

ET108/2006 – Et. Unică/2006 – Lucrare în extenso

Ministerul Educației și Cercetării
Universitatea Tehnică din Cluj-Napoca
Facultatea de Știința și Ingineria Materialelor
Catedra de Chimie

Programul:	Cercetare de Excelență
Modul: II	Proiecte de Dezvoltare a Resurselor Umane pentru Cercetare
Tipul proiectului:	Proiecte de cercetare de excelenta pentru tinerii cercetători
Cod proiect:	ET108/2006
Denumirea proiectului:	Procese la Interfața fazelor: Modelare matematica, Optimizare numerica, Implementare web, cu aplicatii în separarea și caracterizarea seriilor de compuși chimici biologic activi
Etapă:	Unică/2006

Lucrare în extenso

Cuprins

<i>Etape, obiective și activități</i>	<i>– pag. 02</i>
<i>Activități și rezultate</i>	<i>– pag. 03</i>
<i>Publicații</i>	<i>– pag. 09</i>
<i>Toxicitatea compușilor metal-alchil</i>	<i>– pag. 10</i>
<i>Analiza de corelație cu Pearson, Kendall și Spearman</i>	<i>– pag. 20</i>
<i>Activitatea antialergică a benzamidelor</i>	<i>– pag. 41</i>
<i>Optimizarea fazei mobile la amestecuri de solvenți</i>	<i>– pag. 51</i>
<i>Cromatografia planară</i>	<i>– pag. 63</i>
<i>Concluzii</i>	<i>– pag. 73</i>

Etape, obiective și activități

Pentru etapele anului 2006 obiectivele planificate au fost:

- *Managementul resurselor: umane, materiale, cunoastere (Etapa Unică);*
- *Colectarea datelor experimentale (cromatografie pe strat subtire) și dobândirea de cunoștințe (Etapa Unică).*

Activitățile prevăzute a se desfășura au fost:

- *Documentare folosind resursele Springer Verlag (Etapa Unică);*
- *Elaborare specificatii pentru mediile de experimentare cantitativa si pentru echipamentele de achizitionat (Etapa Unică);*
- *Separarea prin cromatografie pe strat subtire a steroizilor (Etapa Unică);*
- *Participări la manifestări științifice și dobândirea de competențe complementare (Etapa Unică).*

Activitățile au fost realizate și obiectivul planificat a fost atins.

Activități și rezultate

- **Documentare folosind resursele Springer Verlag (Etapa Unică):**

Următoarele lucrări au fost selectate, obținute și pe baza lor a fost actualizată documentația pentru modelarea proceselor la interfața fazelor:

- Sven Groß, Volker Reichelt and Arnold Reusken, A finite element based level set method for two-phase incompressible flows, *Computing and Visualization in Science*, DOI:10.1007/s00791-006-0024-y (Published online: 20 October 2006);
- Myungjoo Kang, Hyeseon Shim and Stanley Osher, Level Set Based Simulations of Two-Phase Oil–Water Flows in Pipes, *Journal of Scientific Computing*, DOI:10.1007/s10915-006-9103-y (Published online: 17 October 2006);
- J. Jodlbauer, P. Zöllner and W. Lindner, Determination of zeranol, taleranol, zearalenone, α - and β -zearalenol in urine and tissue by high-performance liquid chromatography-tandem mass spectrometry, *Chromatographia*, Volume 51, Numbers 11-12 / June, 2000, DOI:10.1007/BF02505405 (Accepted: 4 January 2000, Online Date 12 October, 2006).

Documentația detaliată se găsește în *lucrarea în extenso*.

- **Elaborare specificații pentru mediile de experimentare cantitativa si pentru echipamentele de achizitionat (Etapa Unică):**

Analiza cantitativă este bazată pe măsurarea unei proprietăți care este corelată direct sau indirect, cu cantitatea de constituent ce trebuie determinată dintr-o probă. În mod ideal, nici un constituent, în afară de cel căutat, nu ar trebui să contribuie la măsurătoarea efectuată. Din nefericire, o astfel de selectivitate este rareori întâlnită.

Pentru a proceda la o analiză cantitativă, trebuie urmate o serie de etape:

1. Obținerea unei probe semnificative prin metode statistice;
2. Prepararea probei;
3. Stabilirea procedurii analitice în funcție de:
 - a. Metode:
 - i. chimice;
 - ii. fizice cu sau fără schimbări în substanță;
 - b. Condiții:

ET108/2006 – Et. Unică/2006 – Lucrare in extenso

- i. determinate de metoda de analiză aleasă;
 - ii. determinate de substanța cercetată;
- c. Cerințe:
- i. rapiditate, exactitate, costuri;
 - ii. posibilitatea de amortizare;
4. Evaluarea și interpretarea rezultatelor.

Practic, după natura analizei, există 7 tipuri de metode de analiză: (1) gravimetrice; (2) volumetrice; (3) optice; (4) electrice; (5) de separare; (6) termice; (7) de rezonanță. În general, (1) și (2) sunt metode chimice, iar (3-7) sunt instrumentale (bazate pe relații între o proprietate caracteristică și compoziția probei). Adeseori, în analiză se cuplează după sau mai multe dintre aceste procedee de bază. O altă clasificare a metodelor de analiză se poate face după implicarea componentilor în reacții chimice, în metode stoechiometrice și metode nestoechiometrice.

Separarea diferitelor substanțe dintr-un amestec constituie una dintre cele mai importante probleme ale chimiei analitice. Metoda cromatografică se bazează pe repetarea echilibrului de repartiție a componentelor unui amestec între o fază mobilă și una staționară. Datorită diferențelor în repartiție are loc deplasarea, cu viteză diferită, a componentelor purtate de faza mobilă de-a lungul fazei staționare.

În general, metodele de separare cromatografice se împart în două categorii: în prima intră cele care se bazează pe interacțiunea diferită a componentilor cu faza staționară (repartiție, adsorbție, schimb ionic și afinitate), iar în a doua cele care se bazează pe mărimea diferită a componentilor (excluziunea sterică). Cromatografia pe strat subțire și pe hârtie posedă două avantaje:

- sunt metode de separare cu costuri reduse;
- separarea se produce într-un mod similar cu cea de lichide de înaltă performanță (TLC - thin layer chromatography; HPLC - high pressure liquid chromatography; HPTLC - metoda integratoare - cuplarea TLC (analiza calitativă, alegerea solvenților potriviți, alegerea compoziției optime a fazei mobile) cu HPLC (analiza cantitativă folosind informația obținută prin TLC).

Documentația detaliată se găsește în lucrarea [Horea Iustin NAȘCU, Lorentz JĂNTSCHI, Chimie analitică și instrumentală (in Romanian), AcademicDirect & AcademicPres, Internet & Cluj-Napoca, 320 p., ISBN(10) 973-744-046-3 & ISBN(13) 978-

ET108/2006 – Et. Unică/2006 – Lucrare in extenso

973-744-046-4 (AcademicDirect) && ISBN (10)973-86211-4-3 & ISBN(13) 978-973-86211-4-5 (AcademicPres), 2006 (November), în curs de apariție] și în *lucrarea în extenso*.

Echipamente de achiziționat: calculator tip server pentru *optimizarea numerică a proceselor la interfața fazelor*.

