
	Total	8.71897	36			
GLbIAcDR	Între	728.1592	1	728.1592	9.103054	0.004733
	În	2799.673	35	79.99065		
	Total	3527.832	36			
Prop	Între	8.193264	1	8.193264	45.67429	$7.85 \cdot 10^{-8}$
	În	6.278461	35	0.179385		
	Total	14.47172	36			



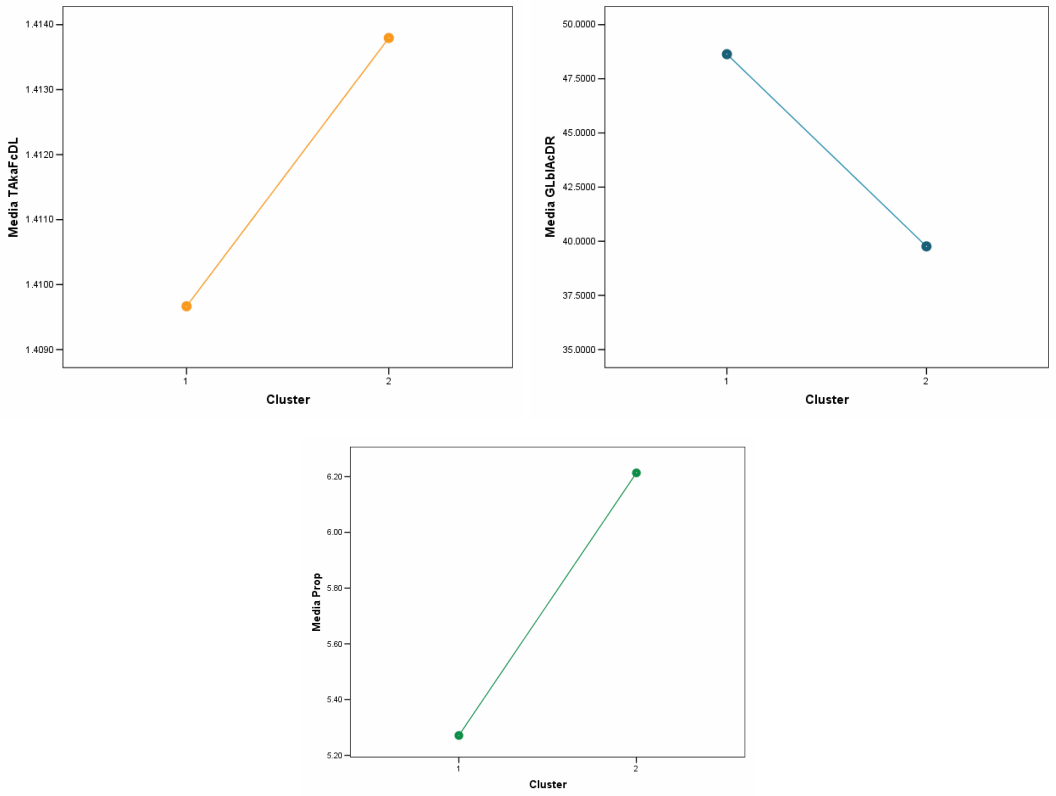


Figura 3. Contribuții medii în interiorul clusterilor

Aplicarea testului Welch de comparare a mediilor a pus în evidență următoarele:

- Diferență semnificativă statistic în ceea ce privește mediile în clusteri pentru descriptorul TEuIFFDL (Statistica Welch = 68.992, df1 = 1, df2 = 31.80, p = $1.81 \cdot 10^{-9}$)
- Diferență semnificativă statistic în ceea ce privește mediile în clusteri pentru descriptorul GLCIicdI (Statistica Welch = 19.284, df1 = 1, df2 = 34.493, p = $1.07 \cdot 10^{-4}$)
- Diferență semnificativă statistic în ceea ce privește mediile în clusteri pentru descriptorul GLbIAcDR (Statistica Welch = 8.797, df1 = 1, df2 = 24.395, p = 0.007)
- Diferență semnificativă statistic în ceea ce privește mediile în clusteri pentru descriptorul Prop (Statistica Welch = 44.792, df1 = 1, df2 = 30.11, p = $2.01 \cdot 10^{-7}$).

Distribuția valorilor în cadrul claselor pentru variabilele cu contribuție semnificativă statistic la clasificare sunt redate în Figura 4.

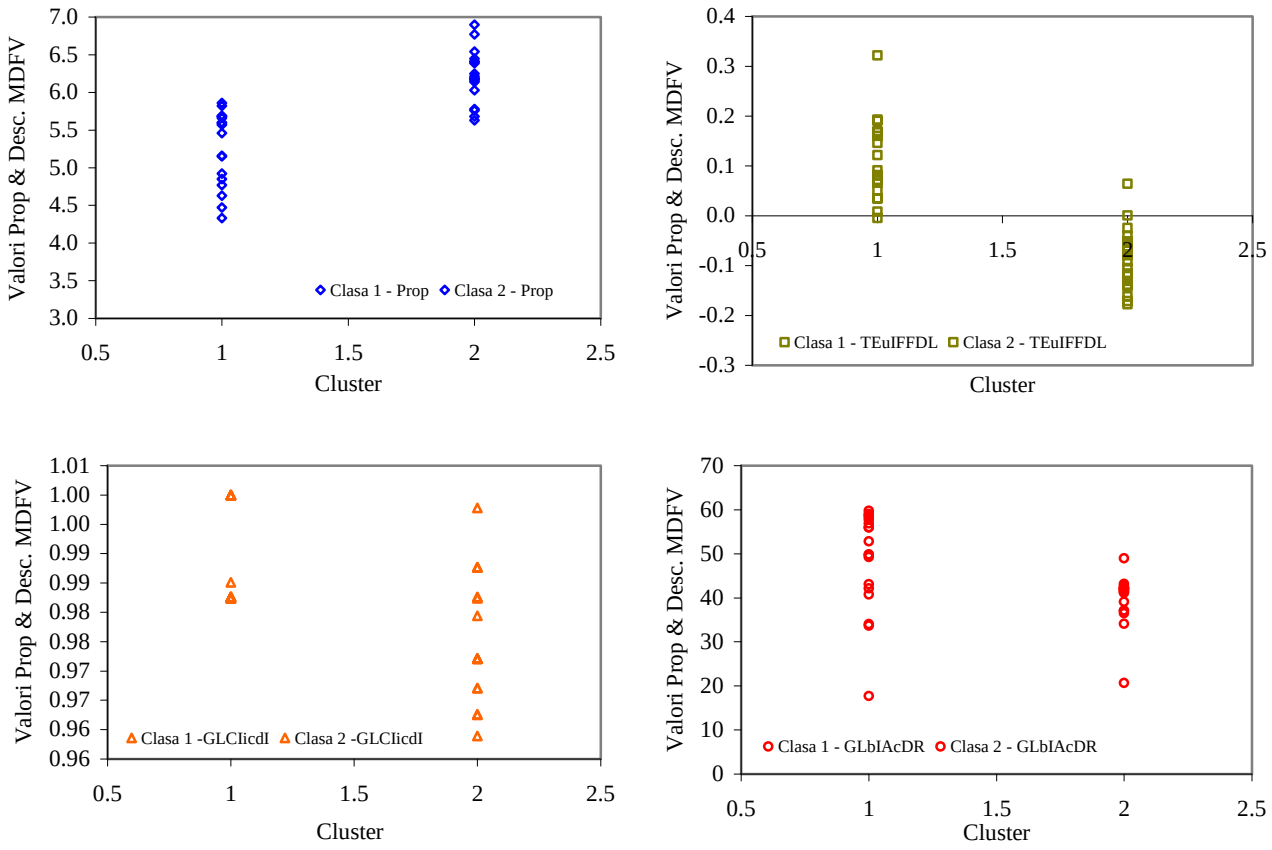


Figura 4. Distribuția valorilor variabilelor cu contribuție semnificativă statistic în clasificare

Următoarele concluzii se pot desprinde pe baza analizei de clusterizare realizată pe derivații de carbochinone:

- Analiza ierarhică de cluterizare a permis identificarea numărului optim de clusteri: în ceea ce privește proprietatea măsurată a derivaților de carbochinonă clasificarea optimă se face prin utilizarea a 3 clase (mediile celor trei clase sunt: 4.7850 - 5.6757 - 6.3467).
- Utilizarea metodei k-means (știut fiind că numărul optim de clusteri este egal cu 3) clasifică diferit compușii pe baza proprietății măsurate cu mediile pe cele trei clase egale cu: 4.48 – 5.52 – 6.37.
- Atât metode ierarhică de clasificare cât și metoda k-medii s-au dovedit a fi semnificative statistic la un prag de semnificație de 5%.
- Analiza de clasificare a compușilor pe baza valorilor proprietății măsurate și a descriptorilor moleculari ca și variabile a identificat un număr optim de 2 clase.
- Analiza de clasificare a compușilor pe baza valorilor proprietății măsurate și a descriptorilor moleculari atunci când se investighează moleculele a evidențiat un model semnificativ statistic dar cu diferen e semnificative statistic a mediilor doar a 3 descriptori MDFV și respectiv a proprietății de interes.

4.1.1.2. Compuşi organici – traversare barieră hemato-encefalică

Analiza de clasificare pentru compuşii organici care traversează bariera hemato-encefalică s-a realizat pentru modelul următor. Modelul a fost ob inut în conformitate cu principiile de analiză care se regăsesc în [17-34]:

[17] Bolboacă SD, Jäntschi L. Modelling the property of compounds from structure: statistical methods for models validation. Environmental Chemistry Letters 2008;6:175-181.

[18] Bolboacă SD. Assessment of Random Assignment in Training and Test Sets using Generalized Cluster Analysis Technique. Appl Med Inform 2010;28(2):9-14.

[19] Bolboacă SD, Jäntschi L. Dependence between determination coefficient and number of regressors: a case study on retention times of mycotoxins. Studia Universitatis Babes-Bolyai Chemia. Submitted manuscript.

[20] Jäntschi L, Bolboacă SD. Observation vs. Observable: Maximum Likelihood Estimations according to the Assumption of Generalized Gauss and Laplace Distributionss. Leonardo El J Pract Technol 2009;8(15):81-104.

[21] Jäntschi L, Bolboacă SD. Distribution Fitting 2. Pearson-Fisher, Kolmogorov-Smirnov, Anderson-Darling, Wilks-Shapiro, Kramer-von-Misses and Jarque-Bera statistics. Bulletin of University of Agricultural Sciences and Veterinary Medicine Cluj-Napoca. Horticulture 2009;66(2): 691-697.

[22] Bolboacă SD, Jäntschi L. Structure-Property Based Model for Alkanes Boiling Points. International Journal of Pure and Applied Mathematics 2008;47(1): 23-30.

[23] Stoenoiu CE, Bolboacă SD, Jäntschi L. Model Formulation & Interpretation - From Experiment to Theory. International Journal of Pure and Applied Mathematics 2008;47(1):9-16.

[24] Bolboacă SD, Pică EM, Cimpoiu CV, Jäntschi L. Statistical Assessment of Solvent Mixture Models Used for Separation of Biological Active Compounds. Molecules 2008;8(13):1617-1639.

[25] Bolboacă SD, Jäntschi L. Modelling Analysis of Amino Acids Hydrophobicity. MATCH Communications in Mathematical and in Computer Chemistry 2008;60(3):1021-1032.

[26] Jäntschi L, Bolboacă SD. A Structural Modelling Study on Marine Sediments Toxicity. Marine Drugs 2008;6(2):372-388.

[27] Bolboacă SD, Jäntschi L. A Structural Informatics Study on Collagen. Chemical Biology & Drug Design 2008;71(2):173-179.

[28] Jäntschi L, Bolboacă SD, Diudea MV. Chromatographic Retention Times of Polychlorinated Biphenyls: from Structural Information to Property Characterization, International Journal of Molecular Sciences, MDPI, 8(11), 1125-1157, 2007

[29] Jäntschi L, Bolboacă SD. Structure versus biological role substituted thiadiazole - and thiadiazoline – disulfonamides. Studii si Cercetari Stiintifice - Seria Biologie 2004;12:50-56.

[30] Jäntschi L, Bolboacă SD. Triazines herbicidal assessed activity. Studii si Cercetari Stiintifice - Seria Biologie 2007;12:57-62.

$$\hat{Y}_{logBB} = 0.5370(\pm0.30) - 8.4411(\pm4.42)\times TLgFAIDI - 497.0205(\pm144.97)\times GAmIAaDI + 4.1129(\pm1.55)\times TAgFIADL - 3.1303(\pm1.26)\times TAgPIADL$$

$R = 0.7816$ (95%CI_r [0.6791-0.8541]), $R^2 = 0.6109$;
$se_{est} = 0.61$; $n_{tr} = 81$; F_{est} (p) = 30 ($6.41\cdot10^{-15}$)
t_{X1} (p) = 3.59 ($5.84\cdot10^{-4}$); t_{X2} (p) = -3.80 ($2.87\cdot10^{-4}$); t_{X2} (p) = -6.83 ($1.85\cdot10^{-9}$);
t_{X4} (p) = 5.30 ($1.11\cdot10^{-6}$); t_{X5} (p) = -4.96 ($4.21\cdot10^{-6}$);
$R_{loo} = 0.7334$; $R^2_{loo} = 0.5378$; $s_{loo} = 0.65$; F_{loo} (p) = 22 ($4.27\cdot10^{-12}$);
R (p) = 0.7816 ($7.31\cdot10^{-18}$); r_{sQ} (p) = 0.7636 ($9.18\cdot10^{-17}$);
ρ (p) = 0.7460 ($8.91\cdot10^{-16}$); τ_a (p) = 0.5568 ($1.37\cdot10^{-10}$); τ_b (p) = 0.5578 ($1.53\cdot10^{-10}$);
τ_c (p) = 0.5499 ($2.16\cdot10^{-10}$); Γ (p) = 0.5589 ($8.86\cdot10^{-5}$)

unde $\hat{Y}_{logBB}$ = proprietatea estimată de modeluk MDFV; TLgFAIDI (X_1), GAmIAaDI (X_2), TAgFIADL (X_3), and TAgPIADL (X_4) = descriptori MDFV [35]; valorile din parantezele rotunde permit prin scădere respectiv adunare obținerea intervalului de încredere de 95% asociat; R = coeficientul de corelație; R^2 = coeficientul de determinare; s_{est} = eroarea standard a estimatului; n_{tr} = volumul eşantionului – setul de învățare; F_{est} (p) = valoarea statisticii Fisher (valoarea probabilității de eroare asociată statisticii F); t = valoarea statisticii Student; R^2_{loo} = pătratul coeficientului de cros validare în analiza lasă unul afară; s_{loo} = eroarea standar a prezisului; F_{loo} = statistica Fisher în analiza lasă-unul-afară; [] = limitele intervalului de confidență la un prag de semnificație de 5%; r = coeficientul de corelație Pearson între proprietatea observată şi valoarea estimată de către model; r_{sQ} = coeficientul de corelație semi-cantitativ [36]; ρ = coeficientul de corelație al rangurilor Spearman [37]; τ_a, τ_b, τ_c = coeficienți de corelație Kendall [38, 39]; Γ = coeficientul de corelație Gamma [40, 41, 42].

---

[31] Jäntschi L, Bolboacă SD. Structure-Activity Relationships on the Molecular Descriptors Family Project at the End. Leonardo El J Pract Technol2007;11:163-180.

[32] Bolboacă SD, Jäntschi L. Antiallergic Activity of Substituted Benzamides: Characterization, Estimation and Prediction. Clujul Medical 2007;LXXX(1):125-132.

[33] Jäntschi L, Bolboacă SD. Modeling the octanol-water partition coefficient of substituted phenols by the use of structure information. International Journal of Quantum Chemistry 2007;107(8):1736-1744.

[34] Jäntschi L, Bolboacă SD. The Jungle of Linear Regression Revisited. Leonardo El J Pract Technol 2007;10:169-187.

[35] Jäntschi L, Stoenoiu CE, Bolboacă SD. A Formula for Vertex Cuts in b-Trees. International Journal of Pure and Applied Mathematics 2008;47(1):17-22.

[36] Bolboacă S, Jäntschi L. Pearson Versus Spearman, Kendall's Tau Correlation Analysis on Structure-Activity Relationships of Biologic Active Compounds. Leonardo J Sci2006;9:179-200.

[37] Spearman C. General intelligence" objectively determined and measured. American Journal of Psychology 1904;15: 201-293.

[38] Kendall MG. A New Measure of Rank Correlation. Biometrika 1938;30:81-89.

[39] Kendall MG. Partial rank correlation. Biometrika 1942;32(3-4):277-283.

Analiza de clasificare s-a realizat pe baza datelor prezentate în Tabelul 12.

Tabelul 12. Date experimentale: Compuşi organici ce traversează bariera hemato-encefalică

Mol	TLgFAIDI	GAmIAaDI	TAgFIADL*	TAgPIADL*	logBBB
002_72108	0.0329	0.0052	-1.0252	-1.5745	-2.00
004_2803	0.0205	0.0020	-1.4967	-2.0460	0.11
005_4992	0.0014	0.0003	-1.1392	-1.6885	0.49
006_3696	0.0008	0.0005	-1.0499	-1.5992	0.83
008_50287	0.0565	0.0034	-0.8908	-1.4401	-0.82
011_241	0.0003	0.0002	0.0000	0.0000	0.37
012_7282	0.0015	0.0001	0.0000	0.0000	1.01
013_11507	0.0015	0.0002	0.0000	0.0000	0.90
014_3776	0.0320	0.0000	0.0000	0.0000	-0.15
015_6560	0.0323	0.0001	0.0000	0.0000	-0.17
018_6278	0.0109	0.0000	0.0000	0.0000	0.40
020_3226	0.0969	0.0001	0.0000	0.0000	0.24
022_9844	0.0332	0.0001	0.0000	0.0000	0.13
023_3562	0.0427	0.0000	0.0000	0.0000	0.35
024_8900	0.0012	0.0002	0.0000	0.0000	0.81
028_947	0.0547	0.0000	-2.1915	-2.7408	0.03
032_31300	0.0647	0.0000	0.0000	0.0000	0.27
033_1140	0.0007	0.0002	0.0000	0.0000	0.37
034_2244	0.0640	0.0003	0.0000	0.0000	-0.50
035_4737	0.0965	0.0005	-2.5462	-3.0955	0.12
037_338	0.0639	0.0002	0.0000	0.0000	-1.10
038_5566	0.0332	0.0005	-0.6629	-1.2122	1.44
039_3121	0.0333	0.0002	0.0000	0.0000	-0.22
040_2520	0.0311	0.0005	-2.2922	-2.8415	-0.70
041_5726	0.1332	0.0012	-1.3222	-1.8715	-0.72
043_5452	0.0012	0.0006	-0.9837	-1.5330	0.24
045_192706	0.0331	0.0022	-0.9077	-1.4571	1.00
050_4926	0.0008	0.0005	-1.0678	-1.6171	1.23
051_4463	0.0328	0.0024	-0.8923	-1.4416	0.00
052_3035905	0.0048	0.0017	-1.0524	-1.6017	-0.16
054_3672	0.0329	0.0002	0.0000	0.0000	-0.18
056_2153	0.0644	0.0037	-1.0929	-1.6422	-0.29
057_1983	0.0635	0.0002	-2.6518	-3.2010	-0.31
058_948	0.0464	0.0000	-2.1915	-2.7408	0.03
059_6348	0.0054	0.0000	0.0000	0.0000	0.60
060_3715	0.0750	0.0006	-1.8896	-2.4389	-1.26
061_5362440	0.1273	0.0038	-0.7268	-1.2761	-0.75
062_4616	0.0735	0.0005	-1.8298	-2.3791	0.61
064_2555	0.0323	0.0005	-1.4135	-1.9629	-0.35
065_2160	0.0008	0.0003	-2.0325	-2.5818	0.88
066_2995	0.0008	0.0005	-1.2238	-1.7731	1.00
069_4205	0.0007	0.0024	-0.7032	-1.2525	0.53

[40] Goodman LA, Kruskal WH. Measures of association for cross classifications. Part I. J Amer Statist Assoc 1954;49:732-764.

[41] Goodman LA, Kruskal WH. Measures of association for cross classifications. Part II. J Amer Statist Assoc 1959;52:123-163.

[42] Goodman LA, Kruskal WH. Measures of association for cross classifications. Part III. J Amer Statist Assoc 1963;58:310-364.

070_21844	0.0647	0.0006	0.0000	0.0000	0.40
073_475100	0.0959	0.0025	-0.7642	-1.3135	-0.02
077_14922095	0.0399	0.0033	-0.9777	-1.5270	-0.66
078_2992532	0.0281	0.0025	-1.5270	-2.0763	-0.18
080_10442225	0.0565	0.0046	-0.8054	-1.3548	-1.54
081_10442293	0.0404	0.0037	-0.9069	-1.4562	-1.12
082_9971484	0.0320	0.0003	-1.5994	-2.1487	-0.46
084_3167851	0.0329	0.0006	-1.7270	-2.2763	0.30
085_2276	0.0007	0.0002	-1.7386	-2.2879	-0.30
086_72747	0.0008	0.0002	-1.4626	-2.0119	-0.06
087_2519	0.0648	0.0037	-0.9751	-1.5244	-2.00
088_2708	0.0535	0.0004	-2.9160	-3.4650	-1.60
093_1775	0.0636	0.0004	-2.3112	-2.8605	-2.20
094_4946	0.0324	0.0003	-3.1241	-3.6730	-1.20
095_444349	0.1646	0.0004	-3.2560	-3.8060	-4.10
096_6575	0.0205	0.0000	0.0000	0.0000	0.34
097_450682	0.0923	0.0003	-2.4979	-3.0472	-0.52
100_8036856	0.0281	0.0005	-1.9209	-2.4702	0.00
101_8620184	0.0324	0.0002	-1.8035	-2.3528	-0.02
103_BBCPD24	0.0004	0.0008	-1.1135	-1.6628	0.44
105_6168	0.0426	0.0000	0.0000	0.0000	0.08
106_T7	0.0008	0.0005	-2.0325	-2.5818	0.85
107_23218171	0.0403	0.0029	-0.9200	-1.4693	-0.73
108_BBCPD18	0.0404	0.0020	-0.8184	-1.3678	-0.27
110_BBCPD16	0.0598	0.0046	-1.2836	-1.8329	-1.57
113_YG16	0.0281	0.0001	-2.1441	-2.6934	-0.42
115_5854406	0.0281	0.0024	-1.3487	-1.8980	-1.40
116_117961	0.0646	0.0003	-1.9691	-2.5184	-0.43
117_4916	0.0330	0.0003	-1.9150	-2.4643	0.25
118_CBZEPO	0.0333	0.0003	-2.2938	-2.8431	-0.34
120_114837	0.0429	0.0048	-0.7204	-1.2697	-0.30
121_8560187	0.0740	0.0037	-0.7077	-1.2570	-1.34
122_8267285	0.1056	0.0041	-0.7132	-1.2626	-1.82
124_7972174	0.0429	0.0002	-1.0488	-1.5981	1.64
125_8083053	0.0645	0.0020	-0.6744	-1.2237	0.16
126_23342331	0.0323	0.0005	-1.7364	-2.2858	0.52
127_23342332	0.0103	0.0004	-1.7167	-2.2660	0.39
129_SKF93319	0.0324	0.0020	-1.1973	-1.7466	-1.30
130_CBZ	0.0333	0.0002	-2.3290	-2.8783	0.00
001_2756 #	0.0292	0.0040	-1.1630	-1.7123	-1.42
003_51671 #	0.0597	0.0056	-1.1622	-1.7115	-1.06
007_5039#	0.0412	0.0030	-1.0968	-1.6461	-1.23
009_91769#	0.0002	0.0007	-1.0264	-1.5757	0.14
010_6569#	0.0322	0.0001	0.0000	0.0000	-0.08
016_7892#	0.0011	0.0001	0.0000	0.0000	0.97
017_580244#	0.0011	0.0001	0.0000	0.0000	1.04
019_3283#	0.0011	0.0001	0.0000	0.0000	0.00
021_702#	0.0322	0.0000	0.0000	0.0000	-0.16
025_8058#	0.0012	0.0002	0.0000	0.0000	0.80
026_3763#	0.0749	0.0001	0.0000	0.0000	0.42
027_7296#	0.0008	0.0002	0.0000	0.0000	0.93
029_8003#	0.0011	0.0001	0.0000	0.0000	0.76
030_1031#	0.0322	0.0001	0.0000	0.0000	-0.16
031_180#	0.0320	0.0000	0.0000	0.0000	-0.15
036_5983#	0.0336	0.0025	-0.8784	-1.4277	0.08
042_3658#	0.0426	0.0004	-1.3182	-1.8675	0.39
044_2118#	0.0108	0.0023	-0.9736	-1.5229	0.04
046_4192#	0.0429	0.0024	-1.0993	-1.6486	0.36

048_5284371#	0.0329	0.0005	-1.3813	-1.9306	0.55
049_2726#	0.0109	0.0005	-1.0757	-1.6250	1.06
053_3043#	0.0639	0.0033	-1.1528	-1.7021	-1.30
055_2206#	0.0327	0.0002	-1.9546	-2.5039	-2.00
063_2554#	0.0323	0.0005	-1.4280	-1.9773	-0.14
067_4184#	0.0007	0.0005	-0.8507	-1.4000	0.99
068_166560#	0.0327	0.0005	-1.4609	-2.0102	0.82
071_3151#	0.0735	0.0024	-0.6160	-1.1653	-0.78
072_5073#	0.0643	0.0016	-0.7477	-1.2970	-0.67
074_55482#	0.0393	0.0052	-1.0394	-1.5887	-1.88
079_104391#35	0.0555	0.0049	-1.3810	-1.9303	-1.15
083_10498206#	0.0319	0.0004	-1.5622	-2.1115	-0.24
089_750#	0.0593	0.0000	-4.3890	-4.9380	-3.50
091_5288826#	0.0639	0.0005	-1.3803	-1.9296	-2.70
092_994#	0.0593	0.0002	-3.6720	-4.2210	-1.30
102_BBCPD23#	0.0003	0.0008	-1.0947	-1.6440	0.69
104_BBCPD26#	0.0002	0.0008	-1.0264	-1.5757	0.22
109_BBCPD19#	0.0404	0.0047	-0.6374	-1.1867	-0.28
111_BBCPD14#	0.0398	0.0039	-0.9918	-1.5411	-0.12
114_YG19#	0.0281	0.0003	-2.4896	-3.0389	-1.30
123_143157#	0.0108	0.0005	-1.4459	-1.9952	1.03
128_ICI17148#	0.0286	0.0025	-1.5629	-2.1122	-0.04

* values different at more than 3 decimals;
# compounds in test set'

Rezultatele obținute în investigarea proprietății exprimată în scară logaritmică (Tabelul 13) pun în evidență existența unui număr optim de clase egal cu 2, respectiv egal cu 4.

Tabelul 13. Sumarizarea rezultatelor: aglomerarea compuşilor

Nr clusteri	CoefAglomLast	CoefAglPrev	Dif
2	115.4226	40.6948	74.7278
3	40.6948	26.1063	14.5885
4	26.1063	14.1194	11.9869
5	14.1194	9.5480	4.5714
6	9.5480	5.7101	3.8378
7	5.7101	3.9018	1.8083

CoefAglUltim = coeficientul de aglomerare cu valoarea mare pentru numărul de clusteri de interes;
CoefAglPrevc= coeficientul de aglomerare anterior;
Dif = diferența dintre ultim şi anterior;

Distribuția compuşilor în funcție de utilizarea unui număr fix de clusteri a fost următoarea:

- 2 clusteri (Figura 5): valorile centrale ale clusterilor -1.85 primul cluster şi 0.12 cel de-al doilea cluster
  - Cluster 1: 21 compuşii (002_72108; 060_3715; 080_10442225; 087_2519; 088_2708; 093_1775; 095_444349; 110_BBCPD16; 115_5854406; 121_8560187; 122_8267285; 129_SKF93319; 001_2756; 007_5039; 053_3043; 055_2206; 074_55482; 089_750; 091_5288826; 092_994 şi 114_YG19).
  - Cluster 2: 101 compuşii (cei care nu au fost menționați anterior).

- Aşa cum reiese din reprezentarea grafică (Figura 5) există 2 compuşi care au fost clasificați ca aparținând primului cluster dar care însă sunt valori extreme. Cu toate acestea, normalitatea proprietății măsurate nu poate fi respinsă la un prag de semnificație de 5% (statistica Kolmogorov-Smirnov = 0.229, p = 0.1889; statistica Chi-Square = 1.6994, p = 0.1924).

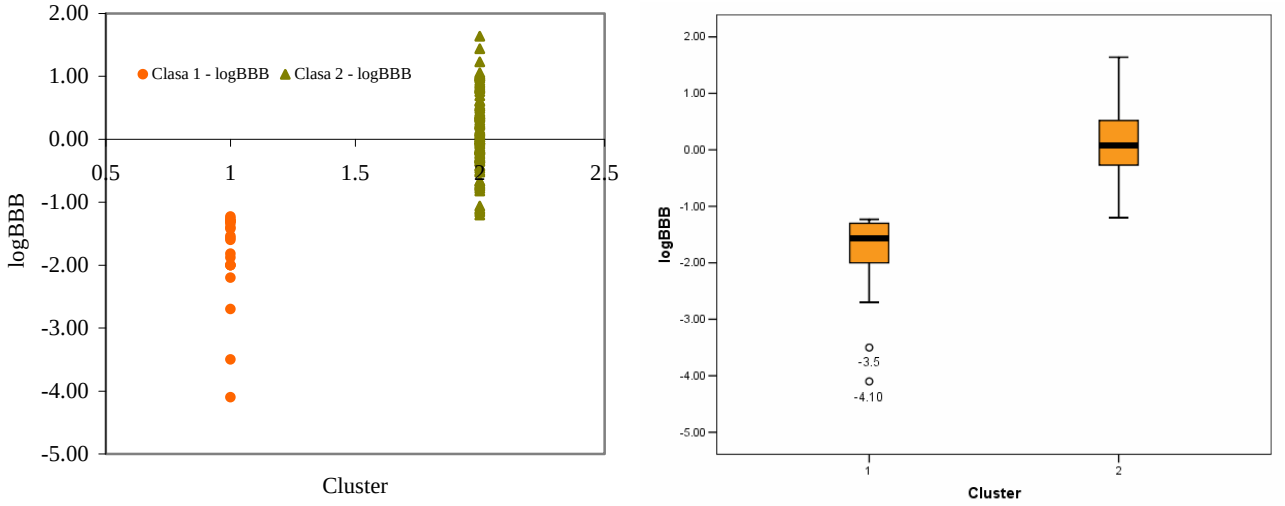


Figura 5. Distribuția valorilor logBBB în funcție de cei 2 clusteri (valorile extreme corespund compuşilor 095_444349 şi respectiv 089_750)

- 4 clusteri (Figura 6):
  - Cluster 1: 18 compuşi (002_72108; 060_3715; 080_10442225; 087_2519; 088_2708; 093_1775; 110_BBCPD16; 115_5854406; 121_8560187; 122_8267285; 129_SKF93319; 001_2756; 053_3043; 055_2206; 074_55482; 091_5288826; 092_994 şi 114_YG19).
  - Cluster 2: 81 compuşi (compuşii nespecificați ca aparținând celorlați clusteri).
  - Cluster 3: 2 compuşi (095_444349 şi 089_750) cu valorile extreme identificate în clusterul 1 al clasificării în 2 clase.
  - Cluster 4: 21 compuşi (006_3696; 012_7282; 013_11507; 024_8900; 038_5566; 045_192706; 050_4926; 065_2160; 066_2995; 106_T7; 124_7972174; 016_7892; 017_580244; 025_8058; 027_7296; 029_8003; 049_2726; 067_4184; 068_166560; 102_BBCPD23 şi 123_143157).

Şi în cazul clasificării în 4 clase există un compus ce poate fi considerat outlier (valoarea proprietății 1.64, clusterul 4) şi respectiv un compus cu valoare extremă (valoarea proprietății 1.44). Dar, nici în acest caz normalitatea datelor experimentale pentru clsuterul 4 nu poate fi respinsă la un prag de semnificație de 5% (statistica Kolmogorov-Smirnov = 0.2255, p = 0.2026; statistica Chi-Square = 0.3617, p = 0.5476)

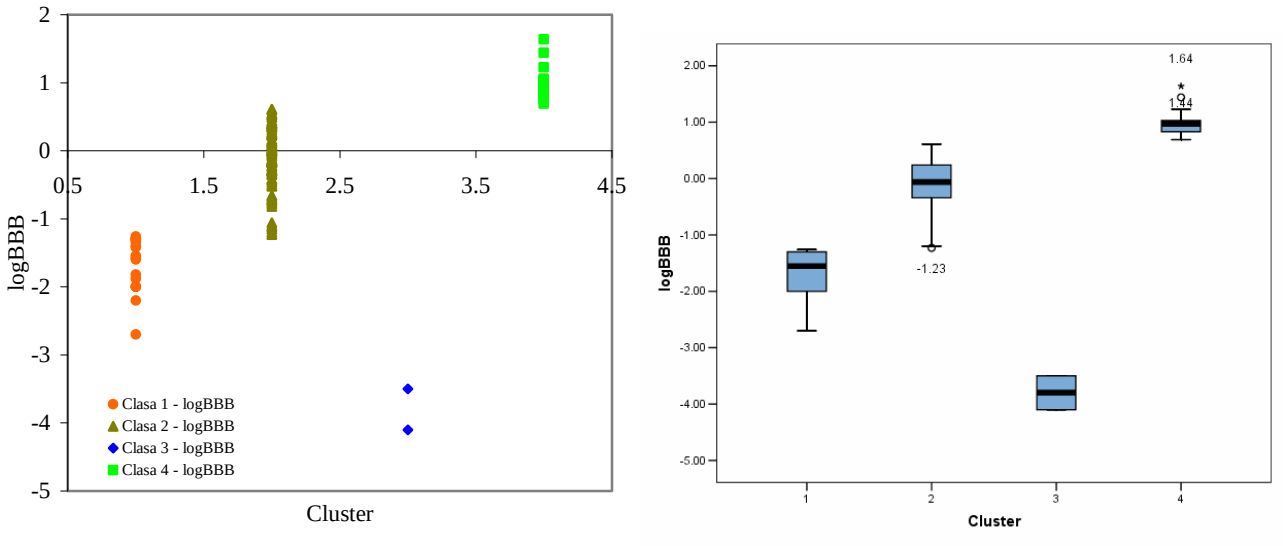


Figura 6. Distribuția valorilor logBBB în funcție de cei 4 clusteri

Parametrii statistici pentru fiecare cluster în parte sunt prezentați în Tabelul 14 pentru analiza cu 2 clusteri şi în Tabelul 15 pentru modelul de clasificare cu 4 clusteri.

Tabelul 14. Parametrii statistici: modelul cu 2 clusteri

Cluster	n	Min	Max	Media	StDev
1	21	-4.10	-1.23	-1.85	0.76
2	101	-1.30	1.64	0.11	0.62

n = volumul eşantionului; Min = valoarea minimă; Max = valoarea maximă;
Media = media artimetică; StDev = deviația standard.

Tabelul 15. Parametrii statistici: modelul cu 4 clusteri

Cluster	n	Min	Max	Media	StDev
1	18	-2.70	-1.26	-1.66	0.40
2	81	-1.23	0.61	-0.12	0.46
3	2	-4.10	-3.50	-3.80	0.42
4	21	0.69	1.64	0.98	0.22

n = volumul eşantionului; Min = valoarea minimă; Max = valoarea maximă;
Media = media artimetică; StErr = eroarea standard.

Modelul de clasificare care utilizează 2 clusteri s-a dovedit a fi semnificativ statistic (Tabelul 16) la fel ca şi modelul care a utilizat 4 clusteri (Tabelul 17). Pentru modelul de clasificare cu două clase varianțele s-au dovedit a fi omogene (statistica Levene = 0.278, df1 = 1, df2 = 120, p = 0.5987).

Tabelul 16. ANOVA: compuşi organici – model cu 2 clusteri

	SS	df	MS	F	p
Între clusteri	67.221	1	67.211	167.290	$1.60 \cdot 10^{-24}$
În clusteri	48.212	120	0.402		
Total	115.423	121			

SS = suma pătrateloer erorilor; df = grade de libertate;
MS = media pătratelor erorilor; F = statistica Fisher;
p = semnificația statisticii Fisher

Tabelul 17. ANOVA: compuşi organici – model cu 4 clusteri

	SS	df	MS	F	p

Între clusteri	94.338	3	31.463	176.497	$1.89 \cdot 10^{-43}$
În clusteri	21.035	118	0.178		
Total	115.423	121			

SS = suma pătrateloer erorilor; df = grade de libertate;
MS = media pătratelor erorilor; F = statistica Fisher;
p = semnificația statisticii Fisher

Egalitatea mediilor pentru logBBB a fost analizată prin aplicarea testului Welch. Valoarea statisticii Welch a fost de 124.408 (df1 = 1, df2 = 25.555, $p = 2.58 \cdot 10^{-11}$) pentru 2 clusteri şi respectiv 224.963 (df1 = 3, df2 = 4.805, $p = 1.36 \cdot 10^{-5}$).

Rezultatul obținut susține existența unei diferențe semnificative statistic între mediile logBBB atât pentru 2 cât şi pentru 4 clase în clasificarea bazată pe valoarea proprietății măsurate.

Analiza rezultatelor testelor ANOVA evidențiază două modele de clasificare semnificative statistic, modelul cu 4 clase fiind însă mai bun în termeni de semnificație.

Analiza de clasificare a fost aplicată de asemenea pe logBBB şi cei 4 descriptiori MDFV utilizați de către modelul qSAR cu cel mai mare grad de performanță. Analiza s-a aplicat prin impunerea de transformare a datelor în intervalul [0, +1] deoarece nu toate datele experimentale au avut aceeaşi unitate de măsură. Analiza a fost aplicată prin aplicarea metodei Wards şi a distanței Euclidiene aplicată pe cazuri.

Rezultatele analizei sunt prezentate în Tabelul 18. Din analiza rezultatelor din Tabelul 18 rezultă că numărul optim de clase este egal cu 2.

Tabelul 18. Coeficienții asociați analizei ierarhice de clusterizare: proprietate & descriptori MDFV

Nr clusteri	CoefAglomLast	CoefAglPrev	Dif
2	15.6439	10.8301	4.8138
3	10.8301	8.9003	1.9297
4	8.9003	7.1415	1.7588
5	7.1415	5.9716	1.1698
6	5.9716	5.4456	0.5260
7	5.4456	4.9549	0.4907

CoefAglUltim = coeficientul de aglomerare cu valoarea
mare pentru numărul de clusteri de interes;
CoefAglPrevc= coeficientul de aglomerare anterior
Dif = diferența dintre ultim şi anterior.

Distribuția compuşilor în funcție per cluster prin impunerea unui număr de 2 clase a fost următoarea:

- Cluster 1: 11 compuşi (057_1983; 088_2708; 093_1775; 094_4946; 095_444349; 097_450682; 055_2206; 089_750; 091_5288826; 092_994 şi 114_YG19)
- Cluster 2: 111 compuşi (restul compuşilor ne-enumerați anterior).

Testul ANOVA a fost aplicat pentru a identifica contribuția semnificativă în clasificare pentru un număr fixat de trei clusteri. Mediile variabilelor incluse în analiză în funcție de cluster au fost următoarele:

Variabile incluse în clasificare	Clasa	
	1	2
TLgFAIDI	0.0648	0.0347
GAmIAaDI	0.0003	0.0013
TAgFIADL	-2.7857	-0.9341
TAgPIADL	-3.3349	-1.3300
logBBB	-1.88	-0.05

Parametrii statistici descriptivi asociați variabilelor sunt prezentați în Tabelul 19.

Tabelul 19. Rezultate statistică descriptivă: clasificare pe baza proprietății şi a valorilor descriptorilor MDFV

Variable	Cluster	Effects	n	m	StDev	StErr	Min	Max	BCVar
TLgFAIDI	1		11	0.0648	0.0379	0.0114	0.0281	0.1646	
	2		111	0.0347	0.0287	0.0027	0.0002	0.1332	
	Total		122	0.0374	0.0307	0.0028	0.0002	0.1646	
	Model	Fixed			0.0295	0.0027			
		Random				0.0187			0.0004
GAmIAaDI	1		11	0.0003	0.0001	0.0000	0.0000	0.0005	
	2		111	0.0013	0.0016	0.0001	0.0000	0.0056	
	Total		122	0.0012	0.0015	0.0001	0.0000	0.0056	
	Model	Fixed			0.0015	0.0001			
		Random				0.0006			0.0000
TAgFIADL	1		11	-2.7857	0.8239	0.2484	-4.3890	-1.3803	
	2		111	-0.9341	0.7123	0.0676	-2.5462	0.0000	
	Total		122	-1.1011	0.8949	0.0810	-4.3890	0.0000	
	Model	Fixed			0.7223	0.0654			
		Random				1.1897			1.6881
TAgPIADL	1		11	-3.3349	0.8238	0.2484	-4.9380	-1.9296	
	2		111	-1.3300	0.9262	0.0879	-3.0955	0.0000	
	Total		122	-1.5108	1.0810	0.0979	-4.9380	0.0000	
	Model	Fixed			0.9181	0.0831			
		Random				1.2852			1.9677
logBBB	1		11	-1.8845	1.1777	0.3551	-4.1000	-0.3100	
	2		111	-0.0528	0.7861	0.0746	-2.0000	1.6400	
	Total		122	-0.2180	0.9767	0.0884	-4.1000	1.6400	
	Model	Fixed			0.8258	0.0748			
		Random				1.1745			1.6436

n = volumul eşantionului; m = media aritmetică; StDev = deviația standard;
StErr = eroarea standard; Min = valoarea minimă; Max = valoarea maximă;
BCVar = varianța între componente

Omogenitatea varianțelor este asigurată la nivelul clusterilor pentru toate variabilele cu excepția GAmIAaDI (statistica Levene = 24.790, df1 = 1, df2 = 120, $p = 2.17 \cdot 10^{-6}$).

Rezultatele testului ANOVA sunt prezentate în Tabelul 20. De remarcat distribuția mediile variabilelor în interiorul clusterilor (Figura 7). Aşa cum rezultă din Tabelul 20 nu există nici un descriptor MDFV fără contribuție semnificativă în clasificare.

Tabelul 20. Testul ANOVA: clasificare în funcție de valorile proprietății şi descriptorilor MDFV

Variabila	Clusteri	SS	df	MS	F	p
TLgFAIDI	Între	0.009	1	0.009	10.452	0.0016
	În	0.105	120	0.001		
	Total	0.114	121			
GAmIAaDI	Între	0.000	1	0.000	4.587	0.0342

	În	0.000	120	0.000		
	Total	0.000	121			
TAgFIADL	Între	34.311	1	34.311	65.770	$4.93 \cdot 10^{-13}$
	În	62.601	120	0.522		
	Total	96.912	121			
TAgPIADL	Între	40.229	1	40.229	47.724	$2.52 \cdot 10^{-10}$
	În	101.155	120	0.843		
	Total	141.384	121			
logBBB	Între	33.581	1	33.581	49.237	$1.45 \cdot 10^{-10}$
	În	81.842	120	0.682		
	Total	115.423	121			

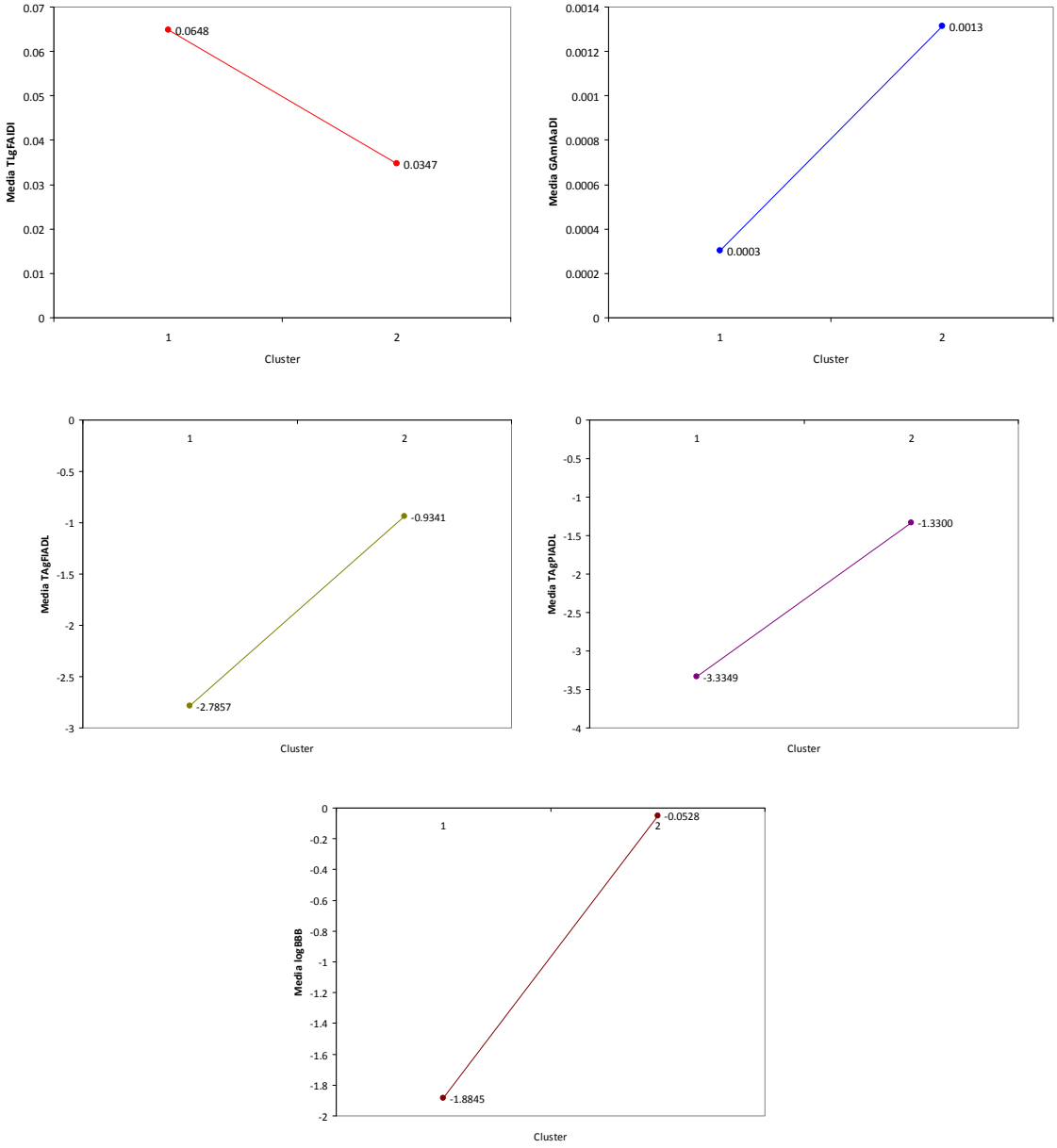


Figura 7. Contribuții medii în interiorul clusterilor

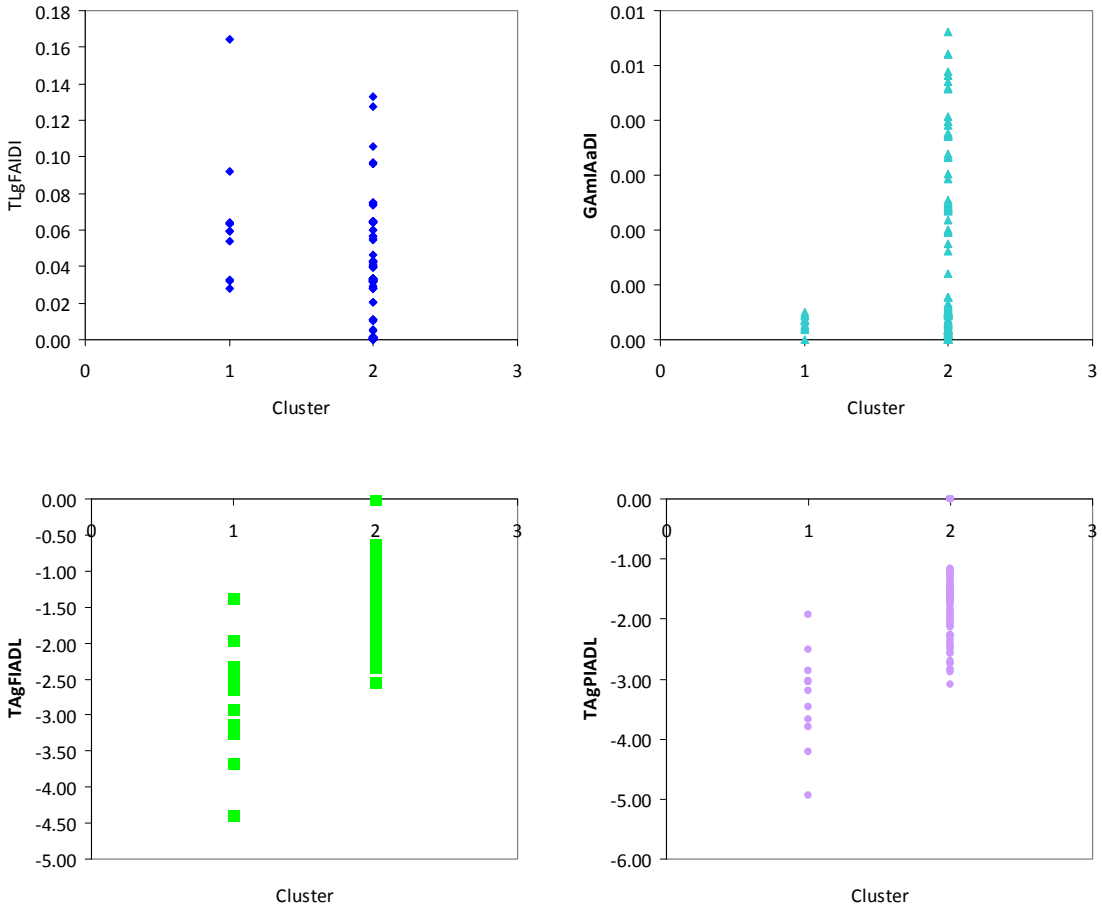
Aplicarea testului Welch de comparare a mediilor a pus în evidență următoarele:

- Diferență semnificativă statistic în ceea ce privește mediile în clusteri pentru descriptorul TLgFAIDI (Statistica Welch = 6.616, df1 = 1, df2 = 11.165, p = 0.026)
- Diferență semnificativă statistic în ceea ce privește mediile în clusteri pentru descriptorul

GAmIAaDI (Statistica Welch = 43.091, df1 = 1, df2 = 119.930, p = $1.40 \cdot 10^{-9}$)

- Diferență semnificativă statistic în ceea ce privește mediile în clusteri pentru descriptorul TAgFIADL (Statistica Welch = 51.722, df1 = 1, df2 = 11.531, p = $1.37 \cdot 10^{-5}$)
- Diferență semnificativă statistic în ceea ce privește mediile în clusteri pentru descriptorul TAgPIADL (Statistica Welch = 57.895, df1 = 1, df2 = 12.644, p = $4.56 \cdot 10^{-6}$)
- Diferență semnificativă statistic în ceea ce privește mediile în clusteri pentru logBBB (Statistica Welch = 25.485, df1 = 1, df2 = 10.901, p = $3.84 \cdot 10^{-4}$).

Distribuția valorilor în cadrul claselor pentru variabilele cu contribuție semnificativă statistic la clasificare sunt redate în Figura 8.



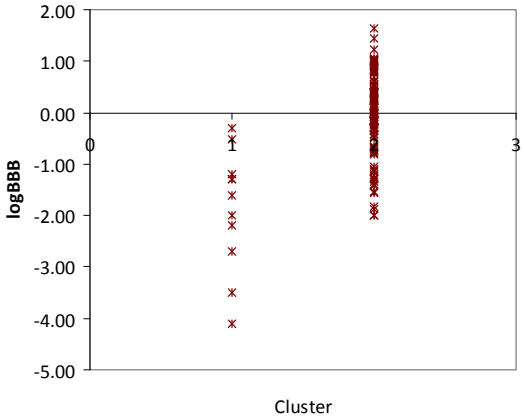


Figura 8. Distribuția valorilor variabilelor cu contribuție semnificativă statistic în clasificare

Următoarele concluzii se pot desprinde pe baza analizei de clusterizare realizată pe compuşii organici cu proprietatea de traversare a barierei hemato-encefalice:

- Analiza ierarhică de cluterizare a permis identificarea numărului optim de clusteri: în ceea ce priveşte logBBB a compuşilor organici investigați clasificarea optimă se face prin utilizarea a 2 sau a 4 clase.
- Utilizarea metodei k-means (ştiut fiind că numărul optim de clusteri este egal cu 2/4) clasifică diferit compuşii pe baza valorilor logBBB.
- Atât metode ierarhică de clasificare cât şi metoda k-medii s-au dovedit a fi semnificative statistic la un prag de semnificație de 5%.
- Analiza de clasificare a compuşilor pe baza valorilor proprietății măsurate şi a descriptorilor moleculari ca şi variabile a identificat un număr optim de 2 clase.
- Analiza de clasificare a compuşilor pe baza valorilor proprietății măsurate şi a descriptorilor moleculari atunci când se investighează moleculele a evidenția un model semnificativ statistic în care fiecare variabilă s-a dovedit a avea o contribuție semnificativă statistic în clasificare.

4.1.1.3. Derivați de sulfonamide - inhibitori ai anhidrazei carbonice II & Taxoizi – inhibiția creşterii celulare

Sulfonamide – inhibitori ai anhidrazei carbonice

Analiza de clasificare pentru s-a realizat pe baza datelor prezentate în Tabelul 21.

Sumarizarea rezultatele obținute în investigarea proprietății de interes în termeni de modalitate de aglomerare în clusteri sunt redate în Tabelul 22.

Tabelul 21. Date experimentale: sulfonamine – inhibitori ai anhidtrazei carbonice

Mol	logKI	TLhFPFdR	GMpFFIdI	TEmFIIDI
s001	1.079	57020	0.004158	2.1796
s002	0	27029	0.010253	4.093
s003	0.579	30290	0.014911	4.608
s004	0.255	25882	0.019949	6.086
s005	0.204	26191	0.012819	4.423
s006	0.278	28274	0.014106	4.7
s007	2.217	83760	0.02023	5.193
s008	2.369	82130	0.027891	6.856
s009	2.238	104750	0.017316	5.1
s010	2.411	103650	0.026936	7.04
s011	1.939	78850	0.016022	4.586
s012	2.423	92850	0.020031	5.14
s013	2.017	92850	0.018626	5.14
s014	1.886	92850	0.017551	5.14
s015	1.146	29532	0.011013	3.0836
s016	0.903	46260	0.010377	3.682
s017	1.579	122670	0.006149	3.774
s018	0.954	70180	0.012339	4.606

Tabelul 22. Sumarizarea coeficienților de aglomerare în analiza de clusterizare ierhică pentru suflbonamide

Nr clusteri	CoefAglomLast	CoefAglPrev	Dif
2	8.5365	3.3920	5.1445
3	3.3920	2.0467	1.3453
4	2.0467	1.2821	0.7647
5	1.2821	1.0105	0.2716
6	1.0105	0.7561	0.2544
7	0.7561	0.5686	0.1875

CoefAglUltim = coeficientul de aglomerare cu valoarea mare pentru numărul de clusteri de interes;
CoefAglPrevc= coeficientul de aglomerare anterior;
Dif = diferența dintre ultim și anterior;

Dendrograma asociată analizei este prezentată în Figura 1.

Un punct clar de demarcare în ceea ce privește diferența este la nivelul 1.3453 (diferență de ordin de mărime) → analiza poate să fie reluată pentru un număr fix de 2 clusteri. În urma anlizei s-a obținut apartenența fiecărui compus la un cluster după cum urmează:

- Cluster 1 (media per cluster egală cu 2.120): 9 compuși (s007; s008; s009; s010; s011; s012; s013; s014 și s017)
- Cluster 2 (media per cluster egală cu 0.600): 9 compuși (restul compușilor nespecificați anterior).

Parametrii statisticii descriptive pentru cei doi clusteri, modelul cu efecte fixe și respectiv random sunt prezantați în Tabelul 23. Figura 10 prezintă distribuția valorilor logKI per cluster, respectiv distribuția mediei per clasă. Distribuția normală a valorilor logKI nu a putut fi respinsă pentru nici unul din clusteri la un prag de semnificație de 5%.

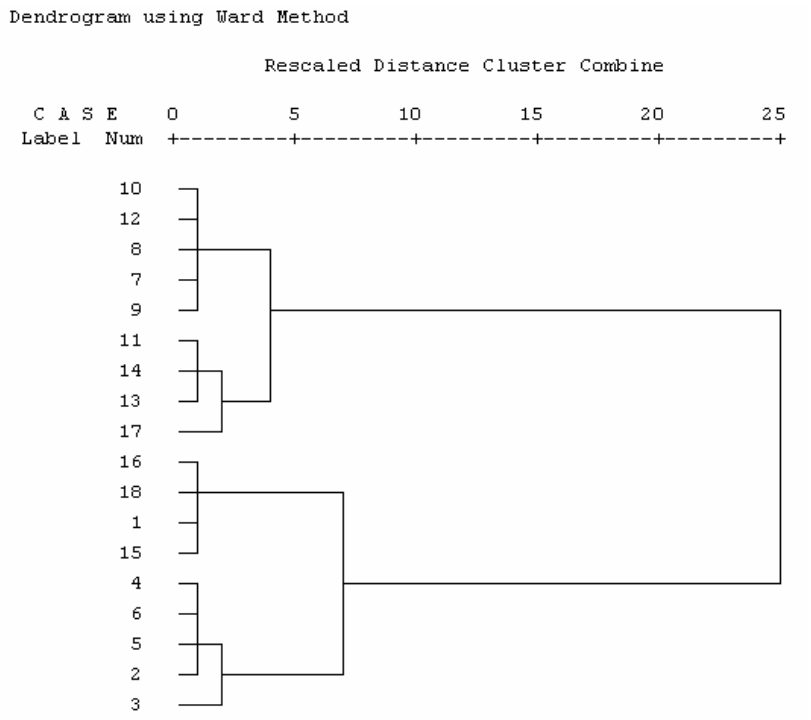


Figura 9. Sulfoamine: dendrograma

Tabelul 23. Parametrii statistici asociaţi clusterilor: modelul cu efecte fixe şi random pentru sulfonamide

Cluster	Effect	m	m	StDev	StErr	Min	Max	BCVar
1		9	2.1199	0.2856	0.0952	1.5790	2.4230	
2		9	0.5998	0.4308	0.1436	0.0000	1.1460	
Total		18	1.3598	0.8587	0.2024	0.0000	2.4230	
Model	Fixed			0.3655	0.0861			
	Random				0.7601			1.14053

n = volumul eşantionului; m = media aritmetică; StDev = deviaţia standard; StErr = eroarea standard; Min = valoarea minimă; Max = valoarea maximă; Media = media artimetică; BCVar = between component variance

Variaţele în cei doi clusteri s-au dovedit a fi omogene (Levene statistic = 3.642, df1 = 1, df2 = 16, p = 0.0744). Rezultatele obţinute în urma aplicării testului ANOVA sunt redate în Tabelul 24.

Tabelul 24. ANOVA: proprietarea sulfonaminelor investigate

	SS	df	MS	F	p
Între clusteri	10.398	1	10.398	77.843	$1.52 \cdot 10^{-7}$
În clusteri	2.137	16	0.134		
Total	12.536	17			

SS = suma pătratelor erorilor; df = grade de libertate; MS = media pătratelor erorilor; F = statistica Fisher; p = semnificaţia statisticii Fisher

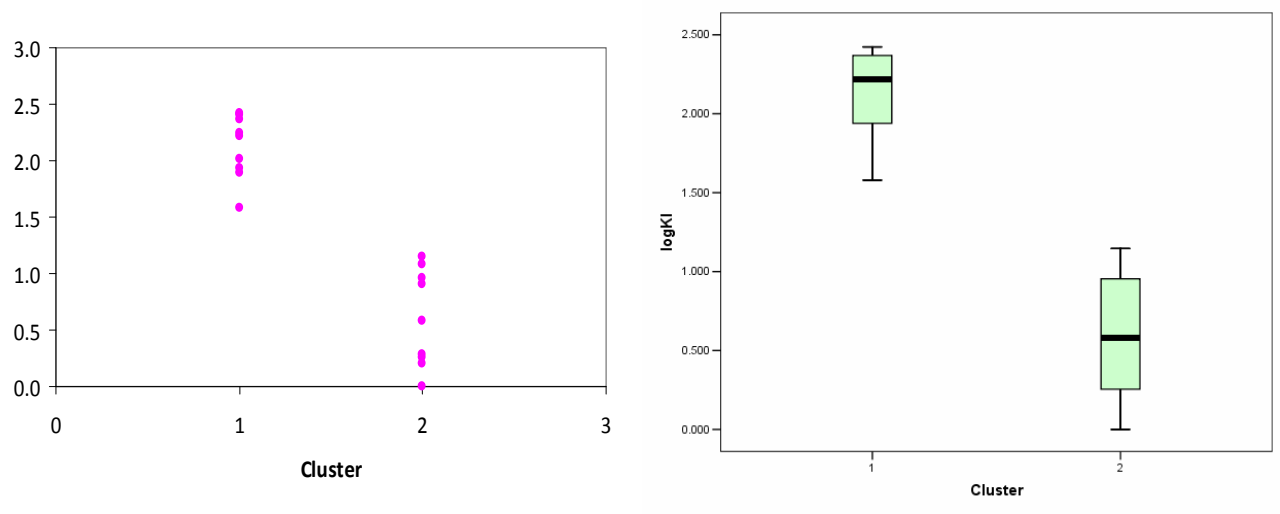


Figura 10. Sulfoamine: distribuția valorilor, respectiv a mediei

Aplicarea testului Welch de comparare a mediilor a pus în evidență o diferență semnificativă statistic între mediile logKI ale celor doi clusteri (Statistica Welch = 77.843, df1 = 1, df2 = 13.894, p = $4.56 \cdot 10^{-7}$).

Analiza de clusterizare s-a aplicat în continuare pentru proprietate şi respectiv cei trei descriptori MDFV ulterior transformării tuturor variabilelor în intervalul [0, 1].

Sumarizarea rezultatele obținute în investigarea proprietății de interes în termeni de modalitate de aglomerare în clusteri sunt redate în Tabelul 25. Dendrograma asociată analizei de clusterizare ierarhică este redată în Figura 11.