Problema majoră în situația actuală cu echipamentele existente este insuficiența memoriei de calcul. Astfel, calculatorul folosit acum pentru calcule, după ce își încarcă serviciile mai posedă foarte puțină memorie de lucru:

193.226.7.211

Up Time
7:09PM up 7 days, 2:08, 0 users, load averages: 0.02, 0.02, 0.00

System Information
Copyright (c) 1992-2005 The FreeBSD Project.
Copyright (c) 1979, 1980, 1983, 1986, 1988, 1989, 1991, 1992, 1993, 1994
The Regents of the University of California. All rights reserved.
FreeBSD 5.4-PRERELEASE #0: Mon Apr 4 10:14:55 EEST 2005
root@academicdirect.ro:/usr/src/sys/i386/compile/NS
Timecounter "i8254" frequency 1193182 Hz quality 0

CPU	user	nice	system	interrupt	idle
Pentium II/Pentium II Xeon/Celeron (400.91-MHz 686-class CPU)	0.0%	0.0%	0.4%	1.2%	98.4%

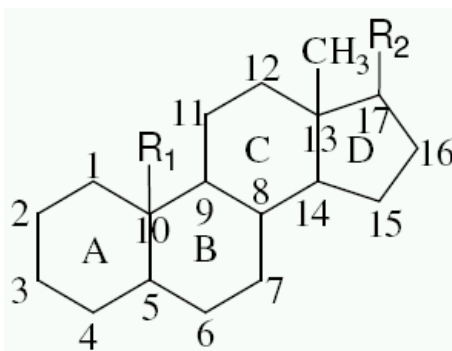
Type = "GenuineIntel" Id = 0x652 Stepping = 2
Features = [FPU, VME, DE, PSE, TSC, MSR, PAE, MCE, CX8, SEP, MTRR, PGE, MCA, CMOV, PAT, PSE36, MMX, FXSR]
cpu0: [ACPI CPU] on acpi0
npx0: [math processor] on motherboard
acpi0: [PTLTD RSDT] on motherboard
Timecounter "i8254" frequency 1193182 Hz quality 0
Timecounter "ACPI-safe" frequency 3579545 Hz quality 1000
Timecounter "TSC" frequency 400911256 Hz quality 800
Timecounters tick every 10.000 msec

RAM				Active	Inact	Wired	Cache	Buf	Free
real memory = 268435456 (256 MB)				108M	58M	68M	12M	35M	656K
avail memory = 256106496 (244 MB)									
cache	current	peak	max						
mbufs	129	-	-						
mbuf	128	9024	-						
sbufs	-	536	2512						

Soluția propusă este achiziționarea unui calculator capabil să suporte mai multă memorie, calculatorul existent (Pentium II/Pentium II Xeon/Celeron) fiind deja la capacitatea maximă de memorie suportată (256 MB).

• **Separarea prin cromatografie pe strat subtire a steroizilor (Etapa Unică):**

Formula structurală generală a steroizilor este:



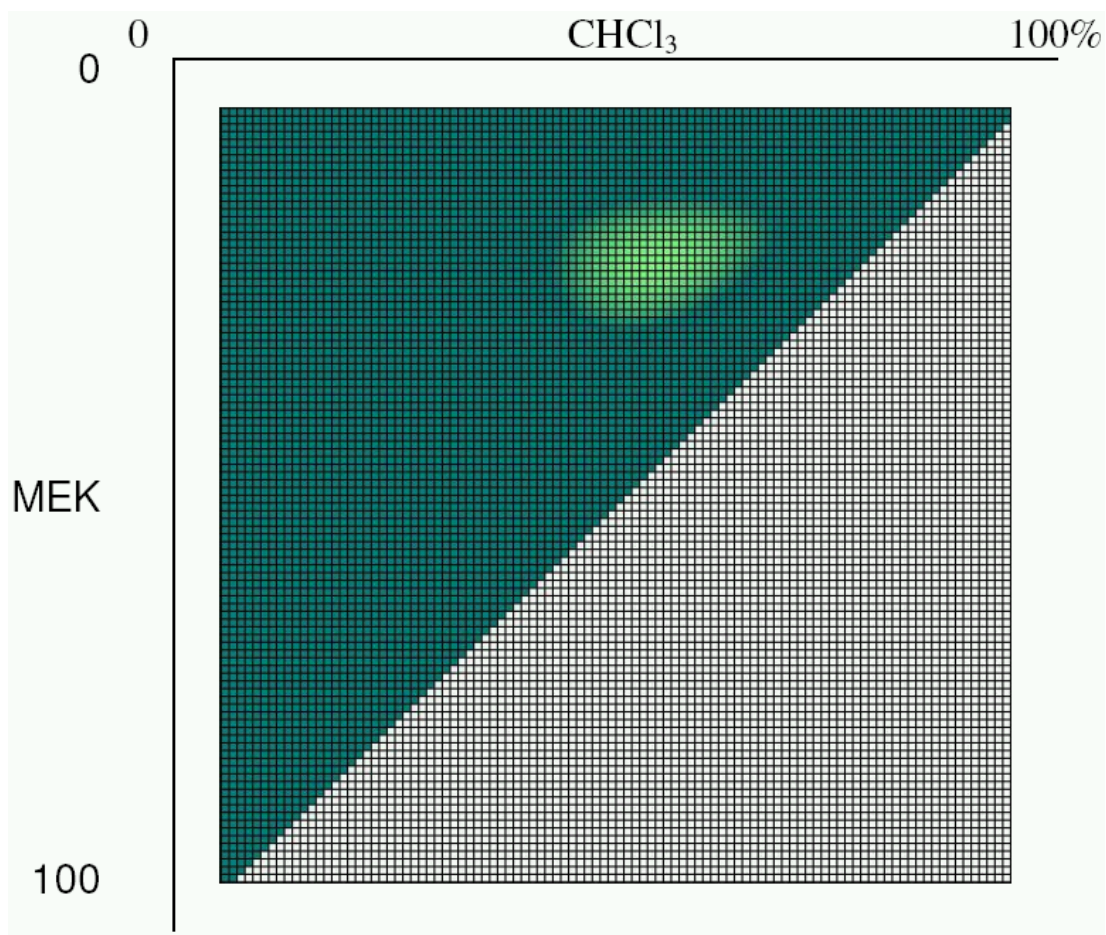
Compușii din clasa steroizilor diferă prin natura substituenților, prin gradul de nesaturare al nucleului tetraciclic, și prin natura lanțurilor R1 și R2. Clasa de izomeri studiată conține compuși cu structură foarte asemănătoare, diferind doar prin numărul și poziția radicalilor hidroxil.

Alegerea fazei mobile și optimizarea compoziției acesteia sunt foarte importante, deoarece separarea cromatografică este dificil de realizat (Sherma J., Handbook of Thin-Layer Chromatography, 3rd Edition, M. Dekker, New York, 2003, p. 913-933).