Tabelul 25. Sumarizarea rezultatelor: coeficienți de aglomerarea prop + MDFV sulfonamide

Nr clusteri	CoefAglomLast	CoefAglPrev	Dif
2	6.6061	3.8359	2.7703
3	3.8359	3.1138	0.7221
4	3.1138	2.3938	0.7200
5	2.3938	1.8595	0.5343
6	1.8595	1.5519	0.3076
7	1.5519	1.2687	0.2832

CoefAglUltim = coeficientul de aglomerare cu valoarea mare pentru numărul de clusteri de interes;
CoefAglPrevc= coeficientul de aglomerare anterior;
Dif = diferența dintre ultim şi anterior;

Rezulatele prezentate în Tabelul 25 au indicat reluarea analizei de clusterizare cu un număr de 2 clusteri.

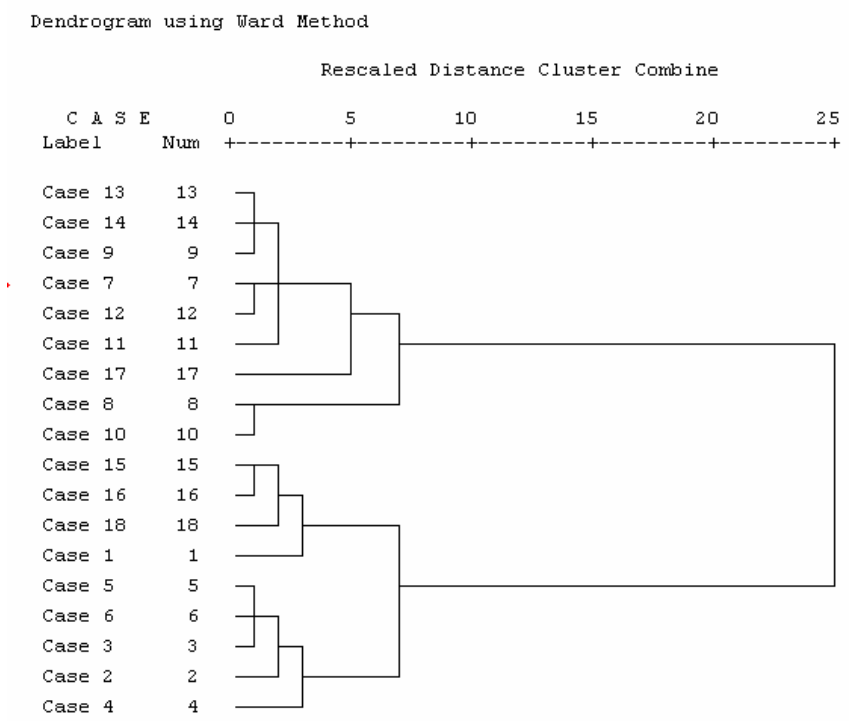


Figura 11. Sulfonamine: dendrograma în analiza ierarhică de clusterizare (prop & descriptori MDFV)

Distribuția compușilor în funcție de utilizarea unui număr fix de 2 clusteri a fost următoarea:

- Cluster 1: 9 compuşi (s007; s008; s009; s010; s011; s012; s013; s014 şi s017)
- Cluster 2: 9 compuşi (restul compușilor nespecificați ca aparținând clusterului 1).

Testul ANOVA a fost aplicat pentru a identifica contribuția în clasificare a fiecărei variabile utilizate iar rezultatele sunt prezentate în Tabelul 26.

Tabelul 26. Rezultate statistică descriptivă: clasificare pe baza proprietății şi a valorilor descriptorilor MDFV

Variabila	Clustrer	Efect	n	Mean	StDev	StErr	Minimum	Maximum	BCVar
logKI	1		9	2.1199	0.2856	0.0952	1.5790	2.4230	
	2		9	0.5998	0.4308	0.1436	0.0000	1.1460	
	Total		18	1.3598	0.8587	0.2024	0.0000	2.4230	
	Model	Fix			0.3655	0.0861			
		Random				0.7601			1.1405
TLhFPFdR	1		9	94929	13703	4568	78850	122670	
	2		9	37851	16193	5398	25882	70180	
	Total		18	66390	32774	7725	25882	122670	
	Model	Fix			15000	3535			
		Random				28539			1.60E+09
GMpFFIdI	1		9	0.0190	0.0064	0.0021	0.0061	0.0279	
	2		9	0.0122	0.0043	0.0014	0.0042	0.0199	
	Total		18	0.0156	0.0063	0.0015	0.0042	0.0279	
	Model	Fix			0.0054	0.0013			
		Random				0.0034			0.0000
TEmFIIDI	1		9	5.3299	1.0251	0.3417	3.7740	7.0400	
	2		9	4.1624	1.1059	0.3686	2.1796	6.0860	
	Total		18	4.7461	1.1962	0.2819	2.1796	7.0400	
	Model	Fix			1.0663	0.2513			
		Random				0.5838			0.5552

Omogenitatea variațelor este asigurată la nivelul clusterilor pentru toate variabilele cu excepția (df1 = 1, df2 = 16, statistica Levene – logKI = 3.642 (p = 0.074); statistica Levene – TLhFPFdR = 0.627 (p = 0.440); statistica Levene – GMpFFIdI = 0.587 (p = 0.455); statistica Levene – TEmFIIDI = 0.065 (p = 0.803)).

Rezultatele testului ANOVA sunt prezentate în Tabelul 27. De remarcat distribuția mediile variabilelor în interiorul clusterilor (Figura 12). Așa cum rezultă din Tabelul 27 nu există nici un descriptor MDFV fără contribuție semnificativă în clasificare.

Tabelul 27. Testul ANOVA: clasificare în funcție de valorile proprietății şi descriptorilor MDFV

Variabila	Cluster	SS	df	MS	F	p
logKI	Între	10.3983	1	10.3983	77.8434	$1.52 \cdot 10^{-7}$
	În	2.1373	16	0.1336		
	Total	12.5356	17			
TLhFPFdR	Între	$1.47 \cdot 10^{10}$	1	$1.47 \cdot 10^{10}$	65.1601	$4.93 \cdot 10^{-7}$
	În	$3.6 \cdot 10^{9}$	16	$2.25 \cdot 10^{8}$		
	Total	$1.83 \cdot 10^{10}$	17			
GMpFFIdI	Între	$2.06 \cdot 10^{-4}$	1	$2.06 \cdot 10^{-4}$	7.0226	0.0175
	În	$4.68 \cdot 10^{-4}$	16	$2.93 \cdot 10^{-5}$		
	Total	$6.74 \cdot 10^{-4}$	17			
TEmFIIDI	Între	6.1341	1	6.1341	5.3953	0.0337
	În	18.1908	16	1.1369		
	Total	24.3249	17			

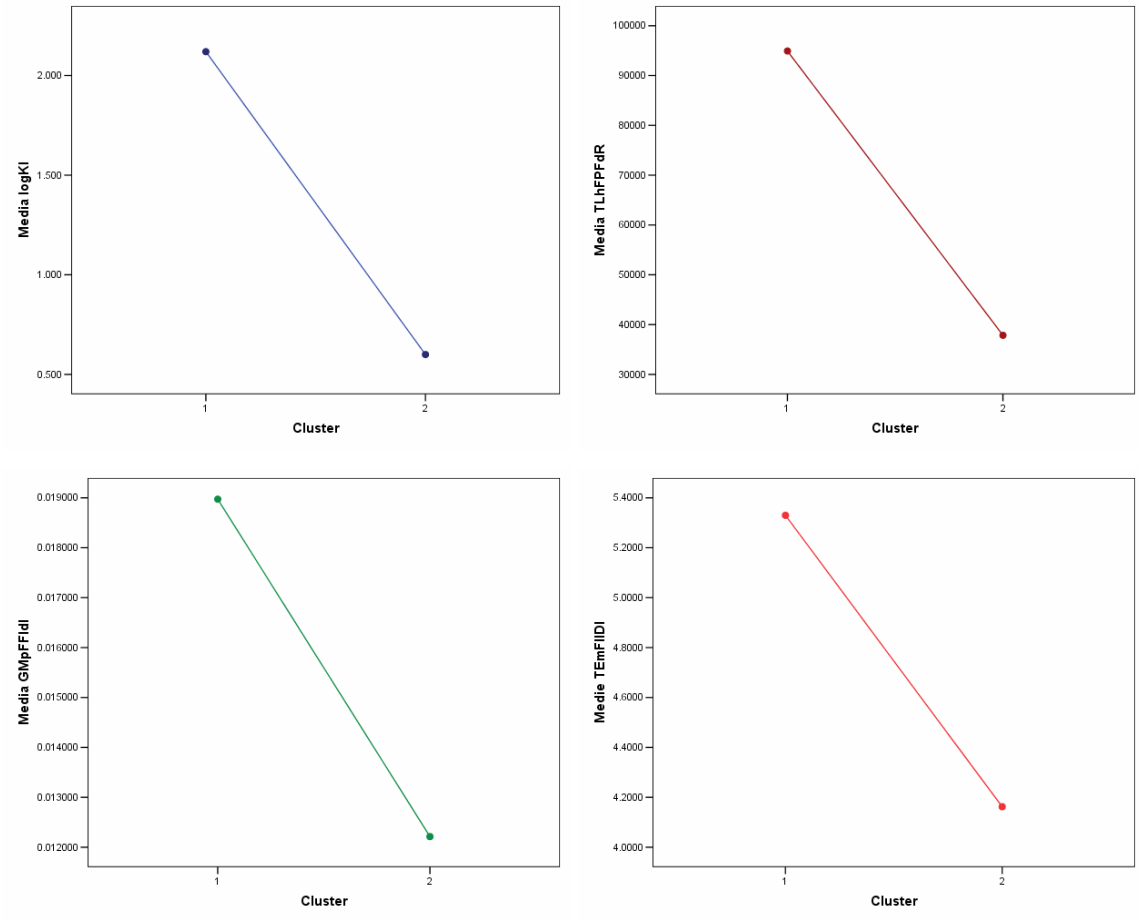


Figura 12. Contribuții medii în clusteri (prop & descriptori MDFV)

Aplicarea testului Welch de comparare a mediilor a pus în evidență următoarele diferențe semnificative statistic la un prag de semnificație de 5%:

- Mediile în clusteri pentru logKII (Statistica Welch = 77.843, df1 = 1, df2 = 13.894, p = $4.56 \cdot 10^{-7}$)
- Mediile în clusteri pentru descriptorul TLhFPFdR (Statistica Welch = 65.160, df1 = 1, df2 = 15.574, p = $5.95 \cdot 10^{-7}$)
- Mediile în clusteri pentru descriptorul GMpFFIdI (Statistica Welch = 7.023, df1 = 1, df2 = 13.959, p = 0.0191)
- Mediile în clusteri pentru descriptorul TEmFIIDI (Statistica Welch = 5.395, df1 = 1, df2 = 15.909, p = 0.0338).

Distribuția valorilor în cadrul claselor pentru variabilele cu contribuție semnificativă statistic la clasificare sunt redate în Figura 13.

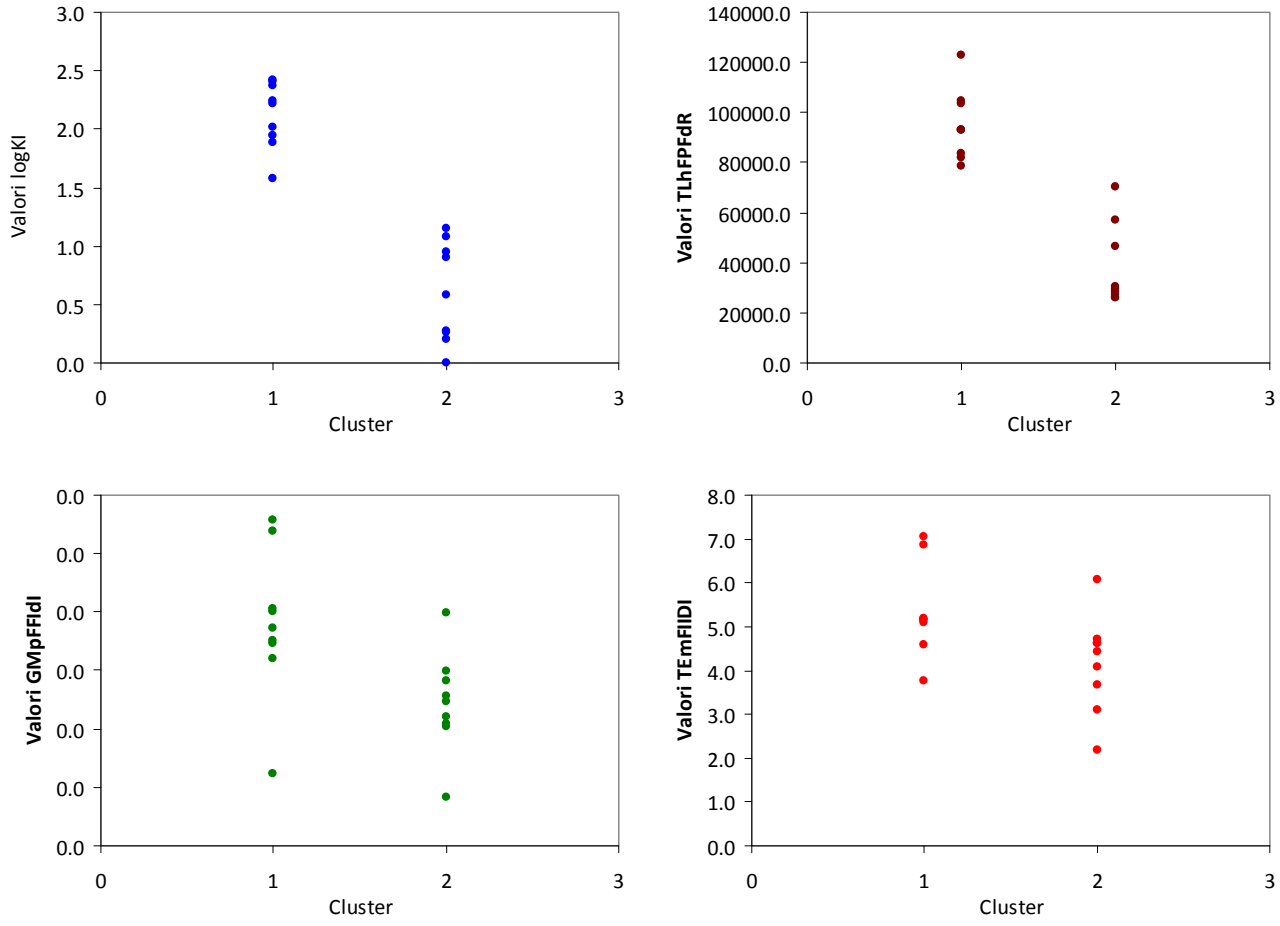


Figura 133. Distribuția valorilor variabilelor cu contribuție semnificativă statistic în clasificare (prop & descriptori MDFV)

Următoarele concluzii se pot desprinde pe baza analizei de clusterizare realizată pe compuşii organici cu proprietatea de traversare a barierei hemato-encefalice:

- Analiza ierarhică de cluterizare a permis identificarea numărului optim de clusteri: clasificarea optimă se face atât în ceea ce priveşte logKI cât şi în ceea ce priveşte logKI şi descriptorii MDFV

ai modelului cu 2 clusteri

- Utilizarea metodei k-means (ştiut fiind că numărul optim de clusteri este egal cu 2) clasifică identic compuşii indiferent dacă clasificarea se realizează doar pe baza valorii logKI sau pe baza valorilor logKI şi a descriptorilor din model.
- Atât metode ierarhică de clasificare cât şi metoda k-medii s-au dovedit a fi semnificative statistic la un prag de semnificație de 5%.
- Toate variabilele (logKI şi descriptori MDFV) s-au dovedit a avea o contribuție semnificativă statistic în clasificare.
- Clasificarea în cazul sulfonaminelor cu activitate inhibitorie a anhidrazei carbonice este indicată a se realiza utilizând doar valorile logKI deoarece clasificarea este identică în cazul utilizării valorilor logKI sau a valorilor logKI & a descriptorilorMDFV.

Analiza de clasificare a compuşilor pe baza valorilor proprietății măsurate şi a descriptorilor moleculari atunci când se investighează moleculele a evidenția un model semnificativ statistic în care fiecare variabilă s-a dovedit a avea o contribuție semnificativă statistic în clasificare.

Taxoizi – inhibitori ai creşterii celulare

Analiza de clasificare pentru s-a realizat pe baza datelor prezentate în Tabelul 28 [43].

Sumarizarea rezultatele obținute în investigarea proprietății de interes în termeni de modalitate de aglomerare în clusteri sunt redate în Tabelul 29.

Tabelul 28. Date experimentale: taxoizi – inhibitori ai creşterii celulare

Mol	logIC50	TAcAIiDR	TQKCPfdL	TMiIPpdL
tax001	1.66	71930000.00	8.05	3.30
tax002	1.37	71930000.00	8.09	3.30
tax003	0.77	71930000.00	8.12	2.48
tax004	1.18	71930000.00	8.02	2.48
tax005	1.09	71930000.00	8.16	2.48
tax007	1.39	71930000.00	7.98	2.48
tax008	1.74	71930000.00	8.16	3.30
tax009	0.77	71930000.00	8.19	2.48
tax010	-1.20	19881000.00	7.28	2.48
tax011	-1.28	26462000.00	7.43	2.48
tax012	-1.00	17061000.00	6.99	2.48
tax013	-1.54	22708000.00	7.38	2.48
tax014	-1.32	19881000.00	7.28	2.48
tax015	-1.60	14493000.00	7.38	2.48
tax016	-0.34	19881000.00	6.93	2.48
tax017	-0.64	34350000.00	7.61	2.48
tax018	-2.00	19881000.00	7.57	2.48
tax019	-1.78	19881000.00	7.17	1.10
tax020	-0.62	26462000.00	7.38	2.48
tax021	-1.20	14493000.00	7.17	2.48

[43] Bolboacă SD, Jäntschi L. Structure-activity relationships of taxoids: a molecular descriptors family approach. Archives of Medical Science 2008;4(1):7-15.

tax022	-0.48	26462000.00	6.87	2.48
tax023	-1.36	14493000.00	7.22	2.48
tax024	-2.00	19881000.00	7.66	2.48
tax025	-1.90	19881000.00	7.43	2.48
tax026	-1.91	14493000.00	7.17	2.48
tax027	-1.18	19881000.00	7.28	2.48
tax028	-0.59	34350000.00	7.66	2.48
tax029	-1.85	26462000.00	7.90	2.48
tax030	-1.91	26462000.00	7.66	2.48
tax031	-1.57	19881000.00	7.38	2.48
tax032	-2.00	19881000.00	7.48	2.48
tax033	-0.64	26462000.00	7.22	2.48
tax034	-2.00	26462000.00	7.78	2.48
tax035	-1.32	19881000.00	7.38	2.48

Tabelul 29. Sumarizarea coeficienților de aglomerare în analiza de clusterizare ierhică pentru taxoizi

Nr clusteri	CoefAglomLast	CoefAglPrev	Dif
2	22.2224	9.3170	12.9053
3	9.3170	5.6183	3.6988
4	5.6183	3.3163	2.3020
5	3.3163	2.5600	0.7563
6	2.5600	1.9775	0.5825
7	1.9775	1.5050	0.4725

CoefAglUltim = coeficientul de aglomerare cu valoarea mare pentru numărul de clusteri de interes;
CoefAglPrevc= coeficientul de aglomerare anterior;
Dif = diferența dintre ultim și anterior;

Dendrograma asociată analizei este prezentată în Figura 1.

Un punct clar de demarcare în ceea ce privește diferența este la nivelul 3.6988 (diferență de ordin de mărime) → analiza poate să fie reluată pentru un număr fix de 2 clusteri. În urma analizei s-a obținut apartenența fiecărui compus la un cluster după cum urmează:

- Cluster 1 (media per cluster egală cu 1.25): 8 compuși (tax001; tax002; tax003; tax004; tax005; tax007; tax008 și tax009)
- Cluster 2 (media per cluster egală cu -1.36): 26 compuși (restul compușilor nespecificați anterior).

Parametrii statisticii descriptive pentru cei doi clusteri, modelul cu efecte fixe și respectiv random sunt prezantați în Tabelul 30. Figura 15 prezintă distribuția valorilor $\log IC_{50}$ per cluster, respectiv distribuția mediei per clasă. Distribuția normală a valorilor $\log IC_{50}$ nu a putut fi respinsă pentru nici unul din clusteri la un prag de semnificație de 5%.

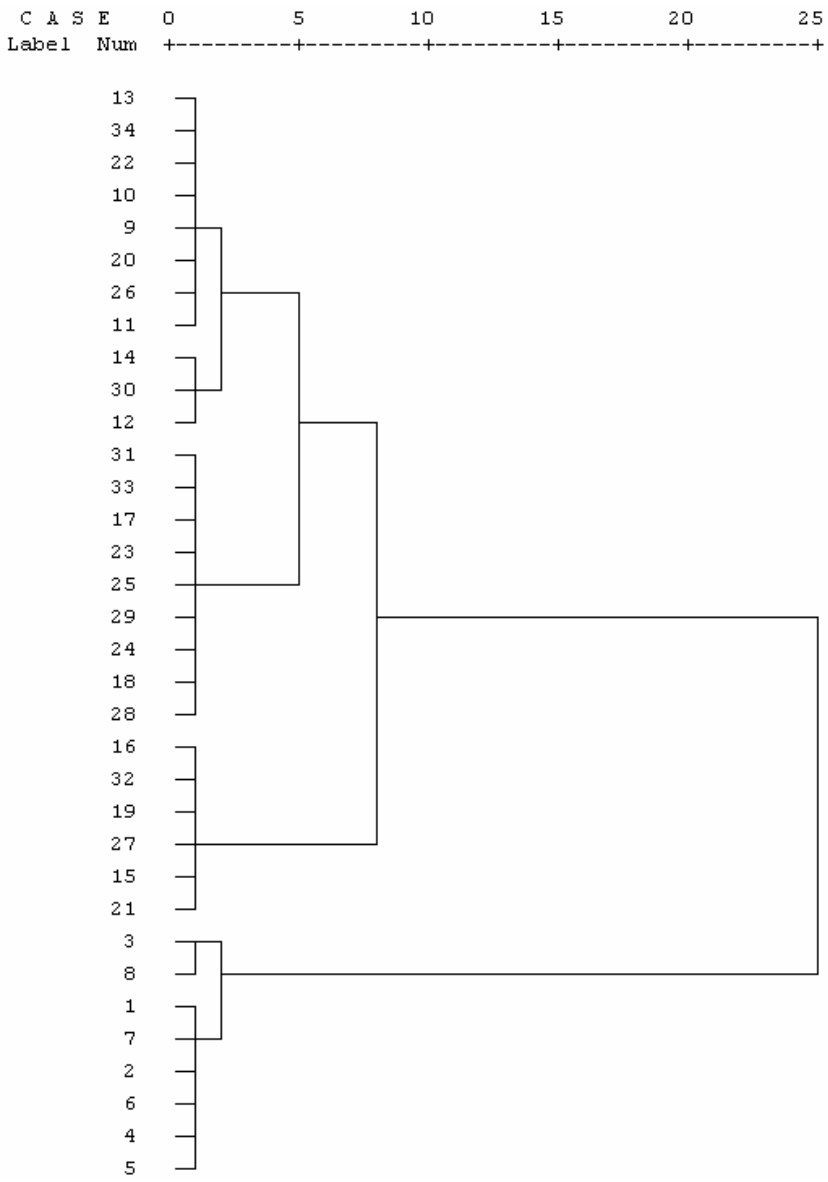


Figura 14. Taxoizi: dendrograma – analiza ierarhică de clasificare

Tabelul 30. Parametrii statistici asociați clusterilor: modelul cu efecte fixe şi random pentru taxoizi

Cluster	Effect	n	m	StDev	StErr	Min	Max	BCVar
1		8	1.2463	0.3652	0.1291	0.77	1.74	
2		26	-1.3550	0.5404	0.1060	-2.00	-0.34	
Total		34	-0.7429	1.2263	0.2103	-2.00	1.74	
Model	Fix			0.5072	0.0870			
	Random				1.4696			3.3622

n = volumul eşantionului; m = media aritmetică; StDev = deviația standard; StErr = eroarea standard; Min = valoarea minimă; Max = valoarea maximă; Media = media artimetică; BCVar = between component variance

Variațele în cei doi clusteri s-au dovedit a fi omogene (Levene statistic = 1.938, df1 = 1, df2 = 32, p = 0.1735). Rezultatele obținute în urma aplicării testului ANOVA sunt redate în Tabelul 31.

Tabelul 31. ANOVA: proprietarea taxoizilor investigați

	SS	df	MS	F	p
Între clusteri	41.40	1	41.40	160.89	$5.02 \cdot 10^{-14}$

În clusteri	8.23	32	0.26		
Total	49.63	33			

SS = suma pătratelor erorilor; df = grade de libertate;
MS = media pătratelor erorilor; F = statistica Fisher;
p = semnificația statisticii Fisher

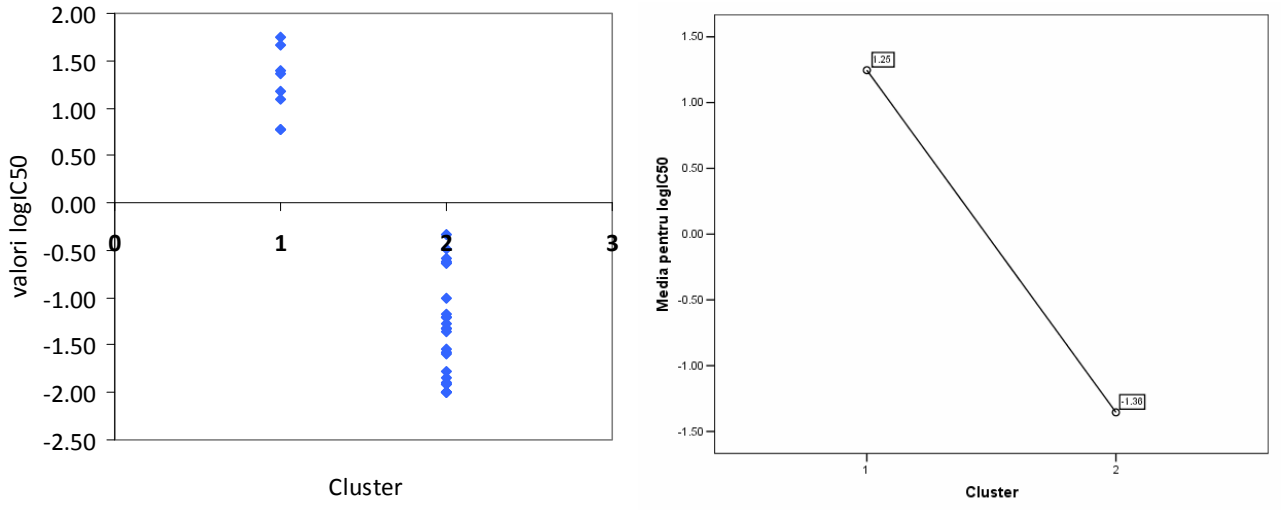


Figura 15. Sulfoamine: distribuția valorilor, respectiv a mediei

Aplicarea testului Welch de comparare a mediilor a pus în evidență o diferență semnificativă statistic între mediile logKI ale celor doi clusteri (Statistica Welch = 242.54, df1 = 1, df2 = 17.399, p = $1.18 \cdot 10^{-11}$).

Analiza de clusterizare s-a aplicat în continuare pentru proprietate şi respectiv cei trei descriptori MDFV ulterior transformării tuturor variabilelor în intervalul [0, 1].

Sumarizarea rezultatele obținute în investigarea proprietății de interes în termeni de modalitate de aglomerare în clusteri sunt redate în Tabelul 32. Dendrograma asociată analizei de clusterizare ierarhică este redată în Figura 16.

Tabelul 32. Sumarizarea rezultatelor: coeficienți de aglomerarea prop + MDFV taxoizi

Nr clusteri	CoefAglomLast	CoefAglPrev	Dif
2	11.5254	5.4248	6.1006
3	5.4248	4.4371	0.9877
4	4.4371	3.8170	0.6201
5	3.8170	3.2028	0.6142
6	3.2028	2.6451	0.5578
7	2.6451	2.1404	0.5047

CoefAglUltim = coeficientul de aglomerare cu valoarea mare pentru numărul de clusteri de interes;
CoefAglPrevc= coeficientul de aglomerare anterior;
Dif = diferența dintre ultim și anterior;

Rezulatele prezentate în Tabelul 25 au indicat reluarea analizei de clusterizare cu un număr de 2 clusteri.

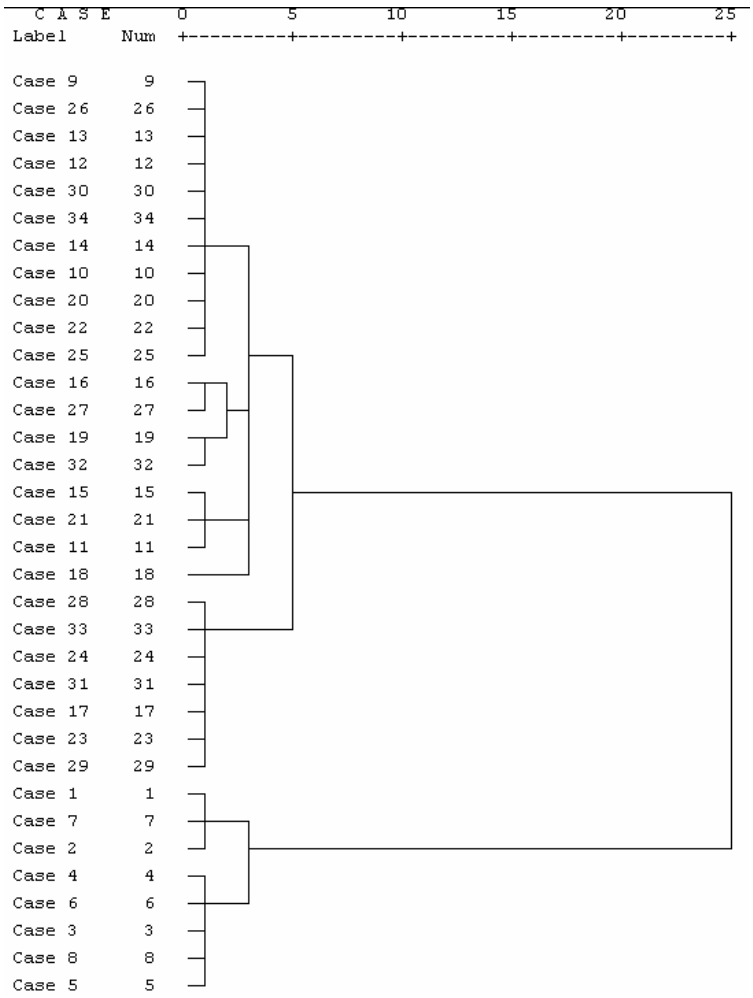


Figura 16. Taxoizi: dendrograma în analiza ierarhică de clusterizare (prop & descriptori MDFV)

Distribuția compușilor în funcție de utilizarea unui număr fix de 2 clusteri a fost următoarea:

- Cluster 1: 8 compuși (tax001; tax002; tax003; tax004; tax005; tax007; tax008 and tax009)
- Cluster 2: 24 compuși (restul compușilor nespecificați ca aparținând clusterului 1).

Testul ANOVA a fost aplicat pentru a identifica diferențe semificative statistic a variabilelor în clusteri iar rezultatele sunt prezentate în Tabelul 33. Omogenitatea variațelor este asigurată la nivelul clusterilor doar pentru $logIC_{50}$ (df1 = 1, df2 = 32, statistica Levene = 1.938 (p = 0.174). Următoarele rezultate au fost obținute pentru descriptorii MDFV:

- TAcAIiDR: statistica Levene = 15.869 (p = 0.000367)
- TQKCPfdL: statistica Levene = 5.297 (p = 0.028018)
- TMiIPpdL: statistica Levene = 9.138 (p = 0.004899)

Tabelul 33. Rezultate statistică descriptivă: clasificare pe baza proprietății și a valorilor descriptorilor MDFV

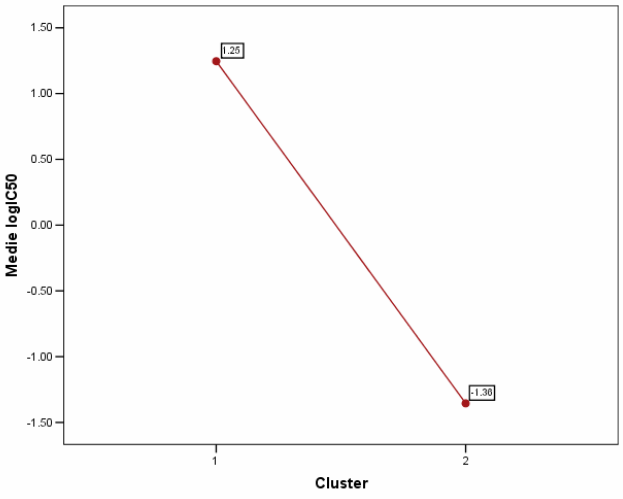
Variabila	Cluster	Efect	n	m	StDev	StErr	Min	Max	BCVar
logIC50	1		8	1.2463	0.3652	0.1291	0.77	1.74	
	2		26	-1.3550	0.5404	0.1060	-2	-0.34	
	Total		34	-0.7429	1.2263	0.2103	-2	1.74	
	Model	Fixe			0.5072	0.0870			

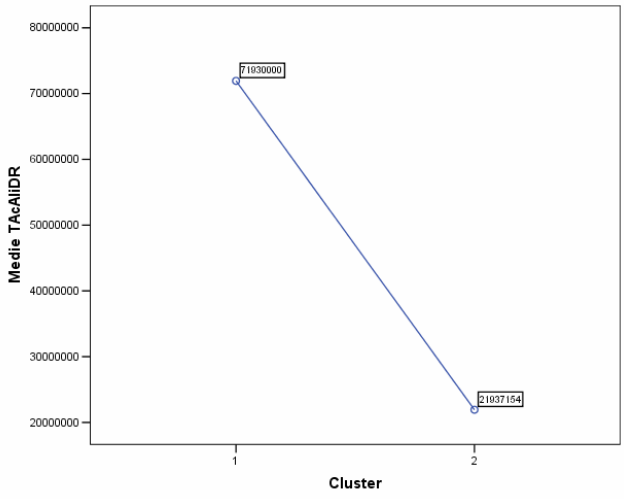
		Random				1.4696			3.3622
TAcAIiDR	1		8	$7.19 \cdot 10^{7}$	0.00	0.00	$7.19 \cdot 10^{7}$	$7.19 \cdot 10^{7}$	
	2		26	$2.19 \cdot 10^{7}$	$5.46 \cdot 10^{6}$	$1.07 \cdot 10^{6}$	$1.45 \cdot 10^{7}$	$3.44 \cdot 10^{7}$	
	Total		34	$3.37 \cdot 10^{7}$	$2.20 \cdot 10^{7}$	$3.78 \cdot 10^{6}$	$1.45 \cdot 10^{7}$	$7.19 \cdot 10^{7}$	
	Model	Fixe			$4.82 \cdot 10^{6}$	$8.27 \cdot 10^{5}$			
		Random				$2.83 \cdot 10^{7}$			$1.25 \cdot 10^{15}$
TQKCPfdL	1		8	8.0938	0.0745	0.0263	7.9780	8.1890	
	2		26	7.3700	0.2529	0.0496	6.8680	7.9020	
	Total		34	7.5403	0.3831	0.0657	6.8680	8.1890	
	Model	Fixe			0.2262	0.0388			
		Random				0.4080			0.2578
TMiIPpdL	1		8	2.7891	0.4198	0.1484	2.4849	3.2960	
	2		26	2.4316	0.2719	0.0533	1.0986	2.4849	
	Total		34	2.5157	0.3422	0.0587	1.0986	3.2960	
	Model	Fixe			0.3103	0.0532			
		Random				0.1967			0.0560

Rezultatele testului ANOVA sunt prezentate în Tabelul 27. De remarcat distribuția mediile variabilelor în interiorul clusterilor (Figura 17). Așa cum rezultă din Tabelul 27, mediile tuturor descriptorilor sunt semnificativ diferite între clusteri.

Tabelul 34. Testul ANOVA: clasificare în funcție de valorile proprietății şi descriptorilor MDFV

Variabila	Cluster	SS	df	MS	F	p
logIC50	Între	41.40	1	41.40	160.89	$5.02 \cdot 10^{-14}$
	În	8.23	32	0.26		
	Total	49.63	33			
TAcAIiDR	Între	$1.53 \cdot 10^{16}$	1	$1.53 \cdot 10^{16}$	657.61	$6.61 \cdot 10^{-23}$
	În	$7.44 \cdot 10^{14}$	32	$2.33 \cdot 10^{13}$		
	Total	$1.60 \cdot 10^{16}$	33			
TQKCPfdL	Între	3.20	1	3.20	62.62	$4.97 \cdot 10^{-9}$
	În	1.64	32	0.05		
	Total	4.84	33			
TMiIPpdL	Între	0.78	1	0.78	8.12	0.0076
	În	3.08	32	0.10		
	Total	3.86	33			





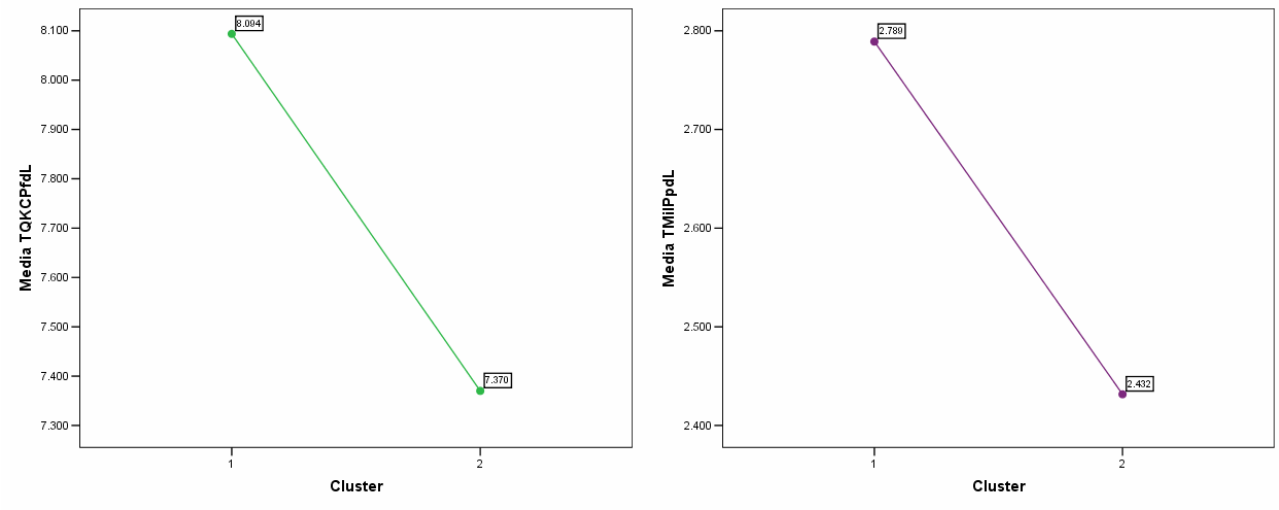
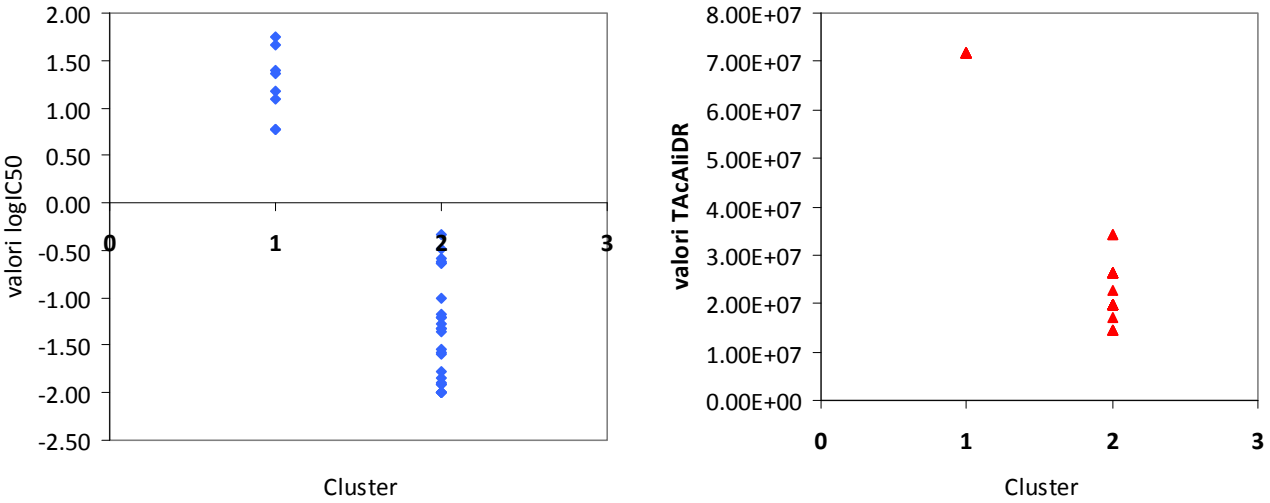


Figura 17. Taxoizi: Contribuții medii în clusteri (prop & descriptori MDFV)

Aplicarea testului Welch de comparare a mediilor a pus în evidență următoarele diferențe semnificative statistic la un prag de semnificație de 5%:

- Mediile în clusteri pentru $\log IC_{50}$ (Statistica Welch = 242.543, df1 = 1, df2 = 17.399, p = $1.18 \cdot 10^{-7}$)
- Mediile în clusteri pentru descriptorul TQKCPfdL (Statistica Welch = 166.153, df1 = 1, df2 = 32.000, p = $3.25 \cdot 10^{-14}$)
- Mediile în clusteri pentru descriptorul TMiIPpdL (Statistica Welch = 5.138, df1 = 1, df2 = 8.882, p = 0.049995)

Distribuția valorilor în cadrul claselor pentru variabilele este redată în Figura 18.



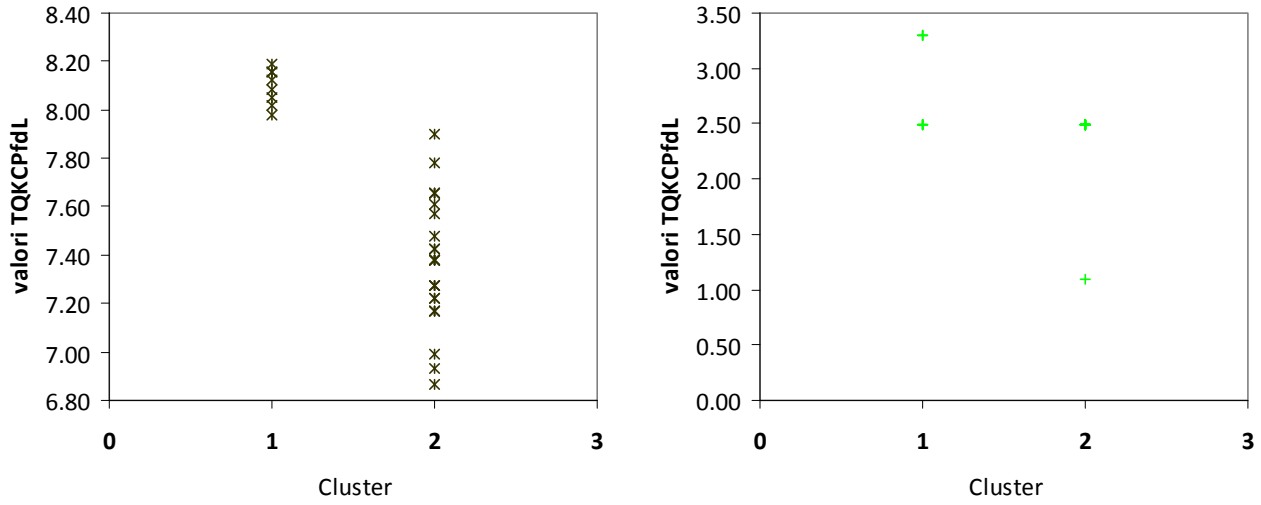


Figura 18. Distribuția valorilor variabilelor cu contribuție semnificativă statistic în clasificare (prop & descriptori MDFV)

Următoarele concluzii se pot desprinde pe baza analizei de clusterizare a taxoizilor:

- Analiza ierarhică de cluterizare a permis identificarea numărului optim de clusteri: clasificarea optimă se face atât în ceea ce privește $\log IC_{50}$ cât și în ceea ce privește $\log IC_{50}$ și descriptorii MDFV cu 2 clusteri
- Utilizarea metodei k-means (știut fiind că numărul optim de clusteri este egal cu 2) clasifică identic compușii indiferent dacă clasificarea se realizează doar pe baza valorii $\log IC_{50}$ sau pe baza valorilor $\log IC_{50}$ și a descriptorilor din model.
- Atât metode ierarhică de clasificare cât și metoda k-medii s-au dovedit a fi semnificative statistic la un prag de semnificație de 5%.
- Toate variabilele ($\log IC_{50}$ și descriptori MDFV) s-au dovedit a avea o contribuție semnificativă statistic în clasificare.
- Clasificarea în cazul taxoizilor cu activitate inhibitorie a anhidrazei carbonice este indicată a se realiza utilizând doar valorile $\log IC_{50}$ deoarece clasificarea este identică în cazul utilizării valorilor $\log IC_{50}$ sau a valorilor $\log IC_{50}$ & a descriptorilorMDFV. Mai mult 2 din descriptorii MDFV s-au dovedit a fi degenerați (au valori identice pentru mai mulți compuși → nu sunt caracterizici pentru caracterizarea $\log IC_{50}$). Modelul identificat pentru taxoizi nu este capabil să explice legătura de liniaritate dintre structura taxoizilor și $\log IC_{50}$ → este necesară căutarea unui nou model în care valorile descriptorilor

Analiza de clasificare a compușilor pe baza valorilor proprietății măsurate ($\log IC_{50}$) a permis clasificarea taxoizilor investigați. Modelul ce redă linearitatea dintre $\log IC_{50}$ și structura compușilor nu este un model valid din moment ce 2 din descriptorii MDFV au valori identice pentru mai multe molecule active.

4.1.1.4. Derivați de triphenilacrilonitrili – afinitate relativă de legare receptori de estrogen

Analiza de clasificare pentru s-a realizat pe baza datelor prezentate în Tabelul 35 [44].

Sumarizarea rezultatele obținute în investigarea proprietății de interes în termeni de modalitate de aglomerare în clusteri sunt redate în Tabelul 36.

Dendrograma asociată analizei este prezentată în Figura 1.

Un punct clar de demarcare în ceea ce privește diferența este la nivelul 0.9617 (diferență de ordin de mărime) → analiza poate să fie reluată pentru un număr fix de 4 clusteri.

Tabelul 35. Date experimentale: triphenilacrilonitrili – afinitate relativă de legare receptori de estrogen

Mol	logRBA	TASaAFDL	GLCACPdL	GMhaAiDR
triph001	-1.046	7.194	-1.6789	13358
triph002	1.556	7.130	0.6603	22774
triph003	0.342	7.270	0.7715	19946
triph004	0.519	7.211	-0.7159	23290
triph005	1.792	7.130	0.7279	24238
triph006	1.869	7.231	-0.8584	39450
triph007	0.785	7.286	0.6316	22890
triph008	2.220	7.304	1.8035	39350
triph009	1.447	7.130	0.7337	23111
triph010	0.398	7.130	-0.8521	21011
triph011	1.968	7.130	0.7519	20622
triph012	1.892	7.304	0.6882	38360
triph013	0.959	7.304	0.6702	29383
triph014	-0.180	7.304	0.7830	22956
triph015	1.230	7.130	-0.6848	24643
triph016	-0.444	7.332	-0.6490	25257
triph017	0.806	7.130	-0.6940	30176
triph018	-2.000	7.440	1.6930	1148.2
triph019	0.531	7.373	0.8650	30626
triph020	2.033	7.130	0.7765	17342
triph021	-0.398	7.543	0.8615	41710
triph022	-2.000	7.296	-2.0017	14537
triph023	-1.398	7.408	-1.0227	23340
triph024	-2.000	7.479	-2.3672	33110
triph025	-1.398	7.350	-0.8356	24907

Tabelul 36. Sumarizarea coeficienților de aglomerare în analiza de clusterizare ierhică pentru trifenilacrilonitrili

Nr clusteri	CoefAglomLast	CoefAglPrev	Dif
2	19.2074	10.1469	9.0606
3	10.1469	5.3291	4.8177
4	5.3291	3.6863	1.6428
5	3.6863	2.7247	0.9617
6	2.7247	1.9173	0.8073
7	1.9173	1.3673	0.5500

CoefAglUltim = coeficientul de aglomerare cu valoarea

[44] Bolboacă SD, Marta MM, Jäntschi L. Binding affinity of triphenyl acrylonitriles to estrogen receptors: quantitative structure-activity relationships. Folia Medica 2010;52(3):37-45.

mare pentru numărul de clusteri de interes;
CoefAglPrevc= coeficientul de aglomerare anterior;
Dif = diferența dintre ultim și anterior;

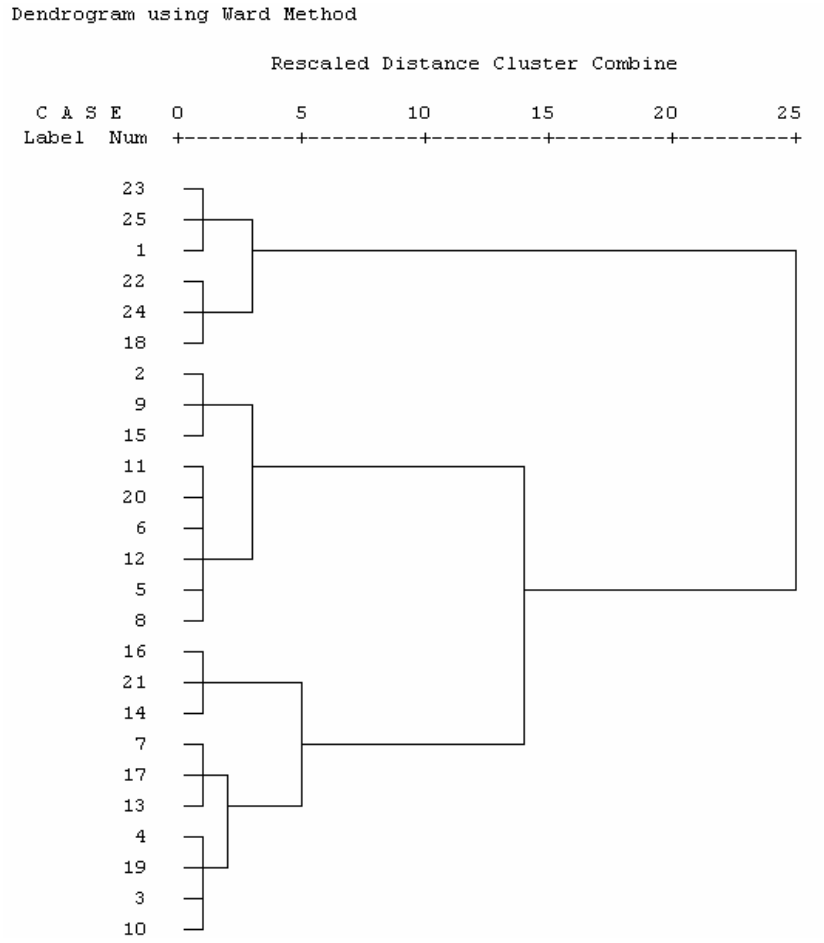


Figura 19. Triphenilacrilonitrili: dendrograma – analiza ierarhică de clasificare

În urma analizei s-a obținut apartenența fiecărui compus la un cluster după cum urmează:

- Cluster 1 (media per cluster egală cu -0.937): 5 compuşi (triph001; triph016; triph021; triph023 şi triph025)
- Cluster 2 (media per cluster egală cu 0.599): 9 compuşi (triph003; triph004; triph007; triph010; triph013; triph014; triph015; triph017 şi triph019)
- Cluster 3 (media per cluster egală cu -2.000): 3 compuşi (triph018; triph022 şi triph024)
- Cluster 4 (media per cluster egală cu 1.847): 8 compuşi (restul compuşilor nespecificați anterior)

Parametrii statisticii descriptive pentru cei 4 clusteri, modelul cu efecte fixe şi respectiv random sunt prezantați în Tabelul 37. Figura 20 prezintă distribuția valorilor logRBA per cluster, respectiv distribuția mediei per clasă.

Variațele în cei 4 clusteri s-au dovedit a nu fi omogene (Levene statistic = 3.530, df1 = 1, df2 = 21, p = 0.0326).

Rezultatele obținute în urma aplicării testului ANOVA sunt redate în Tabelul 38.

Aplicarea testului Welch de comparare a mediilor nu a putut fi aplicat deoarece cel puțin pentru un cluster varianța a fost egală cu 0.

Tabelul 37. Parametrii statistici asociați clusterilor: modelul cu efecte fixe şi random pentru trifenilacrilonitrili

Clustrer	Efecte	n	m	StDev	StErr	Min	Max	BCVar
1		5	-0.937	0.493	0.220	-1.398	-0.398	
2		9	0.599	0.408	0.136	-0.180	1.230	
3		3	-2.000	0.000	0.000	-2.000	-2.000	
4		8	1.847	0.250	0.088	1.447	2.220	
Total		25	0.379	1.385	0.277	-2.000	2.220	
	Fixe			0.361	0.072			
Model	Random				0.833			2.406

n = volumul eşantionului; m = media aritmetică; StDev = deviația standard; StErr = eroarea standard; Min = valoarea minimă; Max = valoarea maximă; Media = media artimetică; BCVar = between component variance

Tabelul 38. ANOVA: logRBA trifeniacrilonitrili

	SS	df	MS	F	p
Între clusteri	43.3139	3	14.4380	110.7126	4.96E-13
În clusteri	2.7386	21	0.1304		
Total	46.0525	24			

SS = suma pătratelor erorilor; df = grade de libertate; MS = media pătratelor erorilor; F = statistica Fisher; p = semnificația statisticii Fisher

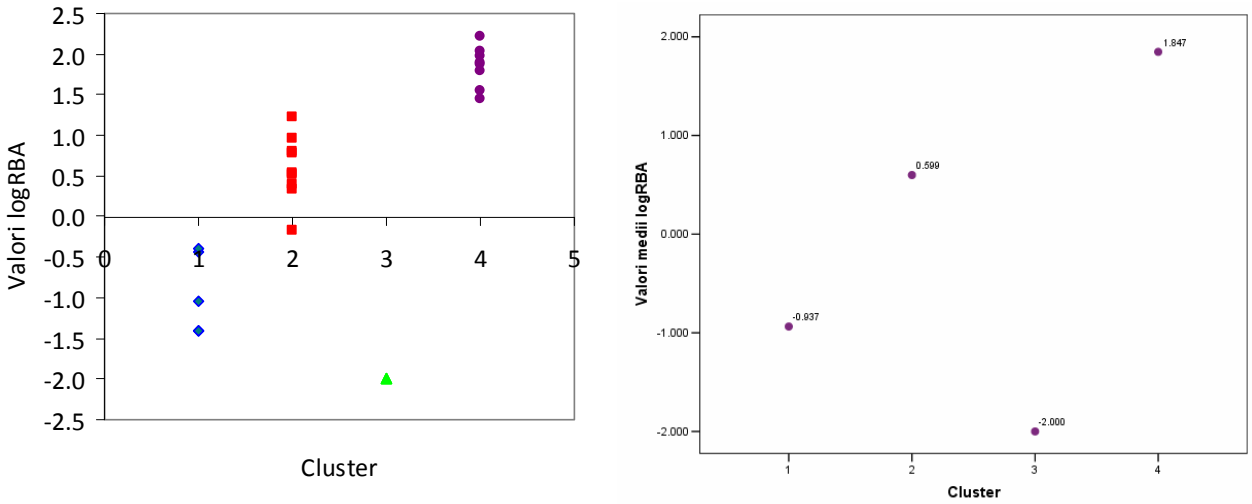


Figura 20. Trifeniacrilonitrili: distribuția valorilor, respectiv a mediei

Analiza de clusterizare s-a aplicat în continuare pentru proprietate şi respectiv cei trei descriptori MDFV ulterior transformării tuturor variabilelor în intervalul [0, 1].

Sumarizarea rezultatele obținute în investigarea proprietății de interes în termeni de modalitate de aglomerare în clusteri sunt redate în Tabelul 39. Un punct clar de demarcare în ceea ce privește diferența este la nivelul 0.7295 (diferență de ordin de mărime) → analiza poate să fie reluată pentru un număr fix de 3 clusteri.

Dendrograma asociată analizei de clusterizare ierarhică este redată în Figura 21.

Tabelul 39. Sumarizarea rezultatelor: coeficienţi de aglomerarea prop + MDFV trifenilacrilonitrili

Nr clusteri	CoefAglomLast	CoefAglPrev	Dif
2	8.7186	6.5900	2.1286
3	6.5900	5.3271	1.2629
4	5.3271	4.5976	0.7295
5	4.5976	3.9127	0.6849
6	3.9127	3.2880	0.6247
7	3.2880	2.7524	0.5356

CoefAglUltim = coeficientul de aglomerare cu valoarea mare pentru numărul de clusteri de interes;
CoefAglPrevc= coeficientul de aglomerare anterior;
Dif = diferența dintre ultim şi anterior;

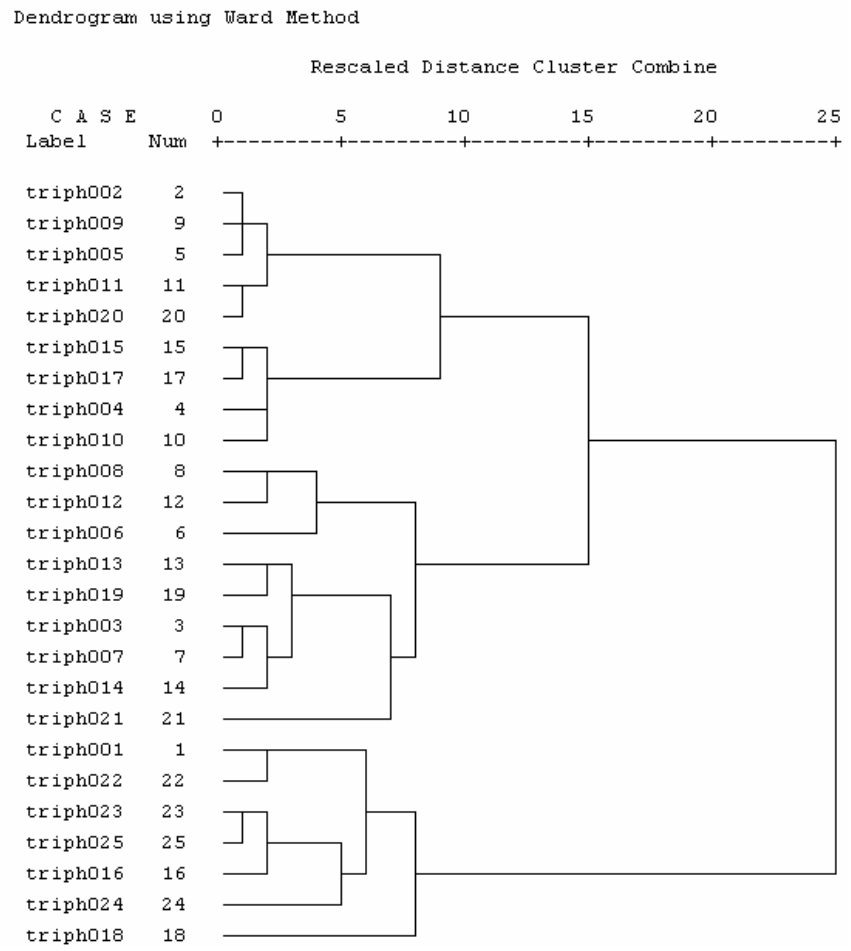


Figura 21. Triphenilacrilonitrili: dendrograma – analiza ierarhică de clasificare (lofRBA + descriptori MDFV)

Alegerea claselor s-a realizat în scopul maximizării diferenței dintre cazurile incluse în fiecare cluster. În urma analizei s-a obținut apartenența fiecărui compus la un cluster după cum urmează:

- Cluster 1: 1 compus (triph018)
- Cluster 2: 5 compuşi (triph006; triph008; triph012; triph021 şi triph024)
- Cluster 3: 19 compuşi (restul compuşilor, nespecificați ca aparținând claselor anterioare)

Testul ANOVA a fost aplicat pentru a identifica diferențe semificative statistic a variabilelor în clusteri iar rezultatele sunt prezentate în Tabelul 33. Omogenitatea variațelor este asigurată la nivelul clusterilor doar pentru $logIC_{50}$ (df1 = 1, df2 = 32, statistica Levene = 1.938 (p = 0.174). Următoarele rezultate au fost obținute pentru descriptorii MDFV:

Tabelul 40. Rezultate statistică descriptivă: clasificare pe baza proprietății şi a valorilor descriptorilor MDFV

Variabila	Cluster	Efecte	n	m	StDev	EtErr	Min	Max	BCVar
logRBA	1		1		.	.	-2.0000		
	2		5	0.7166	1.8434	0.8244	-2.0000	2.2200	
	3		19	0.4158	1.2066	0.2768	-2.0000	2.0330	
	Total		25	0.3793	1.3852	0.2770	-2.0000	2.2200	
	Model	Fixe			1.3450	0.2690			
		Random				0.4939			0.2770
TASaAFDL	1		1		.	.	7.4400		
	2		5	7.3722	0.1321	0.0591	7.2310	7.5430	
	3		19	7.2299	0.0996	0.0229	7.1300	7.4080	
	Total		25	7.2668	0.1225	0.0245	7.1300	7.5430	
	Model	Fixe			0.1063	0.0213			
		Random				0.0789			0.0093
GLCACPdL	1		1		.	.	1.6930		
	2		5	0.0255	1.6436	0.7350	-2.3672	1.8035	
	3		19	-0.0928	0.9572	0.2196	-2.0017	0.8650	
	Total		25	0.0023	1.1242	0.2248	-2.3672	1.8035	
	Model	Fixe			1.1139	0.2228			
		Random				0.2924			0.0579
GMhaAiDR	1		1		.	.	1148		
	2		5	38396	3199	1431	33110	41710	
	3		19	22864	4588	1052	13358	30626	
	Total		25	25101	9066	1813	1148	41710	
	Model	Fixe			4368	874			
		Random				9963			$1.59 \cdot 10^{8}$

Rezultatele testului ANOVA sunt prezentate în Tabelul 41. De remarcat distribuția mediile variabilelor în interiorul clusterilor (Figura 22). Aşa cum rezultă din Tabelul 41, mediile tuturor descriptorilor nu sunt semnificativ diferite între clusteri.