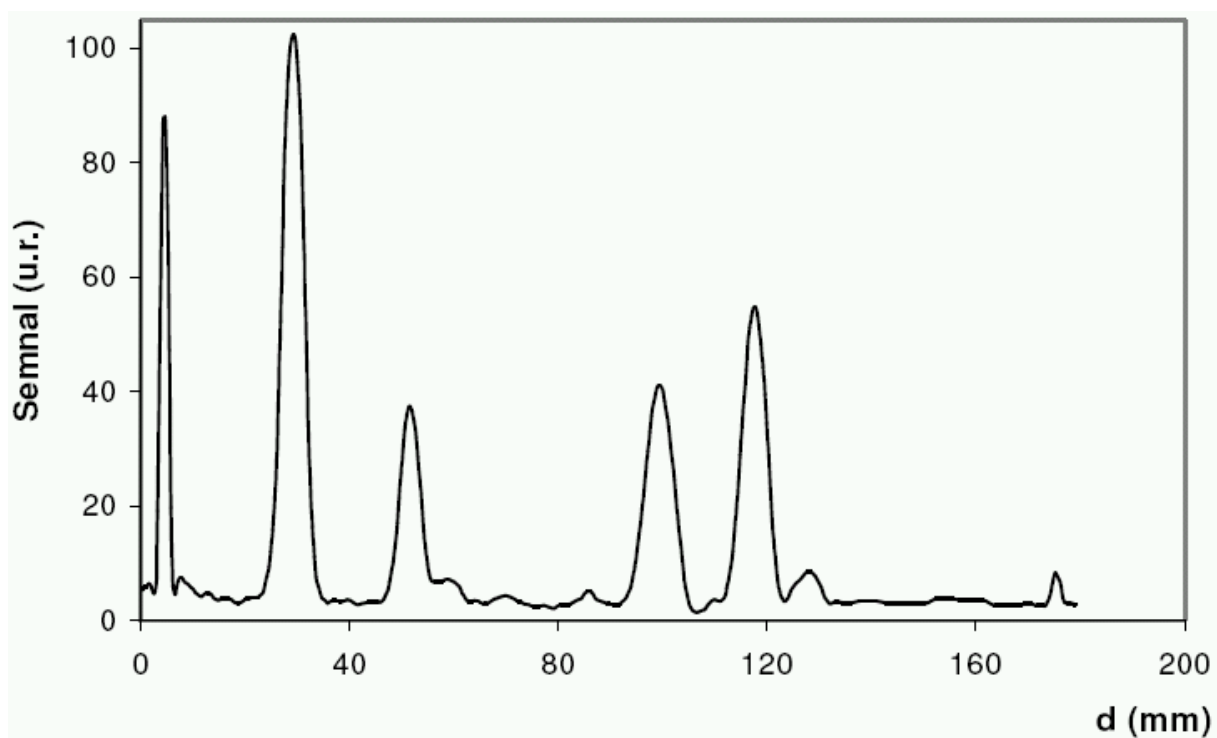
S-a ales faza mobilă folosind metoda taxonomiei numerice (Castro D., Pujalte M.J., Lopez-Cortes L., Garay E., Borrego J.J., Vibrios isolated from the cultured manila clam (*Ruditapes philippinarum*): Numerical taxonomy and antibacterial activities, Journal of Applied Microbiology 93 (3), 2002, pp. 438-447).

S-a optimizat faza mobilă folosind funcții obiectiv (Cimpoi C., Jantschi L., Hodisan T., A new mathematical model for the optimization of the mobile phase composition in HPTLC and the comparison with other models, Journal of Liquid Chromatography and Related Technologies 22 (10), 1999, pp. 1429-1441).

Compoziția optimă a fazei mobile a fost obținută ca fiind: cloroform – ciclohexan - metil-etil-cetonă 54:18:28 v/v (figura următoare).



Separarea cromatografică folosind compoziția de mai sus este:



ET108/2006 – Et. Unică/2006 – Lucrare in extenso

Cromatograma de mai sus dovedește separarea compușilor steroizi (5α androstan- 3β -ol; 5α -androstan- 3α -ol; 5α -androstan- 17β -ol; 5β -androstan- $3\alpha,17\beta$ -diol; 5β -androstan- $3\beta,17\beta$ -diol).

• **Participări la manifestări științifice și dobândirea de competențe complementare (Etapa Unică):**

S-a participat cu lucrări științifice la următoarele conferințe:

1. Lorentz JÄNTSCHI, Sorana-Daniela BOLBOACĂ, Mihaela Ligia UNGURESAN, *Mobile Phase Optimization in Three Solvents High Performance Thin-Layer Chromatography: Methodology and Evaluation*, **6th European Conference on Computational Chemistry**, September 3-7, Book of abstract, **Slovakia, 2006**, Bratislava.
http://lori.academicdirect.org/conferences/6ECCC_HPTL.pdf
2. Sorana BOLBOACĂ, Lorentz JÄNTSCHI, *End-of-Course Examination Methodology on Physical-Chemistry Subject: A Case Study*, **1st European Chemistry Congress**, Teaching Chemistry - Past, Present, and Future, 2006 August 27-31, Budapest, Hungary.
http://lori.academicdirect.org/conferences/EuChemC2006_2_list.pdf
http://lori.academicdirect.org/conferences/EuChemC2006_2.pdf
3. Lorentz JÄNTSCHI, Sorana-Daniela BOLBOACĂ, *Processes Kinetics Modeling: A Numerical Study*, **Ecomaterials and Processes: Characterization and Metrology**, April 19 - 21, 2007, St. Kirik, Plovdiv, Bulgaria.
http://lori.academicdirect.org/conferences/MetEcoMat_Abs_Jantschi&Bolboaca.pdf

Publicații

1. Sorana Daniela BOLBOACĂ, Lorentz JĂNTSCHI, *Modeling of Structure-Toxicity Relationship of Alkyl Metal Compounds by Integration of Complex Structural Information*, **Therapeutics: Pharmacology and Clinical Toxicology**, RP Press, Bucharest, Issue X(1), 110-114, 2006, ISSN 1583-0012, [N.U.R.C. Class C](http://lori.academicdirect.org/articles/Terapeutica.pdf), <http://lori.academicdirect.org/articles/Terapeutica.pdf>
2. Sorana BOLBOACĂ, Lorentz JĂNTSCHI, *Pearson Versus Spearman, Kendall's Tau Correlation Analysis on Structure-Activity Relationships of Biologic Active Compounds*, **Leonardo Journal of Sciences**, AcademicDirect, Internet, [Issue 9](http://ljs.academicdirect.ro/A09/179_200.pdf), 179-200, 2006, ISSN 1583-0233, [DOAJ](http://ljs.academicdirect.ro/A09/179_200.pdf), http://ljs.academicdirect.ro/A09/179_200.pdf
3. Lorentz JĂNTSCHI, Sorana Daniela BOLBOACĂ, *Antiallergic Activity of Substituted Benzamides: Characterization, Estimation and Prediction*, **Clujul Medical**, trimisă spre publicare, http://lori.academicdirect.org/articles/Paper_19654_ClujulMedical_.pdf
4. Lorentz JĂNTSCHI, Sorana Daniela BOLBOACĂ, Mihaela Ligia UNGUREȘAN, *Mobile Phase Optimization in Three Solvents High Performance Thin-Layer Chromatography: Methodology and Evaluation*, **International Journal of Quantum Chemistry**, trimisă spre publicare, http://lori.academicdirect.org/articles/IJQC_6ECCC_MobilePhaseOpt.pdf
5. Horea Iustin NAȘCU, Lorentz JĂNTSCHI, *Chimie analitică și instrumentală (in Romanian)*, AcademicDirect & AcademicPres, **Internet & Cluj-Napoca**, 320 p., ISBN(10) 973-744-046-3 & ISBN(13) 978-973-744-046-4 (AcademicDirect) && ISBN (10)973-86211-4-3 & ISBN(13) 978-973-86211-4-5 (AcademicPres), **2006 (November)**, în curs de apariție, <http://ph.academicdirect.org/>

Modeling of Structure-Toxicity Relationship of Alkyl Metal Compounds by Integration of Complex Structural Information

Abstract

Alkyl metal compounds are ubiquitously toxins, known to be immunotoxic and/or neurotoxic. Starting with the complex information offered by the molecular structure of ten alkyl metal compounds, their toxicity was modeled by applying an original methodology. The obtained models were evaluated and validated by means of correlation coefficients, statistical parameters of models, and by cross-validation correlation coefficients. The model with the highest predictive ability ($r^2_{cv(100)} = 0.9965$) shows that the toxicity of alkyl metal compounds is alike geometrical and topological, depends on the mass and partial charge, and is related with mechanical work of property and its field.

Keywords

Structure-Activity Relationships (SAR); Alkyl metal compounds; Complex information integration; Molecular Descriptors Family (MDF)

Modelarea relației structură-toxicitate a compușilor metal alchil prin integrarea informațiilor structurale complexe

Rezumat

Compușii metal alchil sunt toxine ubicuitare cu proprietăți imunotoxice sau/și neurotoxice. Pornind de la informațiile complexe oferite de structura moleculară a zece compuși metal alchil, toxicitatea acestora a fost modelată prin aplicarea unei metodologii originale. Evaluarea și validarea modelelor obținute s-a realizat prin studiul coeficienților de corelație, a parametrilor statistici asociați modelelor și a coeficienților de corelație încrucișată. Modelul cu capacitatea cea mai bună de predicție ($r^2_{cv(100)} = 0.9965$) evidențiază că, toxicitatea compușilor metal alchil este deopotrivă de natură topologică și geometrică, depinde de masă și sarcina parțială, și este în relație directă cu lucrul și câmpul proprietății.

Cuvinte cheie

Relații Structură-Activitate (SAR); Compuși metal alchil; Integrarea informațiilor complexe; Familia Descriptorilor Moleculari (MDF)

Introducere

Compușii metal alchil sunt toxine ubicuitare cu proprietăți fungicide [Chandra & all, 1987], erbicide [Crowe, 2004], insecticide și bacteriostatice [Mehrotra&Singh, 2004]. Se cunoaște astăzi că, expunerea la compuși metal alchil este imunotoxică [Ade & all, 1996; Dacasto & all, 2001] sau/și neurotoxică [Aschner & Aschner, 1992]. De exemplu, expunerea la trietil plumb produce modificări histologice la nivelul hipocampului, în timp ce expunerea la trimetil plumb determină modificări histologice la nivelul măduvei spinării [Walsh & all, 1986]. Importanța studierii compușilor metal alchil rezidă astfel din efectele pe care aceștia le au asupra mediului și a organismului uman. O serie de cercetători au studiat corelația dintre acești compuși și efectele lor biochimice, stabilind existența corelației dintre structură și toxicitate [Eng & all, 1991; Laughlin & all, 1984].

Pornind de la informațiile complexe oferite de structura moleculară a zece compuși metal alchil, scopul cercetării a fost de a modela toxicitatea acestora prin aplicarea unei metodologii originale și de a evalua abilitățile modelelor SAR obținute în predicția toxicității compușilor metal alchil.

Material și Metodă

Un eșantion de 10 compuși metal alchil au fost incluși în studiu. Denumirea compușilor, abrevierea lor și toxicitatea măsurată (exprimată ca $\log(LC_{50})$ - $\mu\text{mol/l}$ [Ade & all, 1996]) sunt în tabelul I.

Tabelul I. Denumirea compușilor metal alchil, abrevierea și toxicitatea măsurată

Denumire compus	Abreviere	$\log LC_{50}$ ($\mu\text{mol/l}$)
Dibutil staniu	DBS	1.8457
Dietil plumb	DEP	1.8331
Tributil staniu	TBS	0.3979
Trietil plumb	TEP	1.5211
Trietil staniu	TES	2.1973
Trimetil plumb	TMP	2.4907
Trimetil staniu	TMS	3.4419
Tripentil staniu	TPS	0.5441
Trifenil plumb	TFP	0.5315
Tripetil staniu	TPrS	0.7924

Toxicitatea celor zece compuși metal alchil a fost modelată prin integrarea informațiilor complexe oferite de structura moleculară a compușilor, prin aplicarea unei metodologii proprii, folosind Familia Descriptorilor Moleculari. Metodologia aplicată în modelare a

cuprins *șase etape* [Jäntschi, 2005]. Fiecărei etape îi corespunde unul sau mai multe programe PHP (Hypertext Pre-Processor). Toate calculele au fost realizate pe serverul <http://vl.academicdirect.org>.

Prima etapă a modelării a fost destinată reprezentării tridimensionale a compușilor metal alchil care s-a realizat cu programul HyperChem [HyperChem, 2005]. În **etapa a doua** s-a creat fișierul care conține toxicitatea măsurată ($\log(LC_{50})$, exprimată în $\mu\text{mol/l}$) a compușilor metal alchil. Structurile tridimensionale ale compușilor și toxicitatea măsurată asociată acestora au intrat în **etapa a treia**, de generare, calculare și filtrare a membrilor familie descriptorilor moleculari [Diudea & all, 2001; Jäntschi & all, 2000]. În generarea listei familiei de descriptori moleculari s-au luat în considerare următoarele caracteristici, regăsite în denumirea fiecărui descriptor: geometria sau topologia moleculei (litera șapte în denumirea descriptorului), proprietatea atomică (masa atomică relativă, sarcina atomică parțială, cardinalitatea, electronegativitatea, electronegativitatea de grup, numărul de atomi de hidrogen legați direct - litera șase), descriptorul de interacțiune (litera cinci), modelul de suprapunere a interacțiunii descriptorilor (litera patra), metoda de fragmentare moleculară (litera trei), metoda de cumulare a proprietăților de fragmentare (litera a doua) și procedura de linearizare aplicată în generarea descriptorului global molecular (prima literă). Odată generată lista descriptorilor moleculari, s-a trecut la **etapa a patra**, de căutare și identificare a celor mai semnificative modele mono- și/sau multi-variate SAR. Cele mai performante modele SAR identificate au intrat în **etapa cinci**, de validare, etapă în care fiecare compus metal alchil a fost exclus pe rând din analiză, s-au recalculat valorile coeficienților, și s-a prezis pe baza modelului obținut toxicitatea compusului exclus. Analiza de validare a modelelor SAR a avut ca rezultat calcularea coeficientului de validare încrucișată (r^2_{cv}), parametrului Fisher și semnificația analizei de validare încrucișată (F_{pred} , $p_{\text{pred}}(\%)$) [Leave-one-out Analysis, 2005]. **Etapa șase** a constat în analiza modelelor SAR, analiză care s-a realizat pe baza următoarelor criterii: coeficientul de determinare, probabilitatea de model SAR greșit, coeficientul de validare încrucișată, probabilitatea unui model de validare încrucișată greșit și stabilitatea modelului dată de diferența dintre coeficientul de determinare al modelului SAR și coeficientul de validare încrucișată cu valoare cât mai mică. Tot în această etapă s-a realizat compararea performanțelor modelului mono-variat și a celui bi-variat prin analiza de corelare a coeficienților de corelație (testul Steiger [Steiger, 1980]).

Rezultate

În urma integrării cunoștințelor complexe, au fost identificate cele mai performante modele mono- și bi-variate SAR. Cele două modele SAR și statisticile asociate acestora sunt în tabelul II. În tabelul II s-a notat cu \hat{Y} toxicitatea estimată, și cuprinde valorile coeficienților de corelație dintre fiecare descriptor în parte și toxicitatea măsurată ($r(\text{desc}, \log(\text{LC}_{50}))$), eroarea standard (ErStd), valorile extreme ale intervalului de încredere asociat coeficienților (CI_I 95% - valoarea inferioară, CI_S 95% - valoarea superioară), parametrul Student (t) și probabilitatea testului Student (p_t).