Tabelul 41. Testul ANOVA: clasificare în funcție de valorile proprietății şi descriptorilor MDFV

Variabila	Cluster	SS	df	MS	F	p
logRBA	Între	6.2552	2	3.1276	2	0.2007
	În	39.7972	22	1.8090		
	Total	46.0525	24			
TASaAFDL	Între	0.1114	2	0.0557	5	0.0170
	În	0.2485	22	0.0113		
	Total	0.3599	24			
GLCACPdL	Între	3.0330	2	1.5165	1	0.3138
	În	27.2991	22	1.2409		
	Total	30.3321	24			
GMhaAiDR	Între	$1.55 \cdot 10^{9}$	2	$7.76 \cdot 10^{8}$	41	$4.06 \cdot 10^{-8}$
	În	$4.2 \cdot 10^{8}$	22	$1.91 \cdot 10^{7}$		
	Total	$1.97 \cdot 10^{9}$	24			

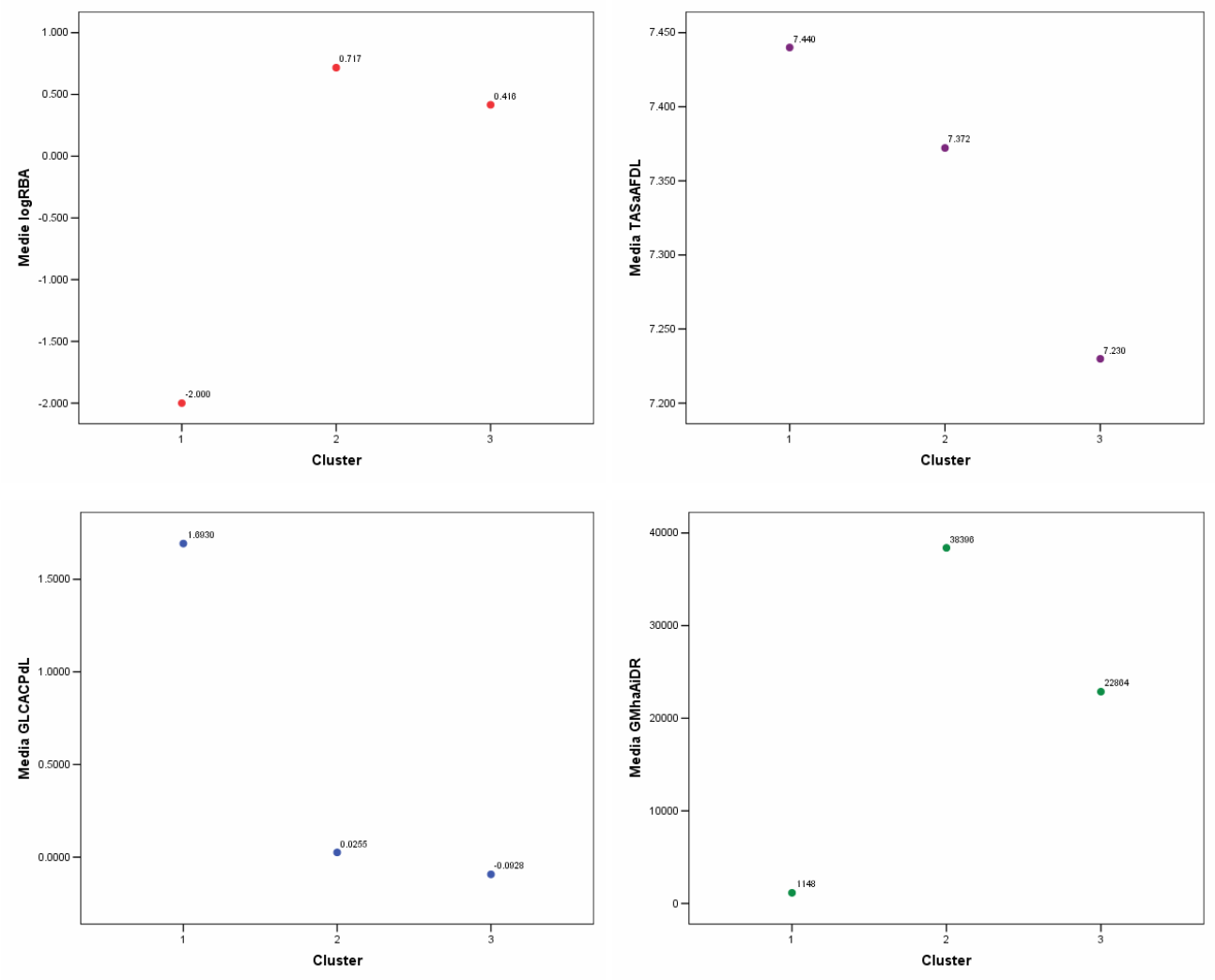
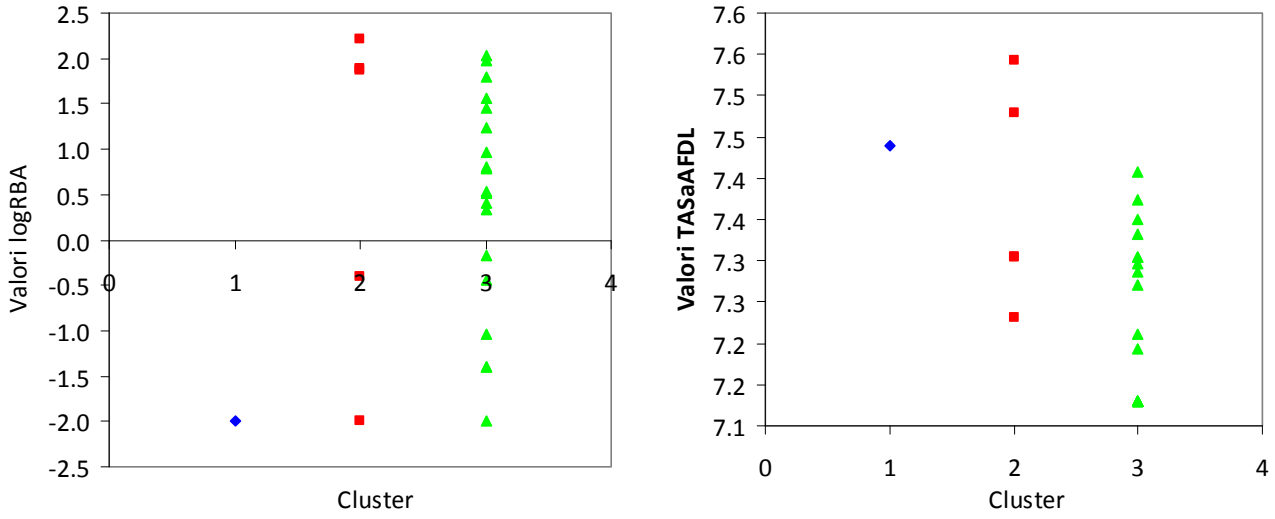


Figura 22.Trifenilacrilonitrili: Contribuții medii în clusteri (prop & descriptori MDFV)

Testul Welch nu a putut fi aplicat datorită distribuției compușilor în clusteri.

Distribuția valorilor în cadrul claselor pentru variabilele este redată în Figura 23.



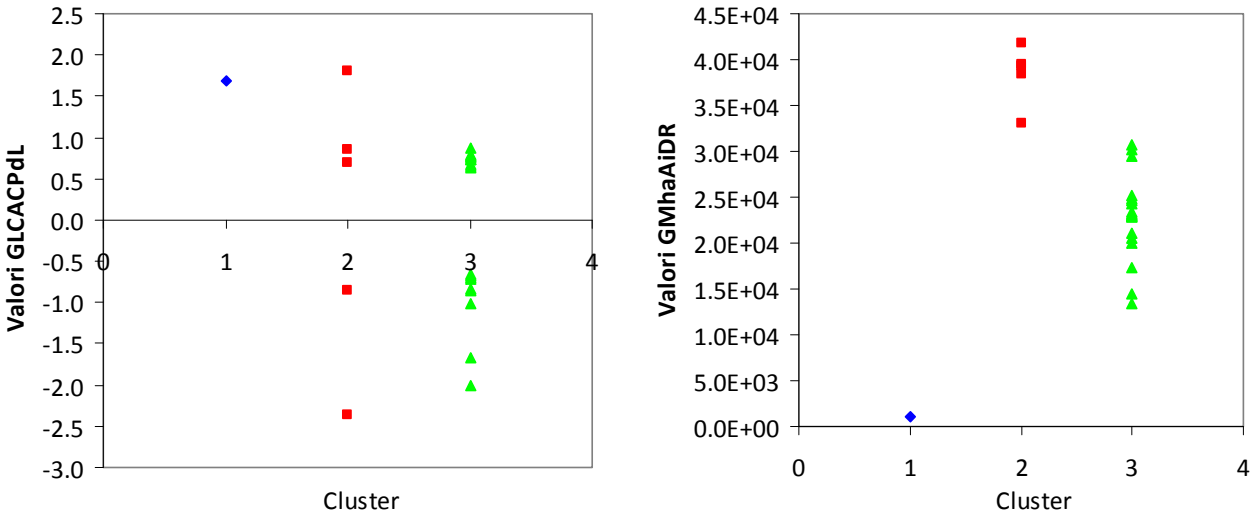


Figura 23. Distribuția valorilor variabilelor în clase (prop & descriptori MDFV)

Următoarele concluzii se pot desprinde pe baza analizei de clusterizare a derivaților de trifenilacrilonitrililor investigați:

- Analiza ierarhică de cluterizare a permis identificarea numărului optim de clusteri: clasificarea optimă se face în ceea ce privește logRBA cu 4 clase iar în ceea ce privește logRBA și descriptorii MDFV cu 3 clase.
- Utilizarea metodei k-means (știut fiind că numărul optim de clusteri este egal cu 4, respectiv 3) clasifică diferit compușii investigați. De remarcat includerea în prima clasă doar a compușilor cu valorare logRBA negativă în cazul clasificării bazat doar pe logRBA și respectiv a valorilor negative extreme, cea maximă în clasa a doau și cele minime (3 valori de -2.000 în clasa a treia). Al patrulea cluster conține doar valori pozitive.
- Valorile medii per clusteri s-au dovedit a nu fi semnificativ statistic diferite pentru logRBA și GLCACPdL.

4.1.2. Analiza factorilor pe baza descriptorilor modelului matematic

Analiza factrorilor se utilizează pentru a identifica variabile, sau factori, capabili să explice modelul de corelație într-un set de variabile observate (în cazul de față variabilele observate sunt reprezentate de valorile descriptorilor MDFV). Analiza factorilor se aplică frecvent pentru a reduce datele şi a identifica un număr mai mic de factori capabili a explica varianța observată dar se poate utiliza şi pentru a genera ipoteze în ceea ce priveşte mecanismul de cauzalitate sau pentru a analiza unele aspecte existente în variabile înainte de aplicare altor metode statistice (de exemplu, pentru a identifica existența colinearității înainte de aplicarea analizei de regresie liniară).

Analiza factorilor este o procedură cu un înalt grad de flexibilitate:

- Metode (şapte) diferite de extracție/identificare a clusterilor
- Metode diferite de rotație (cinci)
- Metode diferite (trei) de calculare a scorurilor factorilor; scorurile obținute pot fi salvate ca şi variabile şi incluse ulterior în alte analize.

Tipuri de variabile: Variabile trebuie să fie cantitative continue măsurabile pe scală interval sau rație. Pot fi incluse în analiza variabilelor datele pentru care coeficientul de corelație Pearson este indicat a fi calculat.

Asumpții: Datele trebuie să aibă o distribuție bivariată normală pentru fiecare pereche de variabile iar observațiile trebuie să fie independente

Analiza factorilor a fost aplicată doar asupra descriptorilor MDFV pentru a identifica, dacă există, factori plecând de la valorile descriptorilor. Analiza s-a realizat cu SPSS 16.0.

Analiza descriptivă: Statistica univariată include media aritmetică, deviația standard şi numărul valid de cazuri pentru fiecare variabilă inclusă în analiză. Soluția inițială pune la dispoziție valorile (eigenvalues = varianța totală explicată de fiecare factor) şi procentele variației explicate (procentul din variația totală atribuit fiecărui factor). Matricea de corelație aduce informații cu privire la coeficienți, nivele de semnificație, determinanți, indicele KMO şi testul de sfericitate Bartlett, inversul, şi imaginea reversă.

Indicele KMO (Kaiser-Meyer-Olkin) – test de măsură a adecvabilității eşantionării – testeaza dacă corelația parțială între variabile este mică. Este utilizat pentru a aprecia dacă analiza factorilor este adecvată a fi aplicată.

- Valoarea între 0.5 şi 1 a indicelul KMO pune în evidență faptul că analiza factorilor este adecvată a fi aplicată.
- Valoarea mai mică de 0.5 indică faptul că analiza factorilor nu este adecvată.

Testul de sfericitate Bartlett:

- Ipoteza testului: variabilele nu sunt corelate la nivelul populației (matricea de corelație în populație este de fapt matrice de identitate: fiecare variabilă se corelează perfect cu ea însăşi – r = 1 – dar nu se corelează cu alte variabile)

Procedura aplicată:

Reducerea datelor → Factor

Descriptiv: → Matricea de corelație: coeficienți & KMO şi Bartlett test

→ Statistica: soluția inițială

Opțiuni: → Valori lipsă: excluderea cazurilor perechi

→ Modalitatea de afişare a coeficienților: sortate după mărime & suprimă valorile absolute mai mici de 0.3

Extragerea: → Metoda: Componente principale

→ Analiza: Matricea de corelație

→ Afişarea:Screeplot & soluția factorilor nerotați

→ Extrage: eigenvalues > 1

Rotația: → Metoda: Varimax (metodă de rotație ortogonală care minimizează numărul de variabile care au valoari de încărcare mari pentru fiecare factor; Simplifică interpretarea factorilor.).

4.1.2.1. Derivați de carbochinonă – activitate anti-tumorală

Patru descriptori MDFV au intrat în analiza factorilor pentru derivații de carbochinone. Matricea de corelație obținută este prezentată în Tabelul 42. Aşa cum se observă din matricea de corelație doar 2 din 6 coeficienți de corelație au valori absolute mai mari de 0.3.

Tabelul 42. Matricea de corelație: derivați de carbochinonă (coeficient de corelație dreapta sus / nivel de semnificație stânga jos)

	TEuIFFDL	GLCIicdI	TAkaFcDL	GLbIAcDR
TEuIFFDL		0.314	0.217	0.335
GLCIicdI	0.029		0.114	0.036
TAkaFcDL	0.099	0.251		-0.314
GLbIAcDR	0.021	0.417	0.029	

Rezultatele indicelui KMO şi a testului Bartlett sunt redate în Tabelul 43. Valoarea indicelui KMO indică faptul că analiza factorilor nu este adecvată (valoarea este mai mică de 0.5). Analiza factorilor ar trebui să se încheie aici dar a fost efectuată până la final pentru exemplificare.

Testul Bartelett este semnificativ statistic ceea ce indică faptul că descriptorii MDFV sunt

corelați.

Tabelul 43. KMO şi testul Bartlett: rezultate derivați carbochinone

Kaiser-Meyer-Olkin		0.394
Testul Bartlett	Approx. Chi-Square	15.987
	Grade de libertate	6
	p	0.014

Rezultatele analizei variațelor explicate de factori este redată în Tabelul 44. În conformitate cu rezultatele prezentate în Tabelul 44, sunt de interes valorile eigen mai mari de 1, indicând astfel un număr de 2 factori. De remarcat faptul că fiecare factor în parte reuşeşte să explice în medie până în 35% din varianță, cumulând o explicare de până la 70%. Reprezentarea grafică a valorilor eigen per factori sunt preznetate în Figura 24.

Tabelul 44. Varianța explicată: rezultate pentru derivații de carbochinone (metoda de extragere: analiza componentelor principale)

	Valori Eigen inițiale			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
Factor	Total	% Var	Cumul%	Total	%Var	Cumul%	Total	% of Variance	Cumulative %
1	1.485	37.129	37.129	1.485	37.129	37.129	1.478	36.943	36.943
2	1.323	33.084	70.212	1.323	33.084	70.212	1.331	33.269	70.212
3	0.790	19.749	89.961						
4	0.402	10.039	100.000						

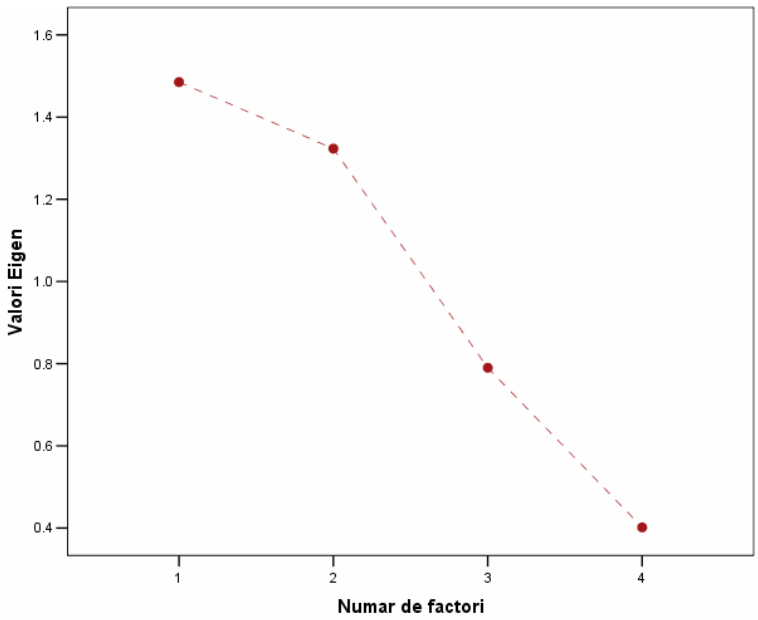


Figura 24. Grafic de tip Scree: derivaşi de carbochinone

Matricea factorilor şi respective matricea factorilor rotați sunt redate în Tabelul 45. Greutatea în primul factor este semnificativă pentru trei descriptori (TEuIFFDL, GLCIicdI şi GLbIAcDR),

respective în cel de-al doilea factor pentru doi descriptori (TAkaFcDL şi GLbIAcDR). Contribuția rămâne semnificativă pentru primii doi descriptori ai primului factor şi respective pentru cei doi descriptorii ai celui de-al doilea factor. Descriptorii cu greutate se pot utiliza mai departe pentru alte analize.

Tabelul 45. Matricea factorilor: derivaţi de carbochinone

Descrriptor MDFV	Matricea factorilor		Matricea factorilor rotaţi	
	Factor 1	Factor 2	Factor 1	Factor 2
TEuIFFDL	**0.8692**	0.0446	**0.8586**	0.1426
GLCIicdI	**0.6496**	0.2572	**0.6897**	-0.1122
TAkaFcDL	0.2084	**0.8402**	0.3433	**0.8340**
GLbIAcDR	**0.5140**	**-0.7412**	0.3834	**-0.7761**

Valorile factorilor pentru fiecare derivate de carbochinonă sunt redate în Tabelul 46. Valorile ambilor factori s-au dovedit a fi normal distribuite la un prag de semnificaţie de 5% (analiză realizată cu EasyFit Professional).

Tabelul 46. Valori ale factorilor identificaţi pentru derivaţii de carbochinonă

Mol	Factor1	Factor2
cqd01	2.14165	-0.36558
cqd02	2.14124	-0.69785
cqd03	1.56105	-0.38285
cqd04	1.60131	0.87542
cqd05	1.14587	-0.11504
cqd06	1.55907	0.82361
cqd07	0.82667	-0.39709
cqd08	1.00241	-2.91032
cqd09	0.41745	1.32062
cqd10	0.66309	0.12288
cqd11	0.4831	-0.35853
cqd12	0.42127	1.3074
cqd13	-0.64942	-0.91549
cqd14	-0.04788	1.69514
cqd15	-0.00309	1.67687
cqd16	0.32074	0.3478
cqd17	-0.28627	-0.66403
cqd18	0.9352	-0.8535
cqd19	-0.62338	-0.67346

Mol	Factor1	Factor2
cqd20	0.06527	-0.78767
cqd21	0.26913	1.63209
cqd22	0.05284	1.80647
cqd23	-0.92467	-0.4003
cqd24	-0.82465	-0.80443
cqd25	-0.7503	-0.87664
cqd26	-0.90484	-0.06156
cqd27	-0.88327	-0.42539
cqd28	-0.96431	-0.31559
cqd29	-0.78689	0.05627
cqd30	-0.64022	-0.09422
cqd31	-0.939	0.5575
cqd32	-1.10161	-0.01155
cqd33	-0.88409	-0.75618
cqd34	-1.30897	1.09188
cqd35	-0.99836	-0.43923
cqd36	-1.03278	1.1354
cqd37	-1.05335	-1.14285

Valorile factorilor identificaţi au fost utilizate în analiza de regresie liniară (metoda includerii trepate a factorilor în analiza de regresie). Statisticile associate modelului de regresie identificat sunt prezantate în Tabelul 47. Modelul de regresie identificat este:

$$\hat{Y} = 5.755 - 0.597*\text{ScorFactor1}$$

Coeficienţii regresiei s-au dovedit a fi semnificativi statistic ($p < 0.05$), Toleranța = 1 şi VIP = 1.

Tabelul 47. Analiza de regresie: factori asociați derivaților de carbochinone

Nr.	R	R^2	R^2_{Adj}	StErr	Change Statistics				Durbin-Watson
					F	df1	df2	p	
1	0.941[a]	0.886	0.883	0.217	271.868	1	35	$4.48 \cdot 10^{-18}$	1.817

[a] regresia realizată cu scorurile factorului 1
R = coeficientul de corelație; R^2 = coeficientul de determinare;
StErr = eroarea standard a estimatului; df = grade de libertate; F = statistica Fisher;
p = nivelul de semnificație

Reprezentarea grafică a relației dintre proprietatea investigate și modelul realizat pe baza unuia din factorii identificați este prezentată în Figura 25.

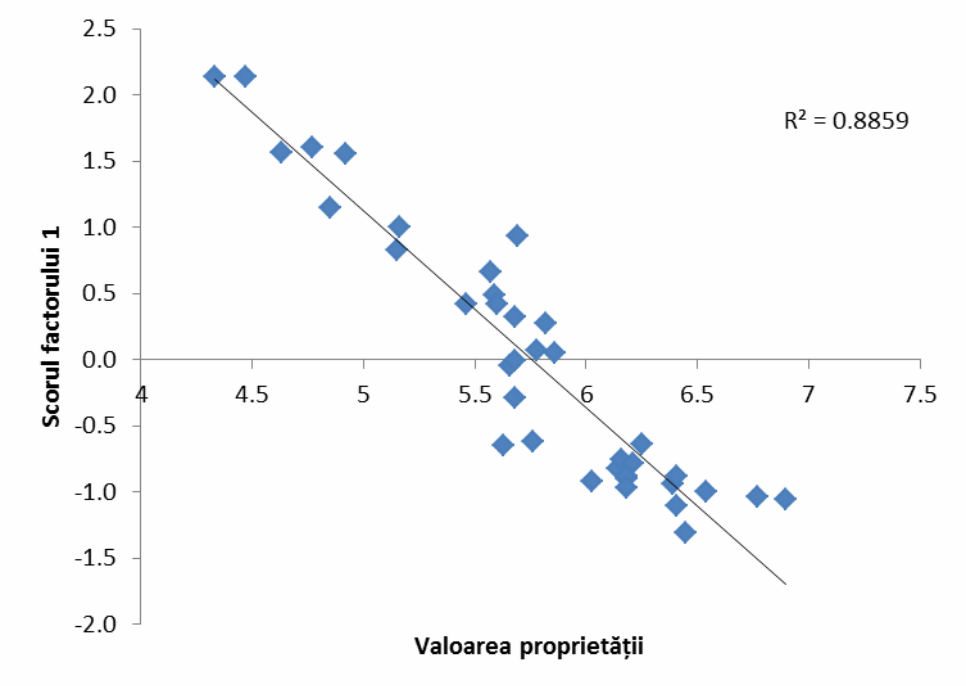


Figura 25. Proprietate vs scoruri associate factorului 1: derivați de carbochinone

Următoarele concluzii se pot desprinde din analiza factorilor pentru derivații de carbochinone cu activitate antitumorală:

- În conformitate cu rezultatele indicelui de KMO analiza factorilor nu este adecvată a fi aplicată pe descriptorii MDFV ai modelului prezentat în [45].
- Aplicarea analizei factorilor identifică existent a doi factori.
- Unul din factorii identificați s-a dovedit a fi în relație de linearitate cu proprietatea investigate, determinarea fiind de 88%. Acest model este semnificativ mai bun în estimare în compara ie cu modelul cu un descriptor [45].

[45] Bolboacă SD, Jantschi L. Raport intermediar 2008: proiect cercetare ID458. 2008; p. 46-69. http://sorana.academicdirect.ro/grants/ID0458/PCE_ID_0458_Extenso_2008.pdf

4.1.2.2. Compuşi organici – traversare barieră hemato-encefalică

Patru descriptori MDFV au intrat în analiza factorilor pentru setul de compuşi organici care traversează bariera hemato-encefalică. Matricea de corelație obținută este prezentată în Tabelul 48. Patru din 6 coeficienți de corelație sunt semnificativi statistic, 3 corelații fiind slabe sau inexistente în conformitate cu regulile empirice de interpretare a coeficientului de corelație.

Tabelul 48. . Matricea de corelație: set compuşi organici (coeficient de corelație dreapta sus / nivel de semnificație stânga jos)

	TLgFAIDI	GAmIAaDI	TAgFIADL	TAgPIADL
TLgFAIDI	1	0.2670	-0.2422	-0.2421
GAmIAaDI	0.0015	1	0.0413	-0.0599
TAgFIADL	0.0036	0.3259	1	0.9881
TAgPIADL	0.0036	0.2560	$1.11 \cdot 10^{-99}$	1

Rezultatele indicelui KMO şi a testului Bartlett sunt redate în Tabelul 49. Valoarea indicelui KMO indică faptul că analiza factorilor pentru acest set de compuşi nu este adecvată (valoarea este mai mică de 0.5). Mai mult, testul Bartelett este semnificativ statistic ceea ce indică faptul că descriptorii MDFV sunt corelați.

Tabelul 49. KMO şi testul Bartlett: rezultate compuşi organici

Kaiser-Meyer-Olkin		0.3509
Testul Bartlett	Approx. Chi-Square	535.38
	df	6
	p	$2.00 \cdot 10^{-112}$

4.1.2.3. Derivați de sulfonamide - inhibitori ai anhidrazei carbonice II & Taxoizi – inhibiția creşterii celulare

Derivați de sulfonamide – inhibitori ai anhidrazei carbonice

Trei descriptori MDFV au intrat în analiza factorilor pentru derivații de sulfonamide. Matricea de corelație obținută este prezentată în Tabelul 50. De remarcat faptul că toți descriptorii au valori absolute ale coeficientului de corelație mai mari de 0.3.

Tabelul 50. Matricea de corelație: derivați de sulfonamide

	TLhFPFdR	GMpFFIdI	TEmFIIDI
TLhFPFdR	1	0.3083	0.3180
GMpFFIdI	0.1067	1	0.9437
TEmFIIDI	0.0992	$2.12 \cdot 10^{-9}$	1

Rezultatele indicelui KMO şi a testului Bartlett sunt redate în Tabelul 51. Valoarea indicelui KMO indică faptul că analiza factorilor este adecvată în cazul setului de compuşi derivați de sulfonamide (valoarea este mai mare de 0.5).

Testul Bartelett este semnificativ statistic ceea ce indică faptul că descriptorii MDFV sunt corelați (Tabelul 51).

Tabelul 51. KMO şi testul Bartlett: rezultate derivați de sulfonamide

Kaiser-Meyer-Olkin		0.551
Test Bartlett	Approx. Chi-Square	35.192
	df	3
	p	$1.1 \cdot 10^{-7}$

Rezultatele analizei variațelor explicate de factori este redată în Tabelul 52. În conformitate cu rezultatele prezentate în Tabelul 52, sunt de interes valorile eigen mai mari de 1, indicând astfel un singur factor. Acest factor este capabil de a explica ~71% din varianță. Reprezentarea grafică a valorilor eigen per factori sunt preznetate în Figura 24.

Tabelul 52. Varianța explicată: rezultate pentru derivații de sulfonamide (metoda de extragere: analiza componentelor principale)

Factor	Valori Eigen inițiale			Extraction Sums of Squared Loadings		
	Total	%Var	Cumul%	Total	%Var	Cumul%
1	2.119	70.633	70.633	2.119	70.633	70.633
2	0.825	27.493	98.126			
3	0.056	1.874	100.000			

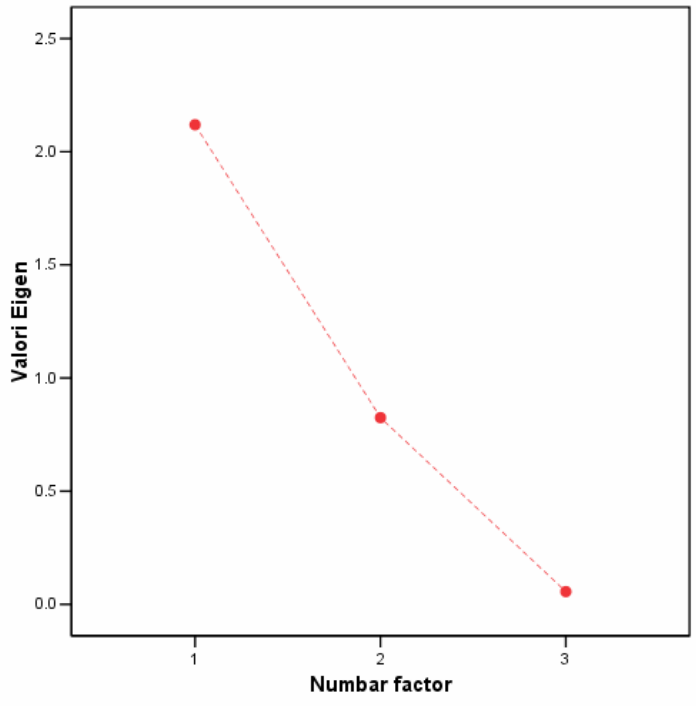


Figura 26. Grafic de tip Scree: derivați de sulfonamide

Valorile factorului pentru derivații de sulfonamine sunt redate în Tabelul 53.

Tabelul 53. Scoruri ale factorului identificat pentru derivații de sulfonamide

Mol	Factor
s001	-1.8619
s002	-0.9331
s003	-0.3796
s004	0.5062
s005	-0.6310
s006	-0.4180
s007	0.6352
s008	1.8002
s009	0.5531
s010	1.9673
s011	0.0663
s012	0.6710
s013	0.5703
s014	0.4933
s015	-1.2410
s016	-0.9313
s017	-0.6101
s018	-0.2569

Scorurile factorului identificat au fost utilizate în analiza de regresie liniară. Statisticile associate modelului de regresie identificat sunt prezantate în Tabelul 47. Modelul de regresie identificat este:

$$\hat{Y} = 5.755 - 0.597*\text{ScorFactor1}$$

Tabelul 54. Analiza de regresie: factori asociați derivaților de sulfonamide

Nr.	R	R^2	R^2_{Adj}	StErr	Change Statistics				Durbin-Watson
					F	df1	df2	p	
1	0.663[a]	0.439	0.404	0.6629	12.522	1	16	0.003	1.162

[a] regresia realizată cu scorurile factorului 1

R = coeficientul de corelație; R^2 = coeficientul de determinare;

StErr = eroarea standard a estimatului; df = grade de libertate; F = statistica Fisher;

p = nivelul de semnificație

Reprezentarea grafică a relației dintre proprietatea investigată şi modelul realizat pe baza unuia din factorii identificați este prezentată în Figura 27.