Tabelul II. Modele SAR și statisticile de regresie

$r(\text{descr}, \log(\text{LC}_{50}))$	CI_I 95%	CI_S 95%	ErStd	t	p_t (%)	
Mono-variat: $\hat{Y} = 0.335 + 0.252 \cdot iFDmCg$						
Intercept	-	0.101	0.569	0.102	3.298	$1.09 \cdot 10^0$
iFDmCg	0.9830	0.213	0.290	0.017	15.165	$3.54 \cdot 10^{-5}$
Bi-variat: $\hat{Y} = 11.210 - 1.639 \cdot IHDmWMt + 0.372 \cdot LAMrEQg$						
Intercept	-	10.413	12.008	0.337	33.239	$5.78 \cdot 10^{-7}$
IHDmWMt	-0.8396	-1.801	-1.478	0.068	-24.009	$5.53 \cdot 10^{-6}$
LAMrEQg	0.9246	0.346	0.397	0.011	34.348	$4.60 \cdot 10^{-9}$

Abrevierea compușilor metal alchil, valorile descriptorilor moleculari folosiți în modelul mono- și bi-variat și toxicitatea estimată (toxicitatea estimată cu modelul mono-variat = $\hat{Y}_{\text{mono-v}}$, toxicitatea estimată cu modelul bi-variat = $\hat{Y}_{\text{bi-v}}$) sunt prezentate în tabelul III.

Tabelul III. Abrevierea compușilor, valorile descriptorilor moleculari și toxicitatea estimată cu ajutorul acestora (cele mai bune modele obținute)

Abreviere	Mono-variat		Bi-variat		
	iFDmCg	$\hat{Y}_{\text{mono-v}}$	IHDmWMt	LAMrEQg	$\hat{Y}_{\text{bi-v}}$
DBS	6.3257	1.9278	5.4095	-1.4056	1.8194
DEP	5.2300	1.6518	5.1908	-2.2845	1.8513
TBS	1.1821	0.6325	5.3537	-5.2584	0.4793
TEP	4.2832	1.4134	5.1084	-3.5195	1.5275
TES	7.0121	2.1006	5.1252	-1.7060	2.1738
TMP	9.3584	2.6915	4.7398	-2.5899	2.4772
TMS	11.868	3.3236	4.7446	0.0947	3.4668
TPS	0.4109	0.4382	5.4053	-5.0309	0.4793
TFP	0.0010	0.3350	5.6086	-3.9080	0.5632
TPrS	2.9605	1.0803	5.2630	-4.9172	0.7548

Statisticile asociate modelului mono- și bi-variat, exprimate prin coeficientul de corelație (r), coeficientul de determinare (r^2), coeficientului de determinare ajustat (r^2_{adj}), eroarea standard a modelului (s_{est}), parametrul Fisher (F_{est}), probabilitatea unui model de regresie greșit exprimată procentual ($p_{\text{est}}(\%)$), coeficientul de validare încrucișată ($r^2_{\text{cv}(\text{loo})}$), parametru Fisher

ET108/2006 – Et. Unică/2006 – Lucrare in extenso

(F_{pred}) și probabilitatea unui model greșit de validare încrucișată ($p_{\text{pred}}(\%)$), eroarea standard a analizei de validare încrucișată (s_{loo}) și diferența dintre coeficientul de determinare și coeficientul de validare încrucișată ($r^2 - r^2_{\text{cv(loo)}}$) sunt în tabelul IV.

Tabelul IV. Statistici asociate modelelor SAR

Caracteristica	Model SAR	
	Mono-variat	Bi-variat
r	0.9830	0.9991
r^2	0.9664	0.9983
r^2_{adj}	0.9622	0.9978
S_{est}	0.1945	0.0473
F_{est}	230	2008
$p_{\text{est}}(\%)$	$3.54 \cdot 10^{-5}$	$2.20 \cdot 10^{-8}$
$r^2_{\text{cv(loo)}}$	0.9466	0.9965
F_{pred}	141	989
$p_{\text{pred}}(\%)$	$2.29 \cdot 10^{-4}$	$2.61 \cdot 10^{-7}$
S_{loo}	0.2454	0.0673
$r^2 - r^2_{\text{cv(loo)}}$	0.0198	0.0018

Reprezentarea grafică a dependenței dintre toxicitatea compușilor metal alchil și structura acestora exprimată prin modelul bi-variat este în figura 1.

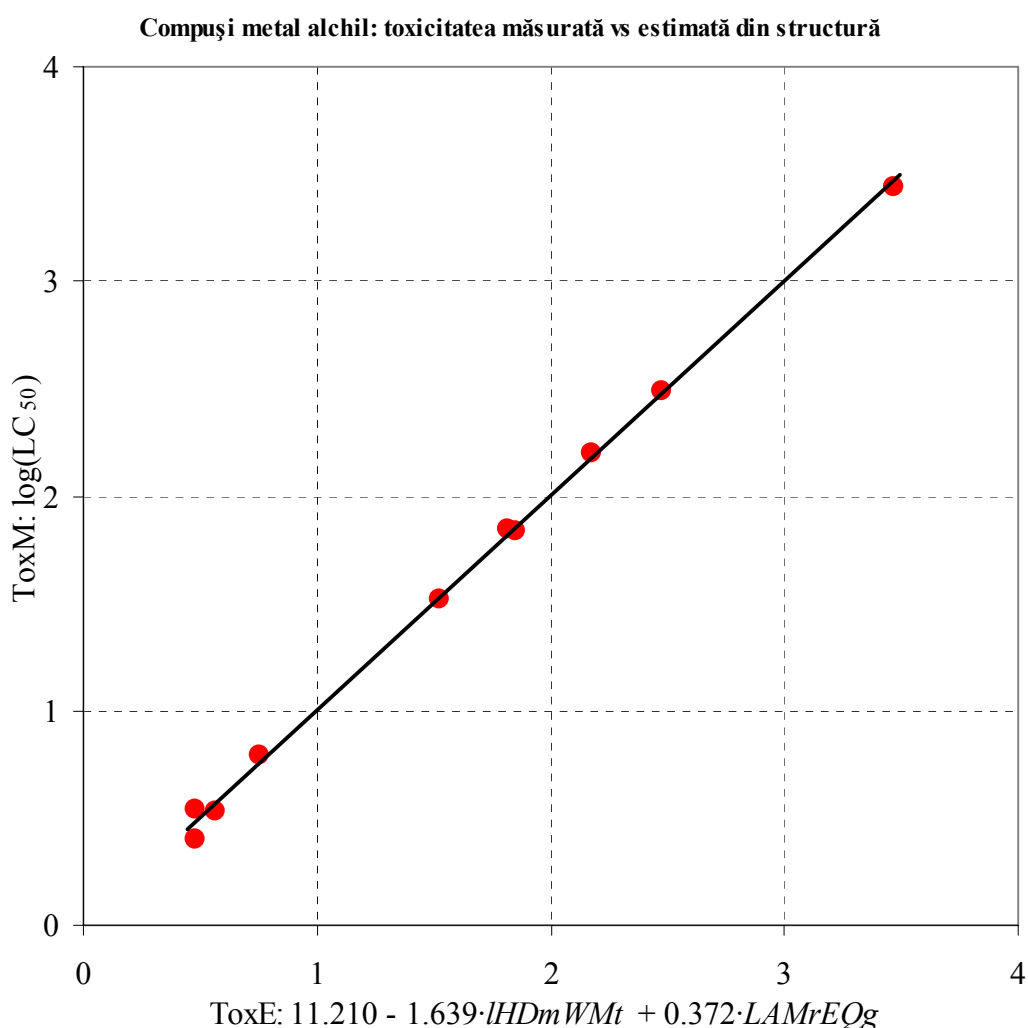


Figura 1. Toxicitatea măsurată (ToxM) vs estimată (ToxE) cu modelul bi-variant

Rezultatele comparării dintre modelele SAR obținute prin integrarea informațiilor complexe ale structurii compușilor metal alchil, respectiv între modelul mono-variant și cel bi-variant sunt în tabelul V. În tabelul V, s-a notat cu $r(\log(LC_{50}), \hat{Y}_{bi-v})$ coeficientul de corelație dintre toxicitatea măsurată ($\log(LC_{50})$) și cea estimată cu modelul bi-variant (\hat{Y}_{bi-v}), cu $r(\log(LC_{50}), \hat{Y}_{mono-v})$ coeficientul de corelație între toxicitatea măsurată și cea estimată cu modelul mono-variant (\hat{Y}_{mono-v}), cu $r(\hat{Y}_{bi-v}, \hat{Y}_{mono-v})$ coeficientul de corelație între toxicitatea estimată cu modelul bi-variant și cea estimată cu modelul mono-variant, cu Z parametrul punctual rezultat din aplicarea testului Steiger și cu p_Z probabilitatea de coincidență a coeficienților de corelație obținuți de cele două modele.

Tabelul V. Rezultate ale comparării modelului bi-variant cu modelul mono-variant

Testul Steiger	
Caracteristica	Valoare

$r(\log(LC_{50}), \hat{Y}_{bi-v})$	0.9991
$r(\log(LC_{50}), \hat{Y}_{mono-v})$	0.9830
$r(\hat{Y}_{bi-v}, \hat{Y}_{mono-v})$	0.9821
Z	3.9038
p_Z (probabilitatea de coincidență, %)	$4.74 \cdot 10^{-3}$

Discuții

Toxicitatea compușilor metal alchil a fost modelată pe baza informațiilor complexe oferite de structura acestora. Rezultatele studiului arată că există o relație între toxicitatea compușilor metal alchil și structura acestora. Trei descriptori moleculari s-au dovedit utili în estimarea și prezicerea toxicității compușilor metal alchil, doi dintre aceștia (iFDmCg, LAMrEQg) considerând geometria moleculei și unul (IHDmWMt) topologia acesteia. În ceea ce privește proprietatea atomică, descriptorul modelului mono-variat consideră cardinalitatea (iFDmCg), în timp ce descriptorii modelului bi-variat consideră masa atomică relativă (IHDmWMt) și sarcina atomică parțială (LAMrEQg). Modelul mono-variat consideră ca descriptor de interacțiune inversul distanței iar modelul bi-variat lucrul proprietății (IHDmWMt) și câmpul acesteia (LAMrEQg). Ambele modele sunt semnificative statistic, având o probabilitate de model greșit mai mică de $3.54 \cdot 10^{-5}$ %.

Valoarea coeficientului de corelație al modelului mono-variat ($r = 0.9830$, tabelul II) susține existența corelației dintre descriptorul iFDmCg și toxicitatea compușilor metal alchil. Aproape nouăzeci și șapte la sută din variația toxicității poate fi explicată prin relația lineară dintre aceasta și descriptorul iFDmCg. Modelul SAR mono-variat este un model stabil ($r^2 - r^2_{cv(100)} = 0.0198$, tabelul IV), capabil să prezică toxicitatea compușilor ($r^2_{cv(100)} = 0.9466$, tabelul IV) cu o probabilitate de a greși egală cu $2.29 \cdot 10^{-4}$ % (tabelul IV). Privind modelul mono-variat în ansamblu putem spune că, toxicitatea compușilor metal alchil este de natură geometrică, depinde de cardinalitatea compușilor și de inversul distanței metrice.

Modelul bi-variat prezintă un coeficient de corelație multiplă foarte aproape de valoarea 1 ($r = 0.9991$, tabelul IV), corelația dintre descriptorii modelului și toxicitatea compușilor fiind puternică. Valoarea coeficientului de determinare a modelului bi-variat ne arată că, aproape sută la sută din variația toxicității compușilor metal alchil poate fi atribuită relației lineare cu descriptorii moleculari IHDmWMt și LAMrEQg. Valoarea coeficientului de validare încrucișată ($r^2_{cv(100)} = 0.9965$, tabelul IV), a probabilității unui model de validare încrucișată greșit ($p_{pred} = 2.61 \cdot 10^{-7}$ %, tabelul IV) și a diferenței dintre coeficientul de determinare și coeficientul de validare încrucișată ($r^2 - r^2_{cv(100)} = 0.0018$, tabelul IV) susțin validitatea

modelului bi-variat și abilitatea acestuia în prezicerea toxicității compușilor metal alchil. Rezultatele analizei de regresia a modelului (valorile intervalului de încredere asociat coeficienților modelului, eroarea standard, valoarea parametrului Student și probabilitatea de a greși asociată testului Student – tabelul II) vin să susțină validitatea acestuia. Dacă ne uităm la valorile coeficienților de corelația dintre toxicitatea compușilor metal alchil cu fiecare descriptor molecular în parte, putem observa că, ambele corelații sunt semnificative ($r > 0.8$, tabelul II). Dar, valoare coeficientului de corelație al modelului bi-variat este mai mare în comparație cu valoarea coeficienților de corelație individuali, ceea ce vine să susțină superioritatea modelului bi-variat care ia în considerare descriptorii IHDmWMT și LAMrEQg în comparație cu posibile modele mono-variate care iau în considerare descriptorul IHDmWMT sau descriptorul LAMrEQg. Din punct de vedere al modelului bi-variat, putem spune că toxicitatea compușilor metal alchil este deopotrivă de natură geometrică și topologică, depinde de masa relativă, sarcina parțială, lucrul și câmpul proprietății.

Atât modelul mono-variat cât și cel bi-variat sunt capabile să estimeze și să prezică toxicitatea compușilor metal alchil. Pentru a vedea însă dacă există o diferență semnificativă între valorile coeficienților de corelație obținuți cu cele două modele, s-a aplicat testul Steiger. Valoarea coeficientului de corelație obținut de modelul bi-variat este semnificativ mai mare în comparație cu valoarea obținută de modelul mono-variat ($p_z = 4.74 \cdot 10^{-3} \%$, tabelul V). Astfel, dacă dorim o predicție a toxicității cât mai apropiată de valoarea reală, vom folosi modelul bi-variat.

În lucrarea [Ade & all, 1996] au fost construite modele distincte pentru fiecare metal. Prin modelul obținut (modelul bi-variat, tabelul II) se arată că modelul structură-activitate este independent de metal – și astfel se integrează informația structurală complexă într-o ecuație capabilă de predicții independente de ligand și de metal.

Modelele de predicție a toxicității compușilor metal alchil își găsesc utilitatea în prezicerea toxicității a compușilor noi, pe baza structurii acestora. Pentru obținerea toxicității unui nou compus, este necesar în primul rând să desenăm, folosind programul HyperChem, structura tridimensională a acestuia. Odată ce avem structura compusului metal alchil ca fișier *.hin, folosind facilitățile programului MDF SAR Predictor [MDF SAR Predictor, 2005], prin alegerea setului corespunzător compușilor metal alchil (setul 52730) și al modelului SAR (mono-variat sau bi-variat) putem prezice toxicitatea compusului de interes. Se deschide astfel calea spre obținerea de informații utile în ceea ce privește toxicitatea noilor compuși metal

alchil, informații obținute strict pe baza structurii compușilor, fără experimente directe, asistată de calculator, modalitatea mai ieftină și mai puțin consumatoare de timp.

Concluzii

Toxicitatea compușilor metal alchil poate fi modelată plecând de la structura acestora. Toxicitatea obținută cu cel mai performant model (modelul bi-variat) este deopotrivă de natură topologică și geometrică, depinde de masa relativă și sarcina parțială, fiind în relație directă cu lucrul proprietății și câmpul acesteia.

Aplicarea metodologiei SAR permite obținerea de modele exacte care deschid calea spre prezicerea toxicității a noi compuși plecând de la structura acestora.