Următoarele concluzii se pot desprinde din analiza factorilor pentru derivații de carbochinone cu activitate antitumorală:

- În conformitate cu rezultatele indicelui de KMO analiza factorilor este adecvat a fi aplicată pe descriptorii MDFV ai modelului prezentat în [46].

[46] Bolboacă SD, Jantschi L. Raport intermediar 2009: proiect cercetare ID458. 2008; p. 145-148. http://sorana.academicdirect.ro/grants/ID0458/PCE_ID_0458_Extenso_2009.pdf

- Analiza factorilor identifică un singur factor.
- Factorul identificat s-a dovedit a fi in relație de linearitate cu proprietatea investigată, determinarea fiind de aproximativ 44%. Acest model este semnificativ mai slab comparative cu cel mai bun model identificat. Acest model este semnificativ mai slab comparative cu cel mai bun model identificat între proprietatea investigată şi scorul factorului identificat.

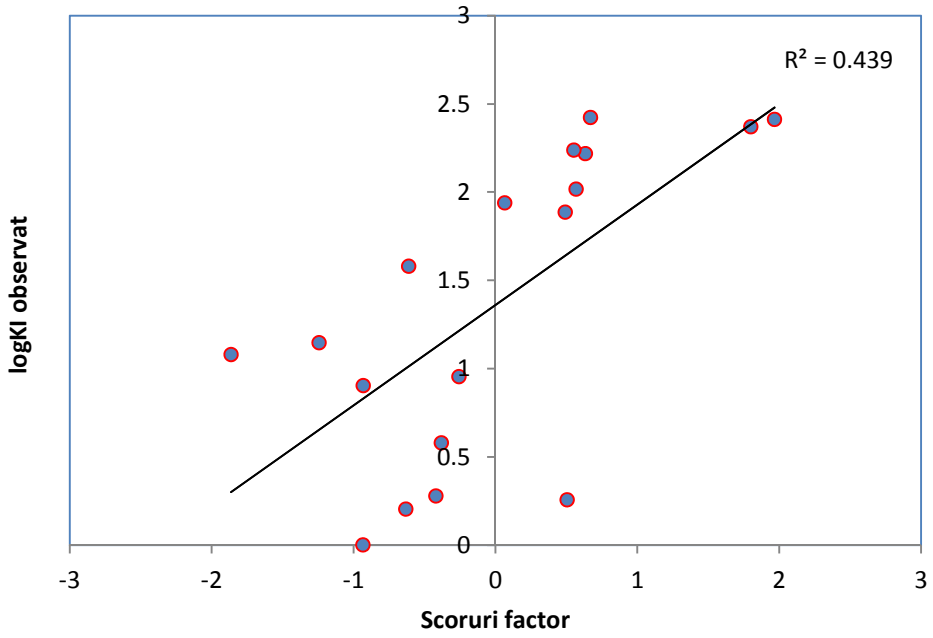


Figura 27. Proprietate vs Scoruri associate factorului: derivați de sulfonamine

Taxoizi – inhibitori ai creşterii celulare

Trei descriptori MDFV au intrat în analiza factorilor pentru taxoizi. Matricea de corelație obținută este prezentată în Tabelul 55. De remarcat faptul că toți descriptorii au valori absolute ale coeficientului de corelație mai mari de 0.3.

Tabelul 55. Matricea de corelație: derivați de sulfonamide

	TAcAIiDR	TQKCPfdL	TMiIPpdL
TAcAIiDR	1	0.8517	0.4507
TQKCPfdL	$8.50 \cdot 10^{-11}$	1	0.4330
TMiIPpdL	$3.73 \cdot 10^{-3}$	$5.26 \cdot 10^{-3}$	1

Rezultatele indicelui KMO şi a testului Bartlett sunt redate în Tabelul 51. Valoarea indicelui KMO indică faptul că analiza factorilor este adecvată în cazul setului de taxoizi (valoarea este mai mare de 0.5).

Testul Bartelett este semnificativ statistic ceea ce indică faptul că descriptorii MDFV sunt

corelați (Tabelul 56).

Tabelul 56. KMO şi testul Bartlett: rezultate taxoizi

Kaiser-Meyer-Olkin Measure		0.6122
Test Bartlett	Approx. Chi-Square	48
	df	3
	p	2.46E-10

Rezultatele analizei variațelor explicate de factori este redată în Tabelul 57. În conformitate cu rezultatele prezentate în Tabelul 57, sunt de interes valorile eigen mai mari de 1, indicând astfel un singur factor. Acest factor este capabil de a explica ~71% din varianță. Reprezentarea grafică a valorilor eigen per factori sunt preznetate în Figura 28.

Tabelul 57. Varianța explicată: rezultate pentru taxoizi (metoda de extragere: analiza componentelor principale)

	Valori Eigen inițiale			Extraction Sums of Squared Loadings		
Factor	Total	%Var	Cumul%	Total	%Var	Cumul%
1	2.1821	72.74	72.74	2.18	72.74	72.74
2	0.6699	22.33	95.06			
3	0.1481	4.94	100			

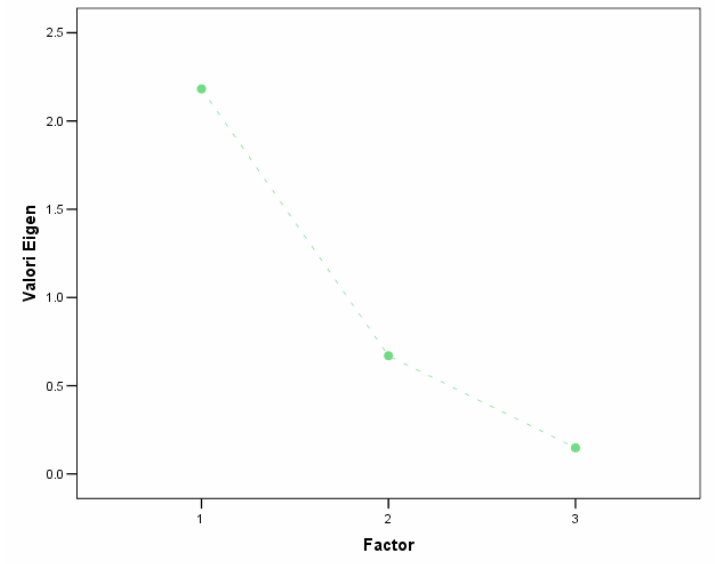


Figura 28. Grafic de tip Scree: taxoizi

Valorile factorului pentru derivații de sulfonamine sunt redate în Tabelul 58.

Scorurile factorului identificat au fost utilizate în analiza de regresie liniară. Statisticile associate modelului de regresie identificat sunt prezantate în Tabelul 59. Modelul de regresie identificat este:

$$\hat{Y} = -0.743 + 1.006 * ScorFactor1$$

Tabelul 58. Analiza de regresie: factori asociați setului de taxoizi

Nr.	R	R^2	R^2_{Adj}	StErr	F	df1	df2	p	Durbin-Watson

1	0.8200	0.6724	0.6622	0.7128	66	1	32	$2.96 \cdot 10^{-9}$	1.699

R = coeficientul de corelație; R^2 = coeficientul de determinare;
StErr = eroarea standard a estimatului; df = grade de libertate; F = statistica Fisher;
p = nivelul de semnificație

Tabelul 59. Scoruri ale factorului identificat pentru taxoizi

Mol	Factor	Mol	Factor
tax001	2.02029	tax020	-0.34656
tax002	2.05883	tax021	-0.80942
tax003	1.3473	tax022	-0.90811
tax004	1.23059	tax023	-0.74886
tax005	1.38474	tax024	-0.16611
tax007	1.18985	tax025	-0.41936
tax008	2.13481	tax026	-0.80942
tax009	1.42218	tax027	-0.58673
tax010	-0.58673	tax028	0.11258
tax011	-0.29261	tax029	0.23041
tax012	-0.95155	tax030	-0.03935
tax013	-0.41886	tax031	-0.47331
tax014	-0.58673	tax032	-0.36651
tax015	-0.57709	tax033	-0.51833
tax016	-0.9655	tax034	0.10048
tax017	0.06413	tax035	-0.47331
tax018	-0.26411		
tax019	-1.98762		

Reprezentarea grafică a relației dintre proprietatea investigată și modelul realizat pe baza unuia din factorii identificați este prezentată în Figura 29.

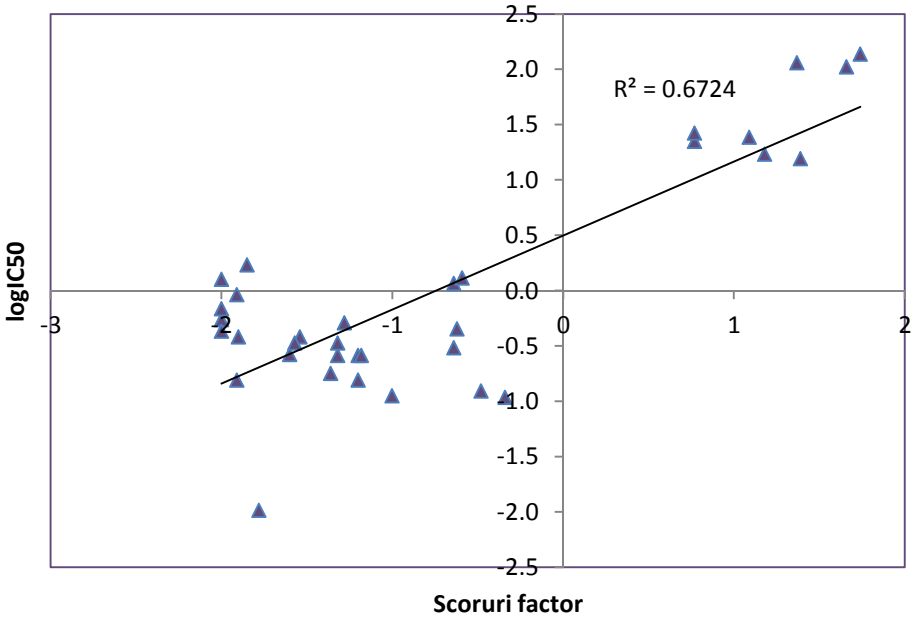


Figura 29. Proprietate vs Scoruri associate factorului: derivați de sulfonamine

Următoarele concluzii se pot desprinde din analiza factorilor pentru derivații de carbochinone cu activitate antitumorală:

- În conformitate cu rezultatele indicelui de KMO analiza factorilor este adecvat a fi aplicată pe

descriptorii MDFV ai modelului prezentat în [47].

- Analiza factorilor identifică un singur factor.
- Factorul identificat s-a dovedit a fi în relație de linearitate cu proprietatea investigată, determinarea fiind de 67%. Acest model este semnificativ mai slab comparative cu cel mai bun model identificat (model cu trei descriptori MDFV) [47].

4.1.2.4. Derivați de trifenilacrilonitril – afinitate relativă de legare receptori de estrogen

Trei descriptori MDFV au intrat în analiza factorilor pentru derivații de trifenilacrilonitril. Matricea de corelație obținută este prezentată în Tabelul 60. De remarcat faptul că toți descriptorii au valori absolute ale coeficientului de corelație mai mici de 0.3.

Tabelul 60. Matricea de corelație: derivați de trifenilacrilonitril

	TASaAFDL	GLCACPdL	GMhaAiDR
TASaAFDL	1	-0.0103	0.2237
GLCACPdL	0.4806		0.0375
GMhaAiDR	0.1413	0.4293	1

Rezultatele indicelui KMO şi a testului Bartlett sunt redate în Tabelul 61. Valoarea indicelui KMO indică faptul că analiza factorilor nu este adecvată în cazul setului de compuşi derivați de trifenilacilonitril (valoarea este mai mare de 0.5), motiv pentru care analiza factorilor se încheie aici. Testul Bartelett nu este semnificativ statistic ceea ce indică faptul că descriptorii MDFV nu sunt corelați (Tabelul 61).

Tabelul 61. KMO şi testul Bartlett: derivați de trifenilacrilonitrili

Kaiser-Meyer-Olkin		0.4963
Testul Bartlett	~Chi-Square	1.1769
	df	3
	p	0.7586

[47] Bolboacă SD, Jantschi L. Raport intermediar 2009: proiect cercetare ID458. 2008; p. 148-152. http://sorana.academicdirect.ro/grants/ID0458/PCE_ID_0458_Extenso_2009.pdf

Obiectivul 4.2. Realizare librărie virtuală

4.2.1. Proiectare implementare aplicație, integrare modele în baza de date, implementare algoritmi de interogare

Scop: Crearea unei librării virtuale pentru seturile de compuşi investigate, librărie care să înglobeze datele obținute în analiza de regresie simplă şi multiplă a proprietăților investigate cu descriptori structurali MDFV.

Utilizatori: Cercetători care doresc aplicarea metodologiei MDFV pe diferite seturi de compuşi.

Modalitate de utilizare: Intranet / Internet.

Restricții de utilizare: utilizarea acestei resurse se face pe bază de parolă pentru secțiunea vizualizării modelelor QSAR.

Pentru fiecare set de date investigat au fost create un număr de 5 tabele în cadrul bazei de date MDFV (vezi Figura 30).

Table	Action	Count	Type	Collation	Size
cqd__logMinEffDose		4,763	MyISAM	latin1_swedish_ci	1.6 MB
cqd_data		37	MyISAM	latin1_swedish_ci	63.6 KB
cqd_mdfv		2,387,280	MyISAM	latin1_swedish_ci	727.2 MB
cqd_prop		1	MyISAM	latin1_swedish_ci	1.3 KB
cqd_qsar		34	MyISAM	latin1_swedish_ci	7.4 KB

Table	Action	Count	Type	Collation	Size
estro__logRBA		3,736	MyISAM	latin1_swedish_ci	4.3 MB
estro_data		144	MyISAM	latin1_swedish_ci	204.3 KB
estro_mdfv		2,387,280	MyISAM	latin1_swedish_ci	2.6 GB
estro_prop		1	MyISAM	latin1_swedish_ci	3.1 KB
estro_qsar		50	MyISAM	latin1_swedish_ci	9.9 KB

Table	Action	Count	Type	Collation	Size
sulfon18__logKI		15,237	MyISAM	latin1_swedish_ci	2.8 MB
sulfon18_data		18	MyISAM	latin1_swedish_ci	30.7 KB
sulfon18_mdfv		2,387,280	MyISAM	latin1_swedish_ci	381.2 MB
sulfon18_prop		1	MyISAM	latin1_swedish_ci	1.1 KB
sulfon18_qsar		34	MyISAM	latin1_swedish_ci	7.1 KB

Table	Action	Count	Type	Collation	Size
taxoids__logIC50		16,255	MyISAM	latin1_swedish_ci	4.9 MB
taxoids_data		34	MyISAM	latin1_swedish_ci	106.1 KB
taxoids_mdfv		2,387,280	MyISAM	latin1_swedish_ci	672.6 MB
taxoids_prop		1	MyISAM	latin1_swedish_ci	1.3 KB
taxoids_qsar		86	MyISAM	latin1_swedish_ci	16.4 KB

Table	Action	Count	Type	Collation	Size
triph__logRBA		4,545	MyISAM	latin1_swedish_ci	1.1 MB
triph_data		25	MyISAM	latin1_swedish_ci	41.6 KB
triph_mdfv		2,387,280	MyISAM	latin1_swedish_ci	508.7 MB
triph_prop		1	MyISAM	latin1_swedish_ci	1.2 KB
triph_qsar		36	MyISAM	latin1_swedish_ci	7.4 KB

Figura 30. Structura tabelară a informației din librăria virtuală

Structura tabelară prezentată în Figura 30 pune în evidență următoarele informații:

- Numărul de compuşi ai setului: ex. 37 compuşi derivați carbochinone.
- Numărul de descriptori MDFV (identic pentru toate seturile de compuşi): 2387280 descriptori.
- Numărul valid de descriptori MDFV (incluşi în analiza de regresie liniară simplă şi multiplă): ex. 4763 descriptori pentru derivații de carbochinone.
- Numărul proprietăților investigate: ex. O singură proprietate pentru derivații de carbochinonă.
- Numărul de modele QSAR identificate a fi valide cu valoarea coeficientului de determinare cât mai apropape de valoarea maximă (1): ex. 34 modele pentru derivații de carbochinonă.

Librăria virtuală a fost proiectată şi implementată. Integrarea modulelor în baza de date şi implementarea algoritmilor de interogare s-a realizat pentru statistica descriptivă, analiza leave-one-out, analiza în setul de învățare şi testare, analiza corelației, calculator şi predictor. În baza de date a fost integrată inclusiv structura 3D a fiecărui compus.

Liniile programului *.php care permite afişarea conținutului librăriei virtuale (Figura 31) sunt după cum urmează:

```
<?
$definition_page["auto_index"]=FALSE;
$definition_page["copyright"]=array("September 2007; August
2010","lori&&sorana");
include("0_mdfv_definitions.php");
echo("<br><br>");
$q=mysql_query("USE `".server_db."`");
if(!array_key_exists("set",$_GET)){
      $q=mysql_query("SELECT `set`,`property` FROM `_jobs` where
`program`='7_prop_bias' AND `t_stop` > 0");
      $n=mysql_num_rows($q);
      if($n==0)die("</body>");
      echo("Descriptors populations:<BR>");
      for($i=0;$i<$n;$i++){
            $r=mysql_fetch_row($q);
            echo("<A
HRef='?set=".$r[0]."&prop=".$r[1]."'>".$r[0]."/".$r[1]."</A>  &nbs
p;  ");
      }
      mysql_free_result($q);
      echo("<BR>");
      $q=mysql_query("SHOW TABLES LIKE '%_prop'");
      $n=mysql_num_rows($q);
      if($n==0)die("</body>");
      echo("Molecules Sets:<UL>");
      for($i=0;$i<$n;$i++){
            $r=mysql_fetch_row($q);
```

```
            $r=explode("_",$r[0]);
            echo("<LI><A HRef='?set=".$r[0]."'>".$r[0]."</A>");
      }
      mysql_free_result($q);
      echo("</UL>");
      $q=mysql_query("SHOW TABLES LIKE '%_qsar'");
      $n=mysql_num_rows($q);
      if($n==0)die("</body>");
      echo("qSARs on Properties (authorization required):<UL>");
      for($i=0;$i<$n;$i++){
            $r=mysql_fetch_row($q);
            $r=explode("_",$r[0]);
            echo("<LI><A
HRef='9_mdfv_clean.php?set=".$r[0]."'>".$r[0]."</A>");
      }
      mysql_free_result($q);
      echo("</UL>");
      die("</body>");
}elseif(array_key_exists("get",$_GET)){
      include("file_get.php");
}elseif(array_key_exists("pdb",$_GET)){
      include("file_pdb.php");
}elseif(array_key_exists("prop",$_GET)){
      define("EPS", 2.22e-16);
      define("MAX_VALUE", 1.2e308);
      define("LOG_GAMMA_X_MAX_VALUE", 2.55e305);
     define("SQRT2PI",  2.50662827463100050241576528481104525300698674060993);
      define("SQRT2", 1.41421356237309504880168872420969807856967187537694);
      define("XMININ", 2.23e-308);
      define("MAX_ITERATIONS", 1000);
      define("PRECISION", 8.88E-016);
      $q=mysql_query("SELECT    *    FROM    `".$_GET["set"]."_prop`    WHERE
`property`='".$_GET["prop"]."'");
      $r=mysql_fetch_row($q);
      array_shift($r);$m=0;
      while(count($r)>0){
            if($r[0]<1e100)$m++;
            array_shift($r);
      }
      $s_m=sqrt($m-2);
      mysql_free_result($q);
      echo("m=".$m."<br>\r\n");
      $q=mysql_query("SELECT `r2` FROM `".$_GET["set"]."__".$_GET["prop"]."`
WHERE 1");
      $n=mysql_num_rows($q);
      echo("<table
border='1'><tr><td>n<td>r2<td>r".$_GET['prop']."<td>t<td>p");
```

```
	if(array_key_exists("p",$_GET)){
		$pp=$_GET["p"];
		if($pp>0.5)$pp=1-$pp;
	}else $pp=2;
	for($i=0;$i<$n;$i++){
		$r=mysql_fetch_row($q);
		$r_1=sqrt($r[0]);
		$t=$r_1*$s_m/sqrt(1.0-$r[0]);
		$p=p_t($m-2,$t);

	if($p<$pp)echo("<tr><td>".$i."<td>".$r[0]."<td>".$r_1."<td>".$t."<td>"
.$p);
	}
	echo("</table>");
	mysql_free_result($q);
	die("</UL></body>");
}elseif(!array_key_exists("property",$_GET)){
	$q=mysql_query("SHOW TABLES LIKE '".$_GET["set"]."_prop'");
	$n=mysql_num_rows($q);
	if($n==0)die("No such set.");
	mysql_free_result($q);
	$q=mysql_query("SHOW TABLES LIKE '".$_GET["set"]."__%'");
	$n=mysql_num_rows($q);
	if($n==0)die("Properties still not available for this set.");
	mysql_free_result($q);
	$q=mysql_query("SHOW TABLES LIKE '".$_GET["set"]."_qsar'");
	$n=mysql_num_rows($q);
	if($n==0)die("Properties still not available for this set.");
	mysql_free_result($q);
	$q=mysql_query("SELECT		DISTINCT		`property`		FROM
`".$_GET["set"]."_qsar`");
	$n=mysql_num_rows($q);
	if($n==0)die("Properties still not available for this set.");
	echo("Properties of ".$_GET["set"]."<UL>");
	for($i=0;$i<$n;$i++){
		$r=mysql_fetch_row($q);
		echo("<LI><A
HRef='?set=".$_GET["set"]."&property=".$r[0]."'>".$r[0]."</A>");
	}
	mysql_free_result($q);
	die("</UL></body>");
}elseif(!array_key_exists("id",$_GET)){
	echo("Set = ".$_GET["set"]."<Br>");
	echo("Property = ".$_GET["property"]."<Br>");
	$columns=array();
	$q=mysql_query("SHOW COLUMNS FROM `".$_GET["set"]."_qsar`");
	for(;$r=mysql_fetch_row($q);)$columns[]=$r[0];
```

```
	mysql_free_result($q);
	echo("<table border='1'><tr>");
	for($i=2;$i<count($columns);$i++){
		echo("<td>".$columns[$i]);
	}
	echo("<td>research");
	$q=mysql_query("SELECT	*	FROM	`".$_GET["set"]."_qsar`	WHERE
`property`='".$_GET["property"]."'");
	for(;$r=mysql_fetch_row($q);){
		$id=array_shift($r);
		array_shift($r);
		echo("<tr><td>".implode("<td>",$r)."<td><A
HRef='?set=".$_GET["set"]."&property=".$_GET['property']."&id=".$id."'>Link"
);
	}
	mysql_free_result($q);
	echo("</table>");

}else{
	if(!array_key_exists("lori",$_GET)){
		die("You need authorization to do this.");
	}
	if(!$_GET["lori"]){
		echo("Options:<UL>");

	$url="?set=".$_GET["set"]."&property=".$_GET["property"]."&id=".$_GET[
"id"]."&lori=";
		echo("<LI><A	HRef='".$url."descriptive_statistics'>Descriptive
Statistics</A><BR><BR>");
		echo("<LI><A		HRef='".$url."leave_one_out'>Leave-One-Out
Analysis</A><BR><BR>");
		echo("<LI><A  HRef='".$url."training_vs_test'>Training  vs.  Text
Experiment</A><BR><BR>");
		echo("<LI><A	HRef='".$url."correlated_correlations'>Correlated
Correlations Analysis</A><BR><BR>");
		echo("<LI><A HRef='".$url."calculator'>Calculator</A><BR><BR>");
		echo("<LI><A HRef='".$url."predictor'>Predictor</A>");
		echo("</UL>");
	}else{
		if(!(file_exists($_GET["lori"].".php")))die("Not Implemented.");
		include($_GET["lori"].".php");
	}
}
function p_t($df,$t){
	$p=$df/2;
	$x=0.5+0.5*$t/pow(pow($t,2)+$df,0.5);
	$beta_gam=exp(-logBeta($p,$p)+$p*log($x)+$p*log(1.0-$x));
```

```
      return(2.0*$beta_gam*betaFraction(1.0-$x,$p,$p)/$p);
}
function betaFraction($x,$p,$q){
      $c=1.0;
      $s_pq=$p+$q;
      $p_p=$p+1.0;
      $p_m=$p-1.0;
      $h=1.0-$s_pq*$x/$p_p;
      if(abs($h)<XMININ)$h=XMININ;
      $h=1.0/$h;
      $f=$h;
      $m=1;
      $d=0.0;
      while(($m<=MAX_ITERATIONS)&&(abs($d-1.0)>PRECISION)){
            $m2=2*$m;$d=$m*($q-$m)*$x/(($p_m+$m2)*($p+$m2));$h=1.0+$d*$h;
            if(abs($h)<XMININ)$h=XMININ;
            $h=1.0/$h;$c=1.0+$d/$c;
            if(abs($c)< XMININ)$c=XMININ;
            $f*=$h*$c;$d=-
($p+$m)*($s_pq+$m)*$x/(($p+$m2)*($p_p+$m2));$h=1.0+$d*$h;
            if(abs($h)<XMININ)$h=XMININ;
            $h=1.0/$h;$c=1.0+$d/$c;
            if(abs($c)<XMININ)$c=XMININ;
            $d=$h*$c;$f*=$d;
            $m++;
      }
      return($f);
}
function logBeta($p,$q){
      global $logBetaCache_res,$logBetaCache_p,$logBetaCache_q;
      if(($p!=$logBetaCache_p)||($q!=$logBetaCache_q)){
            $logBetaCache_p=$p;$logBetaCache_q=$q;

      if(($p<=0.0)||($q<=0.0)||(($p+$q)>LOG_GAMMA_X_MAX_VALUE))$logBetaCache
_res=0.0;
           else  $logBetaCache_res=logGamma($p)+logGamma($q)-logGamma($p+$q);
      }
      return($logBetaCache_res);
}
function logGamma($x){
      global $logGammaCache_res,$logGammaCache_x;
      $lg_d1=-0.57721566490153286051951 74;
      $lg_d2=0.42278433509846713939937 77;
      $lg_d4=1.79175946922805500009402 3;
      $lg_p1=array(4.94523535929672704673 4888,201.81126208567750839155 65,229
0.83837383134639302673 9,11319.67205903380828685045,28557.24636571635335736389,38484.96228443793359990269,26377.48787624195437963534,7225.813979700288197
```

```
698961);
      $lg_p2=array(4.97460784556893203501206 4,542.413859989107049410198 6,155
06.93864978364947665077,184793.2904445632425417223,1088204.76946882876749847
,3338152.967987029735917223,5106661.678927352456275255,3074109.0548505395562
50927);
      $lg_p4=array(14745.02166059939948905062,2426813.369486704502836312,121
475557.4045093227939592,2663432449.630976949898078,29403789566.3455389990687
6,170266573776.5398868392998,492612579337.743088758812,560625185622.39514650
78242);
      $lg_q1=array(67.48212550303777196073036,1113.332393857199323513008,773
8.757056935398733233834,27639.870744033407088985 85,54993.1020622615732979441
4,61611.221800660021278333 52,36351.275915019405072762 87,8785.536302431013170
870835);
      $lg_q2=array(183.0328399370592604055942,7765.049321445005871323047,133
190.3827966074194402448,1136705.821321969608938755,5267964.11743794691757753
8,13467014.54311101692290052,17827365.30353274213975932,9533095.591844353613
395747);
      $lg_q4=array(2690.530175870899333379843,639388.5654300092398984238,413
55999.30241388052042842,1120872109.61614794137657,14886137286.78813811542398
,101680358627.2438228077304,341747634550.7377132798597,446315818741.97132864
62081);
      $lg_c=array(-0.001910444077728,8.4171387781295e-4,-5.952379913043012e-
4,7.936507935003502 48e-4,-
0.002777777777777681622553,0.083333333333333333331554247,0.0057083835261);
      $lg_frtbig=2.25e76;
      $pnt68=0.6796875;
      if($x==$logGammaCache_x) return $logGammaCache_res;
      $y=$x;
      if(($y>0.0)&&($y<=LOG_GAMMA_X_MAX_VALUE)){
            if($y<=EPS){$res=-log($y);}
            elseif($y<=1.5){
                  if($y<$pnt68){$corr=-log($y);$xm1=$y;}
                  else{$corr=0.0;$xm1=$y-1.0;}
                  if(($y<=0.5)||($y>=$pnt68)){
                        $xden=1.0;$xnum=0.0;
                        for($i=0;$i<8;$i++){
                              $xnum=$xnum*$xm1+$lg_p1[$i];
                              $xden=$xden*$xm1+$lg_q1[$i];
                        }
                        $res=$corr+$xm1*($lg_d1+$xm1*($xnum/$xden));
                  }else{
                        $xm2=$y-1.0;$xden=1.0;$xnum=0.0;
                        for($i=0;$i<8;$i++){
                              $xnum=$xnum*$xm2+$lg_p2[$i];
                              $xden=$xden*$xm2+$lg_q2[$i];
                        }
                        $res=$corr+$xm2*($lg_d2+$xm2*($xnum/$xden));
```

```
                }
            }elseif($y<=4.0){
                $xm2=$y-2.0;$xden=1.0;$xnum=0.0;
                for($i=0;$i<8;$i++){
                    $xnum=$xnum*$xm2+$lg_p2[$i];
                    $xden=$xden*$xm2+$lg_q2[$i];
                }
                $res=$xm2*($lg_d2+$xm2*($xnum/$xden));
            }elseif($y<=12.0){
                $xm4=$y-4.0;$xden=-1.0;$xnum=0.0;
                for($i=0;$i<8;$i++){
                    $xnum=$xnum*$xm4+$lg_p4[$i];
                    $xden=$xden*$xm4+$lg_q4[$i];
                }
                $res=$lg_d4+$xm4*($xnum/$xden);
            }else{
                if($y<=$lg_frtbig){
                    $res=$lg_c[6];$ysq=$y*$y;
                    for($i=0;$i<6;$i++)$res=$res/$ysq+$lg_c[$i];
                }else{$res=0.0;}
                $res/=$y;
                $corr=log($y);
                $res=$res+log(SQRT2PI)-0.5*$corr;
                $res+=$y*($corr-1.0);
            }
        }else{
            $res=MAX_VALUE;
        }
        $logGammaCache_x=$x;
        $logGammaCache_res=$res;
        return $res;
}
?>
```

Figura 31. Pagina principală a librăriei virtuale

Liniile programului care implementează analiza de corelație [36] între valoarea observată şi cea estimată a properietății/activității investigate sunt:

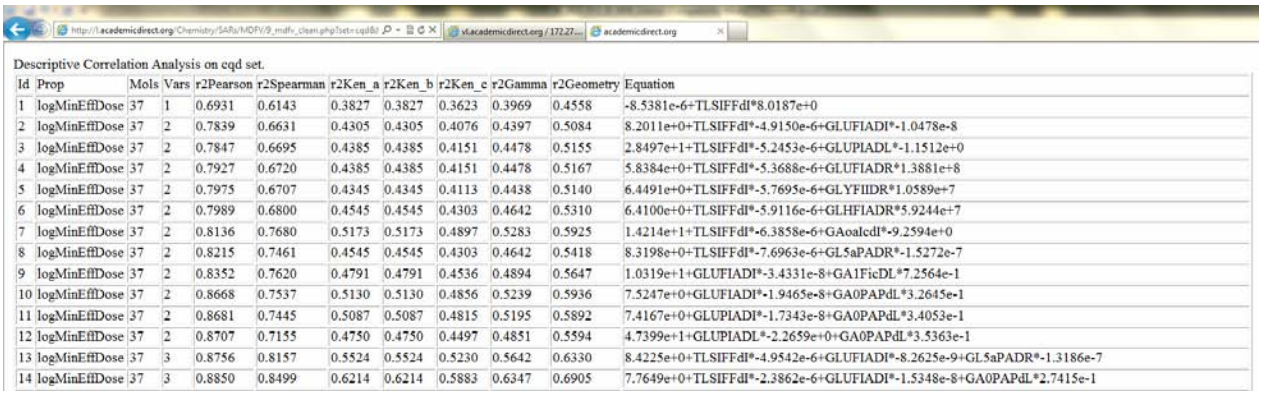
```
<?
include("0_mdfv_definitions.php");
include("Pearson_Spearman_Kendall_Gamma.php");
$q=mysql_query("USE `".server_db."`");
if(!array_key_exists("lori",$_GET))die("You must use an authorization key to
see this.");

$q=mysql_query("SELECT `id` FROM `".$_GET["set"]."_data`");
$n=mysql_num_rows($q);
mysql_free_result($q);
$qSARs=array();
$q=mysql_query("SELECT `id` FROM `".$_GET["set"]."_qsar` ORDER BY `var` ASC,
`r2` ASC");
for(;$r=mysql_fetch_row($q);){
      $qSARs[]=$r[0];
}
mysql_free_result($q);
echo("Descriptive Correlation Analysis on ".$_GET["set"]." set.");
echo("<table border='1'>");
echo("<tr><td>Id<td>Prop<td>Mols<td>Vars<td>r2Pearson<td>r2Spearman<td>r2Ken
_a<td>r2Ken_b<td>r2Ken_c<td>r2Gamma<td>r2Geometry<td>Equation");
for($iq=0;$iq<count($qSARs);$iq++){
      $q=mysql_query("SELECT    *    FROM    `".$_GET["set"]."_qsar`    WHERE
`id`='".$qSARs[$iq]."' LIMIT 1");
      $r=mysql_fetch_row($q);
```

```
	mysql_free_result($q);
	$r[5]=substr($r[5],2);
	$r[4]=trim(sprintf("%.4f",$r[4]));
	$q=mysql_query("SELECT * FROM `".$_GET["set"]."_prop` WHERE
`property`='".$r[1]."' LIMIT 1");
	$prop=mysql_fetch_array($q,MYSQL_ASSOC);
	array_shift($prop);
	mysql_free_result($q);
	$mols=array();$Y_exp=array();
	foreach($prop as $k => $v){if($v<1e100){$mols[]=$k;$Y_exp[]=$v;}}
	unset($prop);
	$r[5]=explode("*",$r[5]);
	for($i=0;$i<count($r[5])-1;$i++){
		$r_d=explode("+",$r[5][$i]);
		if(count($r_d)>1){
			$r[5][$i+1]=$r_d[count($r_d)-1]."*".trim($r[5][$i+1]);
			unset($r_d[count($r_d)-1]);
			$r[5][$i]=trim(implode("+",$r_d));
		}
	}
	unset($r_d);
	$regr_indx=array();$regr_coef=array();$regr_desc=array();
	$regr_coef[0]=array_shift($r[5]);$regr_desc[0]="1";$regr_indx[0]=0;
	for($i=0;$i<count($r[5]);$i++){
		$tmp=explode("*",$r[5][$i]);
		$regr_desc[$i+1]=$tmp[0];
		$regr_coef[$i+1]=$tmp[1];
		$q=mysql_query("SELECT `id` FROM `_mdfv` WHERE `name` LIKE
BINARY '".$tmp[0]."' LIMIT 1");
		$tmp=mysql_fetch_row($q);
		$regr_indx[$i+1]=$tmp[0];
		mysql_free_result($q);
	}
	unset($tmp);
	for($i=0;$i<count($regr_coef);$i++){$regr_coef[$i]=trim(sprintf("%.4e"
,$regr_coef[$i]));}
	$r[5]=$regr_coef[0];
	for($i=1;$i<count($regr_coef);$i++){
		$r[5].="+".$regr_desc[$i]."*".$regr_coef[$i];
	}
	for($i=1;$i<count($regr_desc);$i++){
		$GLOBALS[$regr_desc[$i]]=array();
		for($j=0;$j<count($mols);$j++){

	$GLOBALS[$regr_desc[$i]][$j]=desc_vals($regr_indx[$i],$mols[$j]);
		}
	}
```

```
	$Y_mod=regr_esti($mols,$regr_coef,$regr_desc);
	$r2Pearson=pow(r1($Y_exp,$Y_mod),2);
	$p_Y_exp=pozitii($Y_exp);
	$p_Y_mod=pozitii($Y_mod);
	$r2Spearman=pow(r1($p_Y_exp,$p_Y_mod),2);
	list($r2Ken_a,$r2Ken_b,$r2Ken_c,$r2Gamma)=Kendall_Gamma(array($Y_exp,$
Y_mod),count($Y_exp));
	$r2Geometry=1.0;
	$r2Geometry*=$r2Pearson;
	$r2Geometry*=$r2Spearman;
	$r2Geometry*=$r2Ken_a;
	$r2Geometry*=$r2Ken_b;
	$r2Geometry*=$r2Ken_c;
	$r2Geometry*=$r2Gamma;
	$r2Geometry=pow($r2Geometry,1/6);
	echo("<tr>");
	echo("<td>".$r[0]);
	echo("<td>".$r[1]);
	echo("<td>".$r[2]);
	echo("<td>".$r[3]);
	echo("<td>".trim(sprintf("%.4f",$r2Pearson)));
	echo("<td>".trim(sprintf("%.4f",$r2Spearman)));
	echo("<td>".trim(sprintf("%.4f",$r2Ken_a)));
	echo("<td>".trim(sprintf("%.4f",$r2Ken_b)));
	echo("<td>".trim(sprintf("%.4f",$r2Ken_c)));
	echo("<td>".trim(sprintf("%.4f",$r2Gamma)));
	echo("<td>".trim(sprintf("%.4f",$r2Geometry)));
	echo("<td>".$r[5]);
	unset($r);
	unset($mols);
	unset($Y_exp);
	unset($Y_mod);
	unset($p_Y_exp);
	unset($p_Y_mod);
	unset($regr_indx);
	unset($regr_coef);
	for($i=1;$i<count($regr_desc);$i++){
		unset($GLOBALS[$regr_desc[$i]]);
	}
	unset($regr_desc);
}
echo("</table>");
die("You may try here a top three qualification.");

function desc_vals($id,$mol){
	$q=mysql_query("SELECT `".$mol."` FROM `".$_GET["set"]."_mdfv` WHERE
`id`='".$id."' LIMIT 1");
```

```
	$r=mysql_fetch_row($q);mysql_free_result($q);return($r[0]);
}
function regr_esti(&$mols,&$regr_coef,&$regr_desc){
	$n=count($mols);
	$r=array();
	for($i=0;$i<$n;$i++){
		$r[$i]=$regr_coef[0];
		for($j=1;$j<count($regr_coef);$j++){
			$r[$i]+=$regr_coef[$j]*$GLOBALS[$regr_desc[$j]][$i];
		}
		$r[$i]=sprintf("%.4e",$r[$i]);
	}
	return($r);
}
function disp_array($aa){
	if(!(is_array($aa))){echo("$"."aa schuld be an array!<br>");return;}
	$n=count($aa);
	if($n==0){echo("$"."aa is an empty array!<br>");return;}
	echo("<table border='1'>");
	echo("<tr>");
	for($i=0;$i<$n;$i++){
		echo("<td>".$aa[$i]);
	}
	$m=count($GLOBALS[$aa[0]]);
	for($i=0;$i<$m;$i++){
		echo("<tr>");
		for($j=0;$j<$n;$j++){
			echo("<td>".$GLOBALS[$aa[$j]][$i]);
		}
	}
	echo("</table>");
}
?>
```

Analiza de corelație este astfel disponibila (vezi Figura 32) şi permite alegerea modelului cu puterea cea mai mare de estimare şi respectiv clasificarea modelelor în funcție de puterea de estimare (şapte coeficienți de corelație [36]).



Descriptive Correlation Analysis on eqd set.

Id	Prop	Mols	Vars	r2Pearson	r2Spearman	r2Ken_a	r2Ken_b	r2Ken_c	r2Gamma	r2Geometry	Equation
1	logMinEffDose	37	1	0.6931	0.6143	0.3827	0.3827	0.3623	0.3969	0.4558	-8.5381e-6+TLSIFFdI*8.0187e+0
2	logMinEffDose	37	2	0.7839	0.6631	0.4305	0.4305	0.4076	0.4397	0.5084	8.2011e+0+TLSIFFdI*-4.9150e-6+GLUFIADI*-1.0478e-8
3	logMinEffDose	37	2	0.7847	0.6695	0.4385	0.4385	0.4151	0.4478	0.5155	2.8497e+1+TLSIFFdI*-5.2453e-6+GLUPIADL*-1.1512e+0
4	logMinEffDose	37	2	0.7927	0.6720	0.4385	0.4385	0.4151	0.4478	0.5167	5.8384e+0+TLSIFFdI*-5.3688e-6+GLUFIADR*1.3881e+8
5	logMinEffDose	37	2	0.7975	0.6707	0.4345	0.4345	0.4113	0.4438	0.5140	6.4491e+0+TLSIFFdI*-5.7695e-6+GLYFIIDR*1.0589e+7
6	logMinEffDose	37	2	0.7989	0.6800	0.4545	0.4545	0.4303	0.4642	0.5310	6.4100e+0+TLSIFFdI*-5.9116e-6+GLHFIADR*5.9244e+7
7	logMinEffDose	37	2	0.8136	0.7680	0.5173	0.5173	0.4897	0.5283	0.5925	1.4214e+1+TLSIFFdI*-6.3858e-6+GAoaIcdI*-9.2594e+0
8	logMinEffDose	37	2	0.8215	0.7461	0.4545	0.4545	0.4303	0.4642	0.5418	8.3198e+0+TLSIFFdI*-7.6963e-6+GL5aPADR*-1.5272e-7
9	logMinEffDose	37	2	0.8352	0.7620	0.4791	0.4791	0.4536	0.4894	0.5647	1.0319e+1+GLUFIADI*-3.4331e-8+GA1FicDL*7.2564e-1
10	logMinEffDose	37	2	0.8668	0.7537	0.5130	0.5130	0.4856	0.5239	0.5936	7.5247e+0+GLUFIADI*-1.9465e-8+GA0PAPdL*3.2645e-1
11	logMinEffDose	37	2	0.8681	0.7445	0.5087	0.5087	0.4815	0.5195	0.5892	7.4167e+0+GLUPIADI*-1.7343e-8+GA0PAPdL*3.4053e-1
12	logMinEffDose	37	2	0.8707	0.7155	0.4750	0.4750	0.4497	0.4851	0.5594	4.7399e+1+GLUPIADL*-2.2659e+0+GA0PAPdL*3.5363e-1
13	logMinEffDose	37	3	0.8756	0.8157	0.5524	0.5524	0.5230	0.5642	0.6330	8.4225e+0+TLSIFFdI*-4.9542e-6+GLUFIADI*-8.2625e-9+GL5aPADR*-1.3186e-7
14	logMinEffDose	37	3	0.8850	0.8499	0.6214	0.6214	0.5883	0.6347	0.6905	7.7649e+0+TLSIFFdI*-2.3862e-6+GLUFIADI*-1.5348e-8+GA0PAPdL*2.7415e-1

Figura 32. Analiza corelației: derivați de carbochinină (unde id = numarul de identificare al modelului QSAR în tabelul corespunzător setului investigat, Prop = abrevierea properietății/activității investigate, Mol = volumul eșantionului, Vars = numarul variabileor din modelul QSAR, r2Pearson = coeficient de determinare Pearson, r2Spearman = coeficient de determinare al rangurilor Spearman, r2Ken_a/_b/_c = coeficient de determinare Kendall a, b, respectiv c, r2Gamma = coeficient de determinare Gamma, r2Geometry = coeficient de determinare geometric)

Pentru fiecare model QSAR, prin activarea link-ului se pot obține următoarele informații cu privire la modelul accesat (Figura 33).

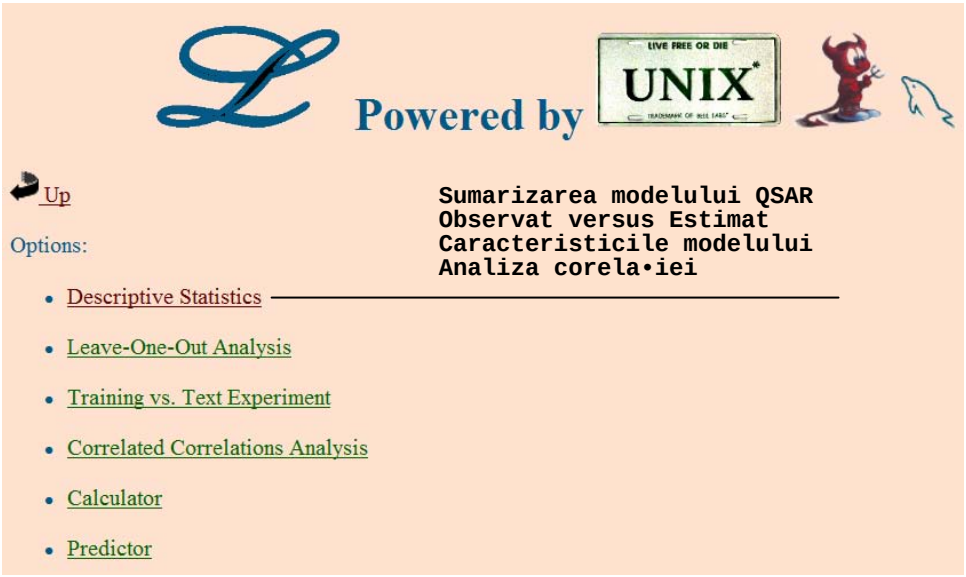


Figura 33. Modalități de analiză a modelelor QSAR prin intermediul librăriei virtuale

Un exemplu de analiză descriptivă a modelului este redată în Figura 34.

Descriptive

Set Name	Molecules Number	Property Name	Molecules Number	Independent Variables	Determination Coefficient	Structure-Activity Relationship
triph	25	logRBA	25	4	0.972192068224109	4.33235125503893E-001+GLLPAiDI*-1.76069288761726E-004+GQCCIPdR* 2.55032658782438E+003+TLWIFaDR*-3.47996599623693E+006+GQbPCcdR* 1.47999037532581E+003

Model

No	Mol	Prop	GLLPAiDI	GQCCIPdR	TLWIFaDR	GQbPCcdR	Estimated	Diff%
1	triph001	-1.046	29282	0.004737	0.0000027989	0.0010712	-0.796	27
2	triph002	1.556	18127	0.004191	0.0000024777	0.0012206	1.114	33
3	triph003	0.342	19038	0.004413	0.0000024777	0.00015404	-0.059	283
4	triph004	0.519	17984	0.004372	0.0000027989	0.001208	0.465	11

Correlation Analysis

No	Prop	GLLPAiDI	GQCCIPdR	TLWIFaDR	GQbPCcdR	Estim
1	-1.046	29282	0.004737	0.0000027989	0.0010712	-0.796
2	1.556	18127	0.004191	0.0000024777	0.0012206	1.114
3	0.342	19038	0.004413	0.0000024777	0.00015404	-0.059
4	0.519	17984	0.004372	0.0000027989	0.001208	0.465
5	1.792	10308	0.003831	0.0000022226	0.0008236	1.873
6	1.869	8593	0.004254	0.0000024777	0.0005085	1.900
7	0.785	9234	0.003986	0.0000024777	0.0001834	0.622
8	2.22	1847.1	0.003508	0.0000022226	0.0006569	2.292
9	1.447	17596	0.003876	0.0000022226	0.0012812	1.382
10	0.398	15876	0.003797	0.0000022226	0.0007627	0.716
11	1.968	18379	0.004154	0.0000022226	0.0013099	1.995
12	1.892	10062	0.003799	0.0000022226	0.0008468	1.869
13	0.959	10217	0.003701	0.0000022226	0.0004425	0.993
14	-0.18	15116	0.003446	0.0000019628	0.00019963	0.025
15	1.23	14987	0.0036	0.0000019628	0.0008629	1.422
16	-0.444	17088	0.0025702	0.0000017175	0.0009487	-0.593
17	0.806	17470	0.003175	0.0000017175	0.000925	0.847
18	-2	26390	0.0024312	0.0000013994	0.0007132	-1.827

Submit Query

Powered by UNIX

Up

Significant Correlation and Probability defined in functions_split.php
Significant Correlation to Reject is set to: 0.29289321881345
Significant Correlation to Accept is set to: 0.70710678118655
Significant Probability to Reject is set to: 0.1
Significant Probability to Accept is set to: 0.01
number of measurements: 25
number of variables: 6

Pearson's quantitative correlation and significance levels from Student's t

r	Prop	GLLPAiDI	GQCCIPdR	TLWIFaDR	GQbPCcdR	Estim
Prop	-	0.6825	0.5255	0.3592	0.3498	**0.986**
GLLPAiDI	**1.7066e-4**	-	0.1445	0.1666	0.1909	0.6922
GQCCIPdR	**6.9805e-3**	0.4908	-	**0.9358**	0.3385	0.533
TLWIFaDR	0.0778	0.4259	**6.7208e-12**	-	0.3176	0.3643
GQbPCcdR	0.0865	0.3606	0.0979	0.1219	-	0.3548
Estim	**2.1552e-19**	**1.2612e-4**	**6.0843e-3**	0.0734	0.0818	-

...

Concordances and Disconcordances

r,p	Prop	GLLPAiDI	GQCCIPdR	TLWIFaDR	GQbPCcdR	Estim
Prop	0:0	0:0	0:4	0:4	0:4	**7:0**
GLLPAiDI	6:1	0:0	0:7	0:7	0:7	0:0
GQCCIPdR	1:1	3:4	0:0	**7:0**	0:4	0:4
TLWIFaDR	0:5	3:4	**7:0**	0:0	0:6	0:4
GQbPCcdR	0:4	0:7	0:6	0:7	0:0	0:4
Estim	**7:0**	6:1	1:1	0:5	0:4	0:0

Intersection

r,p	Prop	GLLPAiDI	GQCCIPdR	TLWIFaDR	GQbPCcdR	Estim
Prop	-	-	-	-	-	-
GLLPAiDI	6:1	-	-	-	-	-
GQCCIPdR	1:5	3:11	-	-	-	-
TLWIFaDR	0:9	3:11	**14:0**	-	-	-
GQbPCcdR	0:8	0:14	0:10	0:13	-	-
Estim	**14:0**	6:1	1:5	0:9	0:8	-

Figura 34. Analiza descriptivă a unui model QSAR corespunzător derivaților de trifenilacrilonitril

4.2.3. Testare mediu virtual

Testarea mediului virtual creat s-a realizat pe parcursul dezvoltării acestuia, în momentul realizării modulelor de interogare precum şi la sfârşitul implementării. Au fost urmărite câteva aspecte:

- minimizarea numărului câmpurilor de tip text;
- minimizarea numărului de clicuri necesare pentru îndeplinirea unui acțiuni specificate;
- minimizarea timpului de răspuns pentru fiecare acțiune.

Pe parcursul dezvoltării şi respectiv în momentul implementării modulelor de interogare mediul virtual a fost testat de către membrii echipei de implementare a proiectului. Testarea finală s-a realizat cu ajutorul unui eşantion format din studenți şi masteranzi.

Protocolul de testare a librăriei virtuale

Scop: stabilirea performanțelor de bază, stabilirea şi validarea măsurilor de performanță şi identificarea conceptelor de desing în scopul îmbunătățirii eficienței şi satisfacției utilizatorului.

Obiective:

1. Determinarea neconcordanțelor de proiectare şi a problemelor de utilizare la nivelul interfeței utilizatorului şi a conținutului. Surse potențiale de eroare:
    a. Erori de navigare: eşecul de a localiza funcțiile, utilizarea excesivă a tastelor pentru îndeplinirea unei funcții, eşecul de urmare a parcursului de ferestre cerut.
    b. Erori de prezentare: eşecul de a localiza şi acționa în mod corespunzător pentru obținerea informației dorite în ecranul identificat, erori de selecție datorate ambiguității etichetelor.
    c. Probleme de utilizare:
2. Testarea mediului virtual în condiții de test controlat cu utilizatori reprezentativi. Datele obținute s-au utilizat pentru a identifica dacă mediul creat îndeplineşte condițiile de eficacitatea, eficiența şi interfața plăcută.
3. Stabilirea performanțelor de referință şi respectiv a nivelului de satisfacție a utilizatorului.

Material şi metode:

- Descrierea eşantionului: ▪ Obiectivul 1: Membrii echipei de cercetare; ▪ Obiectivul 2: Medii echipei de cercetare împreună cu 10 cercetători care nu au participat la dezvoltarea sistemului; ▪ Obiectivul 3: Un eşantion format din 35 studenți şi masteranzi cu cunoştințe prealabile de utilizare a calculatorului.
- Număr sesiuni de test: ▪ Obiectivul 1: 2 (inițial - final (după ultimele modificări identificate ca fiind necesare)); ▪ Obiectivul 2: 1 (Anexa 1); ▪ Obiectivul 3: 2 (2 săptămâni diferență).
- Mediul de testare: toate testele s-au realizat pe aceleaşi echipamente de testare (identitate în

componente hardware şi software). Testarea mediului virtual a fost realizată de către toți participanții cu utilizarea impusă a browser-ului Internet Explorer.

- Instruierea participanților: participanții au fost informați în prima sesiune de testare cu privire la scopul testării, mediul şi modalitatea de testare, precum şi cu privire la necesitatea onestității răspunsurilor.
- Date de colectate: în conformitate cu chestionarul din Anexa 1 (date de testare a mediului virtual) & 2 (date demografice - eşantionul utilizat pentru cel de-al treilea obiectiv).

Rezultatele testării au identificat un mediu virtual performant (timp scurt necesar pentru a realiza o anumită acțiune, număr mic de paşi de urmat pentru a îndeplinit acțiunea specificată), acurat (număr mic de greşeli în îndeplinirea unei acțiuni; nici o eroarea nu a fost fatală – a permis îndeplinirea acțiunii prin punerea la dispoziție a informației corecte), reutilizare intuitivă (la a doua testare participanții şi-au aminte ce anume trebuie să facă ca să îndeplinească acțiunile cerute), răspuns emoțional adecvat (cât de confortabil s-a simțit persoana testată la sfârşitul testului; ar recomanda prietenilor utilizarea sistemului?).

Obiectivul 4.3. Valorificarea rezultatelor

3.1. Documentare, identificare şi selectare compuşi chimici din clasele studiate

Următoarele baze de date au fost utilizate pentru identificarea compuşilor chimici din clasele studiate: PubChem (http://pubchem.ncbi.nlm.nih.gov/), ChemSpider (http://www.chemspider.com/), ChemIDplus (http://chem.sis.nlm.nih.gov/chemidplus/) şi eMolecules (http://www.emolecules.com/). Criteriile de căutare au impus căutarea compuşilor din clasa studiată şi cu activitatea/proprietatea investigată. Au fost identificați compuşi pentru fiecare din clasele de compuşi investigate în cadrul proiectului.

În cele ce urmează se va face exemplificarea pe eşantionul identiicat care a cunținut cel mai mare număr de molecule (compuşi organici ce traversează bariera hemato-encefalică). Compuşi au fost identificați în baza de date PubChem şi pregătiți pentru modelare la fel ca şi compuşii pe baza cărora s-a obținut modelul predictiv (vezi modelul prezentat anterior). Clasificarea compuşilor ca activi, respectiv inactivi a fost luată dintr-o lucrare publicată anterioar [48] (vezi Tabelul 62).

Tabelul 62. Compuşi organici ce traversează bariera hemato-encefalică: denumirea compusului, identificatorul PubMed (CID), clasificarea ca activ vs inactiv observată (Obs) şi prezisă (Pred) pe baza modelului identificat

Nr.	Denumire	CID	Obs	Pred
1	Adenosine	191	1	0
2	Alfentanil	51263	1	0
3	Alosetron	2099	1	1
4	Amiloride	16231	1	0
5	Aripiprazole	60795	1	1
6	Benztropine	2344	1	1
7	Betaxolol	2369	1	0
8	Bisoprolol	2405	1	0
9	Brimonidine	2435	1	1
10	Bromocriptine	31101	1	0
11	Butorphanol	2487	1	0
12	Chloral hydrate	2707	1	1
13	Chlordiazepoxide	2712	1	0
14	Chlorpheniramine	2725	1	1
15	Chlorzoxazone	2733	1	0
16	Citalopram	2771	1	0
17	Clemastine	2781	1	1
18	Clonazepam	2802	1	0
159	Cyclopentolate	2905	0	0
160	Cyclophosphamide	2907	0	0
161	Cytarabine	596	0	0
162	Dantrolene	2952	0	0
163	Dapsone	2955	0	0
164	Delavirdinemesylate	5625	0	0
165	Dexamethasone	5743	0	0
166	Dexpanthenol	4678	0	0
167	Diazoxide	3019	0	0
168	Dibucaine	3025	0	0
169	Dicloxacillin	3041	0	0
170	Digoxin	15478	0	0
171	Diltiazem	3076	0	1
172	Dinoprostone	9691	0	0
173	Disopyramide	3114	0	1
174	Dofetilide	71329	0	1
175	Dorzolamide	3154	0	0
176	Econazole	33745	0	0

[48] Kortagere S, Chekmarev D, Welsh WJ, Ekins S. New predictive models for blood-brain barrier permeability of drug-like molecules. Pharm Res 2008;25:1836-1845.

19	Clorazepate	2809	1	0
20	Clozapine	2818	1	1
21	Cyclobenzaprine	2895	1	1
22	Cyproheptadine	2913	1	1
23	Dezocine	40841	1	0
24	Dipivefrin	3105	1	0
25	Dolasetron	3148	1	1
26	Doxazosin	3157	1	1
27	Doxepin	667477	1	1
28	Dronabinol	2978	1	1
29	Droperidol	3168	1	1
30	Emedastine	3219	1	1
31	Entacapone	5281081	1	0
32	Esmolol	59768	1	0
33	Estazolam	3261	1	1
34	Fexofenadine	3348	1	0
35	Fluoxetine	3386	1	0
36	Flurazepam	3393	1	1
37	Fluvoxamine	5324346	1	0
38	Formoterol	3410	1	0
39	Fosphenytoin	56339	1	0
40	Galantamine	3449	1	1
41	Granisetron	3510	1	1
42	Hydrocodone	411697	1	1
43	Hydromorphone	3648	1	1
44	Isotretinoin	5538	1	1
45	Labetalol	3869	1	0
46	Levobunolol	39468	1	0
47	Levocabastine	54385	1	0
48	Maprotiline	4011	1	0
49	Meperidide	3034126	1	1
50	Metaxalone	15459	1	0
51	Methadone	4095	1	0
52	Methocarbamol	4107	1	0
53	Methoxamine	6082	1	0
54	Methyldopa	4138	1	0
55	Molindone	23897	1	1
56	Nalbuphine	4419	1	0
57	Naratriptan	4440	1	0
58	Nefazodone	4449	1	0
59	Nortriptyline	4543	1	0
60	Ondansetron	4595	1	1
61	Orphenadrine	4601	1	1
62	Oxcarbazepine	34312	1	1
63	Oxycodone	4635	1	1

177	Ephedrine	5032	0	0
178	Eplerenone	443872	0	0
179	Epoprostenol	5280427	0	0
180	Eprosartan	60879	0	0
181	Estramustine	18140	0	0
182	Etidronic acid	3305	0	0
183	Etodolac	3308	0	1
184	Famciclovir	3324	0	0
185	Famotidine	3325	0	0
186	Fenoldopam	3341	0	0
187	Fenoprofen	3342	0	1
188	Flavoxate	3354	0	0
189	Flecainide	3356	0	0
190	Floxuridine	3363	0	0
191	Flunisolide	82153	0	0
192	Fluoxymesterone	6446	0	0
193	Flurbiprofen	3394	0	0
194	Flutamide	3397	0	0
195	Fluvastatin	446155	0	0
196	Fosfomycin	3417	0	1
197	Furosemide	3440	0	0
198	Ganciclovir	3454	0	0
199	Gatifloxacin	5379	0	1
200	Gemcitabine	60750	0	0
201	Gemfibrozil	3463	0	1
202	Glimepiride	3476	0	0
203	Glipizide	3478	0	0
204	Glyburide	3488	0	0
205	Hydralazine	3637	0	1
206	Ibutilide	60753	0	0
207	Idarubicin	42890	0	0
208	Ifosfamide	3690	0	0
209	Imiquimod	57469	0	0
210	Indapamide	3702	0	0
211	Isoetharine	3762	0	0
212	Isosorbide dinitrate	170113	0	1
213	Isradipine	3784	0	0
214	Ketotifen	3827	0	1
215	Lamivudine	3877	0	0
216	Lansoprazole	3883	0	0
217	Latanoprost	5311221	0	0
218	Leflunomide	3899	0	0
219	Letrozole	3902	0	0
220	Levamisole	26879	0	1
221	Lindane	727	0	0

64	Oxymorphone	4639	1	0
65	Paroxetine	4691	1	0
66	Phenelzine	3675	1	0
67	Phenylephrine	6041	1	0
68	Pirbuterol	4845	1	0
69	Pramipexole	4885	1	0
70	Prazosin	4893	1	0
71	Procyclidine	4919	1	1
72	Propoxyphene	10100	1	0
73	Pseudoephedrine	7028	1	0
74	Quazepam	4999	1	0
75	Quetiapine	5002	1	1
76	Rizatriptan	5078	1	0
77	Scopolamine	5184	1	1
78	Secobarbital	5193	1	0
79	Sertraline	5203	1	0
80	Sibutramine	5210	1	1
81	Sufentanil	41693	1	1
82	Sumatriptan	5358	1	0
83	Thiethylperazine	5440	1	1
84	Thiothixene	5454	1	1
85	Tiagabine	5466	1	0
86	Timolol	5478	1	0
87	Tolazoline	5504	1	1
88	Tramadol	5523	1	0
89	Trazodone	5533	1	1
90	Trimethobenzamide	5577	1	1
91	Venlafaxine	5656	1	0
92	Zaleplon	5719	1	0
93	Ziprasidone	60854	1	1
94	Zolpidem	5732	1	0
95	Zolmitriptan	5731	1	0
96	Acarbose	41774	0	0
97	Acetazolamide	1986	0	0
98	Acetylcysteine	581	0	0
99	Acyclovir	2022	0	0
100	Adefovir	60172	0	0
101	Allopurinol	2094	0	0
102	Alprostadil	214	0	0
103	Altretamine	2123	0	0
104	Aminoglutethimide	2145	0	0
105	Amlodipine	2162	0	0
106	Amoxicillin	2171	0	0
107	Ampicillin	2174	0	0
108	Amprenavir	2177	0	0
222	Linezolid	3929	0	1
223	Lisinopril	5362119	0	0
224	Lodoxamide	44564	0	0
225	Loracarbef	3956	0	0
226	Losartan	3961	0	0
227	Lovastatin	53232	0	0
228	Mechlorethamine	4033	0	0
229	Medroxyprogesterone	10631	0	0
230	Melphalan	4053	0	0
231	Mercaptopurine	667490	0	0
232	Meropenem	64778	0	0
233	Mesalamine	4075	0	0
234	Metaproterenol	4086	0	0
235	Metformin	4091	0	0
236	Methimazole	1349907	0	1
237	Methylergonovine	8226	0	0
238	Metoclopramide	4168	0	0
239	Metolazone	4170	0	0
240	Metyrosine	3125	0	0
241	Mexiletine	4178	0	0
242	Miglitol	441314	0	0
243	Milrinone	4197	0	0
244	Minoxidil	4201	0	0
245	Moexipril	91270	0	0
246	Moricizine	34633	0	1
247	Moxifloxacin	4259	0	1
248	Mycophenolic acid	446541	0	0
249	Nabumetone	4409	0	1
250	Naloxone	4425	0	0
251	Naphazoline	4436	0	1
252	Naproxen	1302	0	1
253	Nateglinide	4443	0	0
254	Nedocromil	50294	0	0
255	Nicardipine	4474	0	0
256	Nifedipine	4485	0	0
257	Nimodipine	4497	0	0
258	Nisoldipine	4499	0	0
259	Nitazoxanide	41684	0	0
260	Nitrofurantoin	4509	0	0
261	Nitroglycerin	4510	0	1
262	Nizatidine	4513	0	0
263	Norgestrel	13109	0	0
264	Ofloxacin	4583	0	1
265	Olopatadine	60865	0	0
266	Olsalazine	6816262	0	0