Referințe

Chandra, S., James, B. D., Macauley, B. J. & Magee, R. J. (1987), 'Studies on the fungicidal properties of some organo-tin compounds', *J. Chem. Technol. Biotechnol.*, vol 39, no. 1, pp. 65-73.

Crowe, A. J. (2004), 'Organotin compounds in agriculture since 1980. Part I. Fungicidal, bactericidal and herbicidal properties', *Appl. Organomet. Chem.*, vol. 1, no. 2, pp. 143-55.

Mehrotra, R. C. & Singh, A. (2004), *Biological Applications and Environmental Aspects of Organometallic Compounds*, In: *Organometallic Chemistry. A Unified Approach.*, New Age International, New Delhi, p. 517-534.

Ade, T., Zaucke, F. & Krug, H. F. (1996), 'The structure of organometals determines cytotoxicity and alteration of calcium homeostasis in HL-60 cells', *Fresenius J. Anal. Chem.*, vol. 354, pp. 609-614.

Dacasto, M., Cornaglia, E., Nebbia, C. & Bollo, E. (2001), 'Triphenyltin acetate-induced cytotoxicity and CD4⁺ and CD8⁺ depletion in mouse thymocyte primary cultures', *Toxicology*, vol. 169, no. 3, pp. 227-238.

Aschner, M. & Aschner J.L. (1992), 'Cellular and molecular effects of trimethyltin and triethyltin: Relevance to organotin neurotoxicity', *Neurosci. Biobehav. Rev.*, vol. 16, pp. 427-435.

Walsh, T. J., McLamb, R. L. & Bondy, S. C. (1986), 'Triethyl and trimethyl lead: Effects on behavior, CNS morphology and concentrations of lead in blood and brain of rat', *NeuroToxicology*, vol. 7, no. 3, pp. 21-33.

ET108/2006 – Et. Unică/2006 – Lucrare in extenso

Eng, G., Brinckman, F. E., Olson, G. J., Tierney, E. J. & Bellama, J. M. (1991), 'Total surface areas of Group IVA organometallic compounds: Predictors of toxicity to algae and bacteria', *Appl. Organomet. Chem.*, vol. 5, no. 10), pp. 33-37.

Laughlin, R. B., French, W., Johannesen, R. B., Guard, H. E. & Brinckman, F. E. (1984), 'Predicting toxicity using computed molecular topologies: The example of triorganotin compounds', *Chemosphere*, vol. 13, no. 4, pp. 575-584.

Jäntschi, L. (2005), 'Molecular Descriptors Family on Structure Activity Relationships 1. Review of the Methodology', *Leonardo Electronic Journal of Practices and Technologies*, vol. 6, pp. 76-98.

HyperChem , Molecular Modelling System 2005; Hypercube, Inc., viewed 13 November, 2005, <<http://hyper.com/products>>.

Diudea, M., Gutman, I. & Jäntschi, L. (2001), *Molecular Topology*, Nova Science, Huntington, New York, pp. 332.

Jäntschi, L., Katona, G. & Diudea, M. (2000), 'Modeling Molecular Properties by Cluj Indices', *Commun Math Comput Chem (MATCH)*, Bayreuth, Germany, vol. 41, pp. 151-188.

Leave-one-out Analysis 2005, Virtual Library of Free Software, viewed 20 November, 2005, <http://vl.academicdirect.org/molecular_topology/mdf_findings/loo>.

Steiger, J. H. (1980), 'Tests for comparing elements of a correlation matrix', *Psychol Bull*, vol. 87, pp. 245-251.

MDF SAR Predictor 2005, Virtual Library of Free Software, viewed 20 November, 2005, <http://vl.academicdirect.org/molecular_topology/mdf_findings/sar>.

Pearson versus Spearman, Kendall's Tau Correlation Analysis on Structure-Activity Relationships of Biologic Active Compounds

Abstract

A sample of sixty-seven pyrimidine derivatives with inhibitory activity on *E. coli* dihydrofolate reductase (DHFR) was studied by the use of molecular descriptors family on structure-activity relationships. Starting from the results obtained by applying of MDF-SAR methodology on pyrimidine derivatives and from the assumption that the measured activity (compounds' inhibitory activity) of a biologically active compounds is a semi-quantitative outcome (can be related with the type of equipment used, the researchers, the chemical used, etc.), the abilities of Pearson, Spearman, Kendall's, and Gamma correlation coefficients in analysis of estimated toxicity were studied and are presented.

Keywords

Multiple linear regressions, Correlation coefficients, Molecular Descriptors Family on Structure-Activity Relationships (MDF-SAR)

Introduction

QSAR (Quantitative Structure-Activity Relationships) is an approach which is able to indicate for a given compound or a class of compounds which feature of structure characteristics is correlated with its activity [1]. In QSAR analysis were proposed several approaches for development. Simple and multiple linear regressions is one of the more successful techniques use by many researcher in construct of QSAR models [2-4].

[1] Rogers D., Hopfinger A. J., *Application of Genetic Function Approximation to Quantitative Structure-Activity Relationships and Quantitative Structure-Property Relationships*, J. Chem. Inf. Comput. Sci. 34, 1994, p. 854-866.

[2] Hansch C., Leo A., Stephen R., Eds. Heller, *Exploring QSAR, Fundamentals and Applications in Chemistry and Biology*, ACS professional Reference Book., American

Correlation coefficient is a simple statistical measure of relationship between one dependent and one or more than one independent variables and it is use as a measure of the statistical fit of a regression based model in QSAR [5]. Its squared value (the coefficient of determination) it is most frequently used parameter as a measure of the goodness-of-fit of the model [6-10].

Chemical Society, Washington, D.C., 1995.

[3] Zahouily M., Lazar M., Elmakssoudi A., Rakik J., Elaychi S., Rayadh A., *QSAR for anti-malarial activity of 2-aziridinyl and 2,3-bis(aziridinyl)-1,4-naphthoquinonyl sulfonate and acylate derivatives*, J Mol Model 12(4), 2006, p. 398-405.

[4] Liang G.-Z., Mei H., Zhou P., Zhou Y., Li Z.-L., *Study on quantitative structure-activity relationship by 3D holographic vector of atomic interaction field*, Acta Phys-Chim Sin 22(3), 2006, p. 388-390.

[5] Rosner B., *Fundamentals of Biostatistics*, 4th Edition, Duxbury Press, Belmont, California, USA, 1995.

[6] Katritzky A. R., Kuanar M., Slavov S., Dobchev D.A., Fara D. C., Karelson M., Acree Jr. W. E., Solov'ev V. P., Varnek A., *Correlation of blood-brain penetration using structural descriptors*, Bioorg Med Chem, 14(14), 2006, p. 4888-4917.

[7] Wang Y., Zhao C., Ma W., Liu H., Wang T., Jiang G., *Quantitative structure-activity relationship for prediction of the toxicity of polybrominated diphenyl ether (PBDE) congeners*, Chemosphere 64(4), 2006, p. 515-524.

[8] Roy D. R., Parthasarathi R., Subramanian V., Chattaraj P. K., *An electrophilicity based analysis of toxicity of aromatic compounds towards Tetrahymena pyriformis*, QSAR Comb Sci 25(2), 2006, p 114-122.

[9] Srivastava H. K., Pasha F. A., Singh P. P., *Atomic softness-based QSAR study of testosterone*, Int J Quantum Chem 103(3), 2005, p. 237-245.

[10] Xue C. X., Zhang R. S., Liu H. X., Yao X. J., Hu M. C., Hu Z. D., Fan B. T., *QSAR models for the prediction of binding affinities to human serum albumin using the heuristic method and a support vector machine*, J Chem Inf Comput Sci 44(5), 2004, p. 1693-1700.