109	Amrinone	3698	0	0
110	Anastrozole	2187	0	0
111	Anthralin	2202	0	0
112	Argatroban	92722	0	0
113	Azathioprine	2265	0	0
114	Aztreonam	5362041	0	0
115	Baclofen	2284	0	0
116	Balsalazide	5362070	0	0
117	Beclometasone	20469	0	0
118	Benazepril	2311	0	0
119	Bepridil	2351	0	0
120	Brinzolamide	68844	0	0
121	Budesonide	63006	0	0
122	Bumetanide	2471	0	0
123	Bupivacaine	2474	0	0
124	Calcitriol	6398761	0	0
125	Candesartan	2541	0	0
126	Capsaicine	2548	0	0
127	Captopril	2550	0	0
128	Cefaclor	2609	0	0
129	Cefadroxil	2610	0	0
130	Cefazolin	33255	0	0
131	Cefdinir	6399011	0	0
132	Cefditoren	6437877	0	0
133	Cefixime	54362	0	0
134	Cefmetazole	2626	0	0
135	Cefonicid	43592	0	0
136	Cefoperazone	135784	0	0
137	Cefotaxime	2632	0	0
138	Cefoxitin	37194	0	0
139	Cefpodoxime	6335986	0	0
140	Ceftazidime	157706	0	0
141	Ceftibuten	5282242	0	0
142	Ceftizoxime	2655	0	0
143	Ceftriaxone	5479530	0	0
144	Cefuroxime	2659	0	0
145	Celecoxib	2662	0	0
146	Cephalexin	27447	0	0
147	Chlorpropamide	2727	0	0
148	Chlorthalidone	2732	0	0
149	Cholecalciferol	6221	0	0
150	Cholestyramine	3086319	0	0
151	Ciclopirox	2749	0	0
152	Cidofovir	60613	0	0
153	Cladribine	1546	0	0
267	Oseltamivir	65028	0	0
268	Oxaprozin	4614	0	0
269	Oxybutynin	4634	0	0
270	Pantoprazole	4679	0	0
271	Pemirolast	57697	0	0
272	Penbutolol	37464	0	0
273	Penciclovir	4725	0	0
274	Pentamidine	4735	0	0
275	Pentoxifylline	4740	0	0
276	Perindopril	107807	0	0
277	Pindolol	4828	0	1
278	Pioglitazone	4829	0	0
279	Pramoxine	4886	0	1
280	Procainamide	4913	0	0
281	Procarbazine	4915	0	0
282	Propafenone	4932	0	0
283	Propylthiouracil	657298	0	0
284	Pyridoxine	1054	0	0
285	Quinapril	54892	0	0
286	Quinidine	1065	0	1
287	Ramipril	5038	0	0
288	Rivastigmine	77991	0	1
289	Rofecoxib	5090	0	0
290	Rosiglitazone	77999	0	1
291	Sildenafil	5212	0	0
292	Simvastatin	54454	0	0
293	Streptozocin	5299	0	0
294	Sulfacetamide	5320	0	0
295	Sulfasalazine	5353980	0	0
296	Sulfinpyrazone	5342	0	0
297	Sulindac	5352	0	1
298	Tamsulosin	129211	0	0
299	Tazarotene	5381	0	0
300	Terazosin	5401	0	0
301	Terbutaline	5403	0	0
302	Ticlopidine	5472	0	1
303	Tocainide	38945	0	0
304	Tolazamide	5503	0	1
305	Tolbutamide	5505	0	0
306	Tolmetin	5509	0	0
307	Torasemide	41781	0	0
308	Trandolapril	5484727	0	0
309	Triamcinolone	31307	0	0
310	Triamterene	5546	0	0
311	Valacyclovir	5647	0	0

154	Clindamycin	29029	0	0
155	Clopidogrel	2806	0	1
156	Clotrimazole	2812	0	1
157	Colchicine	2833	0	0
158	Cromolyn	2882	0	0

312	Voriconazole	5231054	0	0
313	Warfarin	6691	0	0
314	Zileuton	60490	0	0
315	Zoledronic acid	68740	0	0

3.2. Predicție activitate pe baza structurii prin folosirea modelelor structură-activitate obținute

Predicția activității/proprietății s-a realizat prin aplicarea modelului matermatic asupra compuşilor identificați. În Tabelul 66 este prezentată abilitatea de predicție a modelului matematic a setului de compuşi organici ce traversează bariera hemato-encefalică, pentru acest set fiind identificat eşantionul cu compoziția cea mai heterogenă şi număr cel mai mare de compuşi.

Abilitățile modelului în clasificarea corectă a compuşilor s-a realizat prin calcularea unui număr de 11 indicatori statistici (Tabelul 67, acuratețea, rate de eroare, probabilitatea inițială de apartenență la o clasa (de compuşi activi sau inactivi, sensibilitatea, specificitatea, rata falşilor negativi, rata falşilor pozitivi, predictivitatea pozitivă, predictivitatea negativă, probabilitatea de clasificare în clasa compuşilor activi, probabilitatea de clasificare în clasa compuşilor inactivi, probabilitatea clasificării greşite ca şi compus activ, probabilitatea clasificării greşite ca şi compus inactiv, rația de probabilitate) şi a intervalelor de confidență asociate acestora. O parte din parametrii utilizați pentru a evalua abilitățile de predicție a modelului cu fost definiți de Cooper şi colab. [49] în timp ce alți parametrii au fost adaptați după parametrii utilizați în evaluarea studiilor medicale de diagnostic [50]. Intervalele de confidență asociate fiecărui parametru au fost calculate sub asumția distribuției binomiale [51-55], prin

[49] Cooper JA, Saracci R, Cole P. Describing the validity of carcinogen screening tests. British Journal of Cancer 1979;39:87-89.

[50] Bolboacă S, Jäntschi L, Achimaş Cadariu A. Creating Diagnostic Critical Appraised Topics. CATRom Original Software for Romanian Physicians. Applied Medical Informatics 2004;14:27-34.

[51] Drugan T, Bolboacă S, Jäntschi L, Achimaş Cadariu A. Binomial Distribution Sample Confidence Intervals Estimation 1. Sampling and Medical Key Parameters Calculation. Leonardo Electronic Journal of Practices and Technologies 2003;3:47-74.

[52] Bolboacă S, Jäntschi L. Optimized Confidence Intervals for Binomial Distributed Samples. International Journal of Pure and Applied Mathematics 2008;47(1):1-8.

[53] Bolboacă SD, Jäntschi L. Communication of Results on Risk Factors Studies: Confidence Intervals. Leonardo Journal of Sciences 2007;10:179-187.

aplicarea unei proceduri de optimizare [56, 57].

Tabelul 63. Indicatori statistici utilizați în analiza predictivității

Parametrul (Abrevierea)	**Formula**	**Definition**
Concordanța (CC) / Acuratețea (AC) / Rata de lipsă a erorii	100*(AP+AN)/n	Fracția totală a compușilor corect clasificați
Rata de eroare (ER)	100* (FP+FN)/n = 1-CC	Fracția totală a compușilor clasificați greșit
Proporția prealabilă de apartenență la o clasă (activ / inactiv) (PPP)	n_i/n	Proporția compușilor ce aparțin clasei i
Sensibilitatea (Se)	100*AP/(AP+FN)	Procentul de compuși activi asignați corect de către model ca aparținând clasei de compuși activi
Rata falșilor negativi (sub-clasificare, FNR)	100*FN/(AP+FN) = 1-Se	Procentul de compuși activi asignați incorect de către model clasei inactive
Specificitatea (Sp)	100*AN/(AN+FP)	Procentul de compuși inactivi asignați corect de către model ca aparținând clasei inactive
Rata falșilor pozitivi (supra-clasificare, FPR)	100*FP/(FP+AN) = 1-Sp	Procentul de compuși inactivi asignați incorect de către model clasei active
Predictivitatea pozitivă (PP)	100*AP/(AP+FP)	Procentul de compuși corect asignați ca fiind activi raportat la totalitatea compușilor clasificați de model ca fiind activi
Predictivitatea negativă (NP)	100*AN/(AN+FN)	Procentul de compuși corect asignați ca fiind inactivi raportat la totalitatea compușilor clasificați de model ca fiind inactivi
Probabilitatea clasificării compușilor ca - activi (PCA) - inactivi (PCIC)	 (AP+FP)/n (FN+AN)/n	 - Probabilitatea de a clasifica un compus ca activ (adevărat& falși pozitivi) - Probabilitatea de a clasifica un compus ca inactiv (adevărat& falși negativi)
Probabilitatea unei clasificări greșite - ca și compus activ (PWCA) - ca și compus inactiv (PWCI)	 FP/(FP+AP) FN/(FN+AN)	 Probabilitatea unei clasificări pozitive false Probabilitatea unei clasificări negative false
Rata șansei (OR)	(AP*AN)/(FP*FN)	Rata clasificării corecte în grupul compușilor activi raportată la rata clasificării incorecte în grupul compușilor inactivi

AP = adevărat pozitivi (compuși activi clasificați de model ca fiind activi); AN = adevărat negativi;
FP = fals pozitivi (compuși inactivi clasificați de model ca fiind activi); FN = fals negativi

Parametrii prezentați în Tabelul 63 se pot folosi atât la diagnosticul unui model QSAR / QSPR

[54] Bolboacă SD, Jäntschi L. Binomial Distribution Sample Confidence Intervals Estimation for Positive and Negative Likelihood Ratio Medical Key Parameters. Annual Symposium on Biomedical and Health Informatics, American Informatics Medical Association, Bethseda, Special Issue: from Foundations to Applications to Policy (Proc. CD, October 22-26, Washington D.C., USA) 2005:66-70.

[55] Bolboacă SD. Binomial Distribution Sample Confidence Intervals Estimation 10. Relative Risk Reduction and RRR-like Expressions. Leonardo Electronic Journal of Practices and Technologies2005;6:60-75.

[56] Bolboacă SD, Jäntschi L. Optimized Confidence Intervals for Binomial Distributed Samples. International Journal of Pure and Applied Mathematics 2008;47(1):1-8.

[57] Jäntschi L, Bolboacă SD. Exact Probabilities and Confidence Limits for Binomial Samples: Applied to the Difference between Two Proportions. TheScientificWorldJOURNAL 2010;10:865-878.

[58] cât şi ca parametrii de evaluare a două modele diferite (ex. model MDF [16, 43, 59, 60] versus model MDFV).

Abilităţile de predicţie a modelului identificat pentru compuşii organici ce traversează bariera hemato-encefalică sunt prezentate în Tabelul 64.

Tabelul 64. Diagnosticul abilităţilor de clasificare a modelului MDFV: compuşi organici ce treversează bariera hemato-encefalică

Parametrul (abrevierea)	Set învăţare (n=81) [95%CI]	Set test (n=41) [95%CI]	Set extern (n=315) [95%CI]
Statistica χ^2 (valoarea p)	10.29 (0.0013)	7.75 (0.0054)	28.24 (p < 0.0001)
Φ	0.3564	0.4347	0.2994
Acurateţea (AC)	69.14 [58.53-78.37]	73.17 [58.32-84.77]	72.70 [67.58-77.39]
Rata erorii (ER)	30.86	26.83	27.30
Probabilitatea a prori de a fi			
- activ	0.482 [0.371-0.592]	0.463 [0.318-0.614]	0.302 [0.253-0.354]
- inactiv	0.519 [0.408-0.630]	0.537 [0.367-0.682]	0.698 [0.644-0.749]
Sensibilitate (Se)	64.10 [48.47-77.70]	84.21 [63.16-95.05]	42.11 [32.54-52.15]
Rata falşilor negativi (FNR)	35.90 [22.30-45.51]	15.79 [4.95-36.84]	57.89 [47.85-67.46]
Specificitate (Sp)	73.81 [59.20-85.15]	63.64 [42.87-81.04]	85.91 [80.80-89.98]
Rata falşilor pozitivi (FPR)	26.19 [14.86-40.80]	36.36 [0.1896-0.5712]	14.09 [10.02-19.20]
Predictivitatea pozitivă (PP)	69.44 [53.32-82.51]	66.67 [46.76-82.76]	56.34 [44.74-67.43]
Predictivitatea negativă (NP)	68.89 [54.49-80.89]	82.35 [59.63-97.48]	77.46 [72.59-81.80]
Probabilitatea de clasificare post-test ca şi			
- activ (PCA)	0.444 [0.340-0.553]	0.585 [0.433-0.726]	0.225 [0.177-0.281]
- inactiv (PCIC)	0.556 [0.447-0.660]	0.415 [0.274-0.567]	0.775 [0.7259-0.818]
Probabilitatea clasificării greşite ca şi compus			
- activ (PWCA)	0.306 [0.175-0.467]	0.333 [0.172-0.532]	0.437 [0.326-0.553]
- inactive (PWCI)	0.311 [0.191-0.455]	0.177 [0.055-0.404]	0.225 [0.177-0.281]
Rata şansei (OR)	5.03 [1.96-13.12]	9.33 [2.18-40.07]	4.43 [2.53-7.76]

Analzia rezultatelor prezentate în Tabelul 64 pune în evidenţă următoarele:

- Modelul MDFV are o acurateţe acceptabilă (~73% în setul extern) dată cu pregădere de abilităţi bune în clasificarea compuşilor inactivi.
- Sensibilitatea mică în setul extern indică faptul că modelul nu este util în clasificarea compuşilor activi, rezultatele fals negative având o pondere neacceptabilă.
- Rata falşilor pozitivi este semnificativ statistic mai mică în comparaţie cu rata falşilor negativi (intervalel de confidenţă nu se suprapun, ceea ce indică o diferenţă semnificativă statistic).

[58] Bolboacă SD, Jäntschi L. Diagnostic of a QSPR Model: Aqueous Solubility of Drug-Like Compounds. Studia Universitatis Babes-Bolyai Chemia 2010;LV(4):68-76.

[59] Jäntschi L, Bolboacă SD. Results from the Use of Molecular Descriptors Family on Structure Property/Activity Relationships International Journal of Molecular Sciences 2007;8(3):189-203.

[60] Bolboacă SD, Jäntschi L. Modelling the Inhibitory Activity on Carbonic Anhydrase I of Some Substituted Thiadiazoleand Thiadiazoline-Disulfonamides: Integration of Structure Information. Computer-Aided Chemical Engineering, Elsevier Netherlands & UK 2007;24:965-970.

- Probabilitatea clasificării greşite ca şi activ este semnificativ statistic mai mare comparativ cu probabilitatea clasificării greşite ca şi compus inactiv.

Parametrii şi indicatorii calulați permit diagnosticul corect şi complet al modelului matematic evaluat. Pentru a uşura activitatea de clasificare a fost realizat un portal care permite calcularea parametrilor şi indicatorilor propuşi (vezi Figura 35).

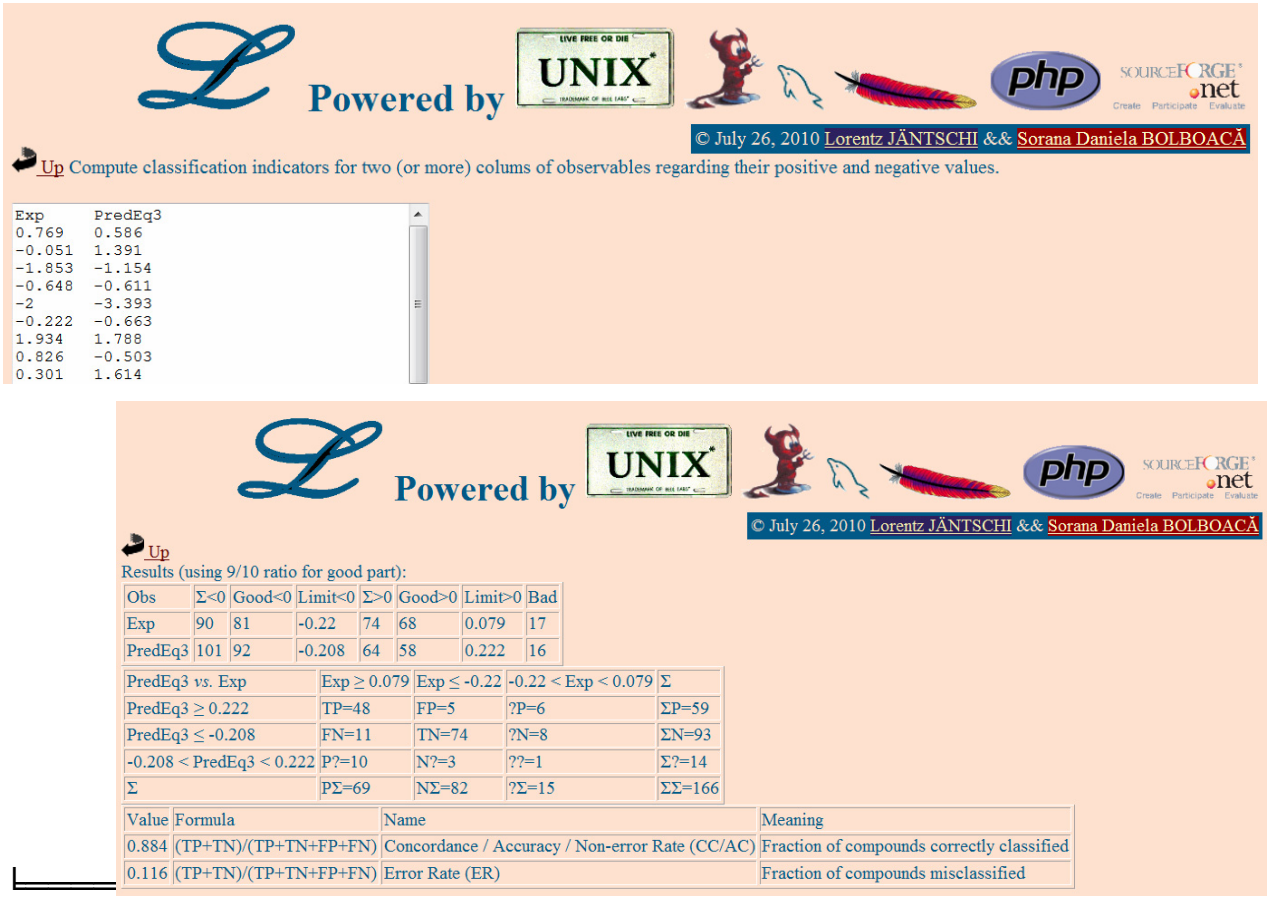


Figura 35. Mediu virtual de clasificare a modelelor QSAR/QSPR

Diseminarea rezultatelor

Publicații 2010

Articole ISI 2010:

- Bolboacă SD, Jäntschi L. Comparison of QSAR Performances on Carboquinone Derivatives. TheScientificWorldJOURNAL 2009;9(10):1148-1166.
- Bolboacă SD, Jäntschi L. Diagnostic of a QSPR Model: Aqueous Solubility of Drug-Like Compounds. Studia Universitatis Babes-Bolyai Chemia 2010;LV(4):68-76.

Articole BDI 2010:

- Bolboacă SD, Marta MM, Stoenoiu CE, Jäntschi L. Molecular Descriptors Family on Vertex Cutting: Relationships between Acelazolamide Structures and their Inhibitory Activity. Applied Medical Informatics 2009;25(3-4):65-74.
- Bolboacă SD, Marta MM, Jäntschi L. Binding affinity of triphenyl acrylonitriles to estrogen receptors: quantitative structure-activity relationships. Folia Medica 2010;52(3):37-45.

Impactul rezultalelor obținute

Principalele rezultate noi, originale obținute în cei trei ani de finanțare a proiectului şi impactul acestora au fost după cum urmează:

1. <u>Standardizarea metodologiilor statistice de evaluare statistică a observabilei:</u>

- Standardizarea metodei de raportare a rezultatelor în analiza de regresie simple şi multiple.

  Jäntschi L, Bolboacă SD, Diudea MV. Chromatographic Retention Times of Polychlorinated Biphenyls: from Structural Information to Property Characterization. International Journal of Molecular Sciences 2007;8(11):1125-1157.
- Analiza normalității datelor observate/experimentale (descriptiv & inferențial) & Identificarea şi îndepărtarea valorilor extreme (descriptiv & inferențial).

  Bolboacă SD, Jäntschi L. Distribution Fitting 3. Analysis under Normality Assumption. Bulletin of University of Agricultural Sciences and Veterinary Medicine Cluj-Napoca. Horticulture 2009;62(2):698-705.
- Metoda de clusterizare în analiza datelor experimentale

  Bolboacă SD, Jäntschi L. Mapping Cigarettes Similarities using Cluster Analysis Methods. International Journal of Environmental Research and Public Health 2007;4(3):233-242.

- Indicatori statistici de analiză a ciclicității

  Bolboacă SD, Jäntschi L. Cyclicity Analysis of Amino-Acids on Type I Collagen Chains. Bulletin of University of Agricultural Sciences and Veterinary Medicine Cluj-Napoca. Animal Science and Biotechnologies 2008;65(1-2):404-409.
- Metode de diagnostic a modelelor qSAR/qSPR prin utilizarea indicatorilor statistici.

  Bolboacă SD, Jäntschi L. Diagnostic of a QSPR Model: Aqueous Solubility of Drug-Like Compounds. Studia Universitatis Babes-Bolyai Chemia 2010;LV(4):68-76.

2. <u>Analiza relatiei structura-activitate pe clase de compusi biologic activi</u>

Jäntschi L, Bolboacă SD, Diudea MV. Chromatographic Retention Times of Polychlorinated Biphenyls: from Structural Information to Property Characterization. International Journal of Molecular Sciences 2007;8(11):1125-1157.

&

Bolboacă SD, Jäntschi L. Structure versus Biological Role of Substituted Thiadiazole- and Thiadiazoline- Disulfonamides. Studii şi Cercetări Ştiinţifice Universitatea Bacău Seria Biologie 2007;12(1):50-56.

&

Bolboacă SD, Jäntschi L. Structure-activity relationships of taxoids: a molecular descriptors family approach. Archives of Medical Science 2008;4(1):7-15.

&

Bolboacă SD, Jäntschi L. A Structural Informatics Study on Collagen. Chemical Biology & Drug Design 2008;71(2):173-179.

&

Bolboacă SD, Jäntschi L. Modelling Analysis of Amino Acids Hydrophobicity. MATCH Communications in Mathematical and in Computer Chemistry 2008;60(3):1021-1032.

3. <u>Dezvoltarea şi implementarea unei metode de modelare a relaţiilor structură-activitate MDFV</u>:

- structura moleculară 2D → 3D;
- ☼ → graf molecular;
- ☼ → reprezentare matriceală (topologie);
- ☼ → proprietăţi atomice;
- ☼ → matrice de adiacenţă;
- ☼ → matrice de distanţă;
- ☼ → fragmentare moleculară prin tăiere de vârf ;
- ☼ ...; ☼ → generarea modelului de structură pentru moleculă

Bolboacă SD, Jäntschi L. Comparison of QSAR Performances on Carboquinone Derivatives. TheScientificWorldJOURNAL 2009;9(10):1148-1166.

&

Bolboacă SD, Marta MM, Stoenoiu CE, Jäntschi L. Molecular Descriptors Family on Vertex Cutting: Relationships between Acelazolamide Structures and their Inhibitory Activity. Applied Medical Informatics 2009;25(3-4):65-740

&

Bolboacă SD, Marta MM, Jäntschi L. Binding affinity of triphenyl acrylonitriles to estrogen receptors: quantitative structure-activity relationships. Folia Medica 2010;52(3):37-45.

4. <u>Taieri de varfuri in grafuri</u>

Jäntschi L, Stoenoiu CE, Bolboacă S. A Formula for Vertex Cuts in b-Trees. International Journal of Pure and Applied Mathematics 2008;47(1):17-22.

<u>Evaluarea utilizarii polinoamelor caracteristice in analiza relatiilor structura-activitate/proprietate</u>

Jäntschi L, Bolboacă SD, Furdui CM. Characteristic and counting polynomials: modelling nonane isomers properties. Molecular Simulation 2009;35(3):220-227.

Măsuri ale dezordinii

Jäntschi L, Bolboacă SD. Entropy due to Fragmentation of Dendrimers, Surveys in Mathematics and its Applications 2009;4:169-177.

Impactul principal al rezultatelor obținute se poate sumariza astfel:

- academic: * formarea a doi cercetatori membrii ai echipei de cercetare (doctoranzi) prin participarea activa la activitatile proiectului si implicarea acestora in toate etapele de derulate a activitatilor; * metoda experimentala in silico cu utilitate educationala atata a studentilor cat si a tinerilor cercetatori.
- economic: dezvoltarea unei noi abordari si metode de caracterizare structura-activitate utila in caracterizarea diversilor compusi terapeutici - realizarea, implementarea si disponibilizarea unei noi metode experimentale in silico cu utilitate in identificarea si analiza a noi potentiali terapeutici activi.
- tehnologic: dezvoltarea unui portal online cu modele MDFV de analiza structura-activitate (http://l.academicdirect.org/Chemistry/SARs/MDFV/, acces autorizat).

Anexe

Anexa 1.

Test de evaluare a utilizabilității librăriei virtuale

Criteriul	Comentarii
Design-ul librăriei virtuale	
Mediul are o hartă care să prezinte secțiunile principale.	
Toate paginile sunt tipăribile iar paginile tipărite sunt acurate și complete.	
Fundalul paginii este alb sau în nuanțe pale cu contrast maxim față de text.	
Textura şi fundalul de tip imagine a fost utilizat doar atunci când nu interferă cu afişarea clară a informației.	
Informațiile din pagină sunt complete și la modificarea caracteristicilor și preferințelor de afișare.	
Terminologia este utilizată consecvent în librăria virtuală.	
Navigare	
Toate hyperlin-urile funcționează sunt funcționale.	
Culorile standard sunt utilizate pentru link-urile nevizualizate.	
Opțiunile de navigare sunt clare și consecvente.	
Link-urile sunt fără ambiguități, clare și specifice, respectiv cât se poate de specifice.	
Posibilitatea de a reveni la pagina anterioară este intuitivă și funcțională.	
Secvențiere din librăria virtuală	
Fiecare pagină are locul său bine stabilit în librăria virtuală.	
Fiecare pagină permite navigarea la alte pagini (ex. Prima pagină, ultima pagină, pagina anterioară, pagina următoare)	
Text	
Textul este structurat în aşa fel încât să permită citirea fără a naviga în pagină, chiar pentru cel mai mic ecran.	
Textul este scris cu respectarea stilului minimalist: compact dar util.	
Nu există mai mult de 2/3 stiluri de fonturi pe aceeași pagină.	
Fontul (stil, coloare, etc.) este uşor de citit atât la ecran cât şi în format tipărit.	
Textul este corect din punct de vedere gramatical.	

	Da	Nu	Nu ştiu	Nu se aplică
Utilitate				
Este util				
Permite control al navigării în librăria virtuală				
Permite realizarea uşoară a activității dorite				
Întruneşte nevoile mele				
Permite realizarea tuturor activităților pe care mă aşteptam să le facă				
Utilizare				
Este uşor de utilizat				
Este simplu de utilizat				
Are interfața prietenoasă				
Necesită urmarea a cât mai puțini paşi posibili pentru a realiza acțiunea dorită				
Este flexibilă				
Mediul virtual se poate utiliza fără efort				
Se poate utiliza şi fără instrucțiuni				
Nu am identificat nici o neconcordanță în timpul utilizării				
Se poate folosi cu succes şi la o nouă utilizare				
Satisfacție				
Sunt mulțumit de această librărie virtuală				
Aş recomnada această librărie virtuală prietenilor				
Funcționează aşa cum te-ai aşteptat să funcționeze				
Este plăcut la utilizare				

Anexa 2.

Test de evaluare a utilizabilității mediului virtual

Stimate participant,

Mulțumim pentru acceptarea participării la evaluarea librăriei virtuale. Testul va alea loc în data ……, în sala ……, orele …… .

Înainte de începerea testului vă rugăm să completați datele generale ale prezentului chestionar.

Date generale

Sexul ☐ F ☐ M

Vârsta ☐ 18-25 ani ☐ 26-39 ani ☐ 40-59 ani ☐ 60-74 ani ☐ 75+

În ultimele 6 luni ați mai participat la un studiu asemănător?

☐ Da ☐ Nu

Date profesionale

Funcția: ☐ Student ☐ Masterand ☐ Altele (specificați) .....

De cât timp ocupați această funcție?

Care din următoarele descriu cel mai înalt nivel al educației dvs?

- o Liceu (fără diplomă de bacalaureat)
- o Liceu (diplomă de bacalaureat)
- o Colegiu (specificați): …….
- o Facultate (cu diplomă de licență)
- o Masterat
- o Doctorat

Utilizați frecvent calculatorul? (Dacă răspunsu dvs. la această întrebare este NU chestionarul se încheie aici pentru dvs. Mulțumin pentru participare.)

☐ Da ☐ Nu

În afară de utilizarea căsuței de e-mail, pentru ce activități utilizați calculatorul?

- o jocuri/divertisment
- o ştiri/ziare/reviste
- o cumpărături/operațiuni bancare
- o design grafic/imagini digitale
- o programare/utilizare pachetului Office
- o Altele (specificați): ……

Expertiza în utilizarea calculatorului şi a Internetului

Câte ore pe săptămână petreceți în fața calculatorului?

☐ 0 – 10 ore ☐ 11-25 ore ☐ 26+ ore

Ce platformă de calculator folosiți deobicei?

☐ Mac ☐ Windows ☐ Altele (specificați): .....

Ce browser de Internet folosiți deobicei?

☐ Firefox ☐ Internet Eplorer ☐ Altele (specificați): .....

Cunoştințe de specialitate (modelare moleculară / relații structură activitate)

Abilități lingvistice (ex. Română (maternă) – Engleză (bine)): ……

Auto-evaluarea expertizei în domeniul modelare moleculară / relații structură-activitate (scala de la 0 = nu am cunoştințe la 10 = expert în domeniu):……

Cunoştințe tehnice (ex. Programare, design web, cercetător, etc.): ……...

Ani de experință: ……..

Utilizator al unor pagini / programe similare (specificați): ……

Informații personale (* = opțional)

- Prenume, nume: ……..
- Adresa*: ……
- Județul de reședință*: ……
- Telefon*: ……
- E-mail: ……..

Evaluarea librăriei virtuale va avea loc în data ....., la orele ....., în locația ..... . Sesiunea de evaluare va fi anunțată prin e-mail cu câteva zile înainte de data stabilită.