A new approach of molecular descriptors family on structure-activity relationships (MDF-SAR) was developed [11], and proved its usefulness in estimation and prediction of: toxicity [12, 13], mutagenicity [12], antioxidant efficacy [14], antituberculosic activity [15], antimalarial activity [16], antiallergic activity [17], anti-HIV-1 potencies [18], inhibition activity on carbonic anhydrase II [19] and IV [20].

[11] Jäntschi L., *Molecular Descriptors Family on Structure Activity Relationships 1. Review of the Methodology*, Leonardo Electronic Journal of Practices and Technologies 6, 2005, p. 76-98.

[12] Jäntschi L., Bolboacă S., *Molecular Descriptors Family on QSAR Modeling of Quinoline-based Compounds Biological Activities*, The 10th Electronic Computational Chemistry Conference 2005; http://bluehawk.monmouth.edu/~rtopper/eccc10_absbook.pdf as on 13 May 2006.

[13] Bobloacă S.D., Jäntschi L., *Modeling of Structure-Toxicity Relationship of Alkyl Metal Compounds by Integration of Complex Structural Information*, Therapeutics, Pharmacology and Clinical Toxicology X(1), 2006, p. 110-114.

[14] Bolboacă S., Filip C., Țigan Ș., Jäntschi L., *Antioxidant Efficacy of 3-Indolyl Derivates by Complex Information Integration*, Clujul Medical LXXIX(2), 2006, p. 204-209.

[15] Bolboacă S., Jäntschi L., *Molecular Descriptors Family on Structure Activity Relationships 3. Antituberculosic Activity of some Polyhydroxyxanthenes*, Leonardo Journal of Sciences 7, 2005, p. 58-64.

[16] Jäntschi L., Bolboacă S., *Molecular Descriptors Family on Structure Activity Relationships 5. Antimalarial Activity of 2,4-Diamino-6-Quinazoline Sulfonamide Derivates*, Leonardo Journal of Sciences 8, 2006, p. 77-88.

[17] Jäntschi L., Bolboacă S., *Antiallergic Activity of Substituted Benzamides: Characterization, Estimation and Prediction*, Clujul Medical LXXIX, 2006, In press.

[18] Bolboacă S., Țigan Ș., Jäntschi L., *Molecular Descriptors Family on Structure-Activity Relationships on anti-HIV-1 Potencies of HEPTA and TIBO Derivatives*, In: Reichert A., Mihalaș G., Stoicu-Tivadar L., Schulz Ș., Engelbrech R. (Eds.), Proceedings of the European Federation for Medical Informatics Special Topic Conference, p. 222-226, 2006.

[19] Jäntschi L., Ungureșan M. L., Bolboacă S.D., *Integration of Complex Structural Information in Modeling of Inhibition Activity on Carbonic Anhydrase II of Substituted Disulfonamides*, Applied Medical Informatics 17(3,4), 2005, p. 12-21.

Several correlation coefficients based on different statistical hypothesis are known and most frequently used today: Pearson correlation coefficient, Spearman rank correlation coefficient and Spearman semi-quantitative correlation coefficient, Kendall tau-a, -b and -c correlation coefficients, Gamma correlation coefficient [5].

Starting from the results obtained by applying of MDF-SAR methodology on a sample of sixty-seven compounds and from the assumption that the measured activity (compounds' inhibitory activity) of a biologically active compounds is a semi-quantitative outcome (can be related with the type of equipment used, the researchers, the chemical used), the abilities of Pearson, Spearman, Kendall's, and Gamma correlation coefficients in analysis of estimated toxicity were studied.

Multi-varied MDF-SAR model of pyrimidine derivatives

A sample of sixty-seven pyrimidine derivatives with inhibitory activity on E. coli dihydrofolate reductase (DHFR) was studied by the use of MDF-SAR methodology.

The set of pyrimidine derivatives (2,4-Diamino-5-(substituted-benzyl)-pyrimidine derivatives) with inhibitory activity on E. coli dihydrofolate reductase (DHFR) was previously studied by Ting-Lan Chiu & Sung-Sau So by the use of neural network approach [21].

By applying the MDF-SAR methodology on the sample of sixty-seven pyrimidine derivatives, a multi-varied model with four descriptors revealed to have good performances in prediction and estimation of inhibitory activity.

The multi-varied MDF-SAR model with four descriptors had the following equation:

$$Y_{\text{est}} = 3.78 + 1.62 \cdot iImrKHt + 2.37 \cdot liMDWHg + 6.40 \cdot IsDrJQt - 8.52 \cdot 10^{-2} \cdot LSPmEQg$$

Analyzing the MDF-SAR model with four descriptors it could be said that inhibitory activity considers compounds geometry (**g**) and topology (**t**), being related with the number of

[20] Jäntschi L., Bolboacă S., *Modelling the Inhibitory Activity on Carbonic Anhydrase IV of Substituted Thiadiazole- and Thiadiazoline- Disulfonamides: Integration of Structure Information*, Electronic Journal of Biomedicine, 2006, In press.

[21] Chiu T.L., So S. S., *Development of neural network QSPR models for Hansch substituent constants. 2. Applications in QSAR studies of HIV-1 reverse transcriptase and dihydrofolate reductase inhibitors*, J Chem Inf Comput Sci 44(1), 2004, p. 154-160.

directly bonded hydrogen's (**H**) of compounds and with the partial charge (**Q**) as atomic properties.

Statistical characteristics of the MDF-SAR model with four descriptors are in table 1 and 2.

Table 1. Statistical characteristics of the multi-varied MDF-SAR model with four descriptors

Characteristic (notation)	Value
Number of variable (v)	4
Correlation coefficient (r)	0.9517
95% Confidence Intervals for r (95% CI _r)	[0.9223, 0.9701]
Squared correlation coefficient (r ²)	0.9058
Adjusted squared correlation coefficient (r ² _{adj})	0.8997
Standard error of estimated (S _{est})	0.1919
Fisher parameter (F _{est})	149*
Cross-validation leave-one-out (loo) score (r ² _{cv-loo})	0.8932
Fisher parameter for loo analysis (F _{pred})	130*
Standard error for leave-one-out analysis (S _{loo})	0.2044
Model stability (r ² - r ² _{cv(loo)})	0.0126
r ² (iImrKHt, liMDWHg)	0.2020
r ² (iImrKHt, IsDrJQt)	0.0047
r ² (iImrKHt, LSPmEQg)	0.1482
r ² (liMDWHg, IsDrJQt)	0.0003
r ² (liMDWHg, LSPmEQg)	0.0212
r ² (IsDrJQt, LSPmEQg)	0.0664

*p < 0.001

Table 2. Statistics of the regression MDF-SAR model with four descriptors

	StdError	t Stat	95%CI _{coefficient}	r(Y _m , desc)
Intercept	0.1999	18.92*	[3.38, 4.18]	n.a.
iImrKHt	0.0709	22.85*	[1.48, 1.76]	0.4803
liMDWHg	0.1500	15.81*	[2.07, 2.67]	0.0558
IsDrJQt	1.4779	4.33*	[3.45, 9.36]	0.0336
LSPmEQg	0.0182	-4.68*	[-0.12, -0.12]	0.0231

StdError = standard error; t Stat = Student tets parameter;
 95% CI_{coefficient} = 95% confidence interval associated with regression coefficients;
 Y_m = measured inhibitory activity; desc = molecular descriptor; * p < 0.001

Graphical representation of the measured versus estimated by MDF-SAR model with four descriptors inhibitory activity is in figure 1.

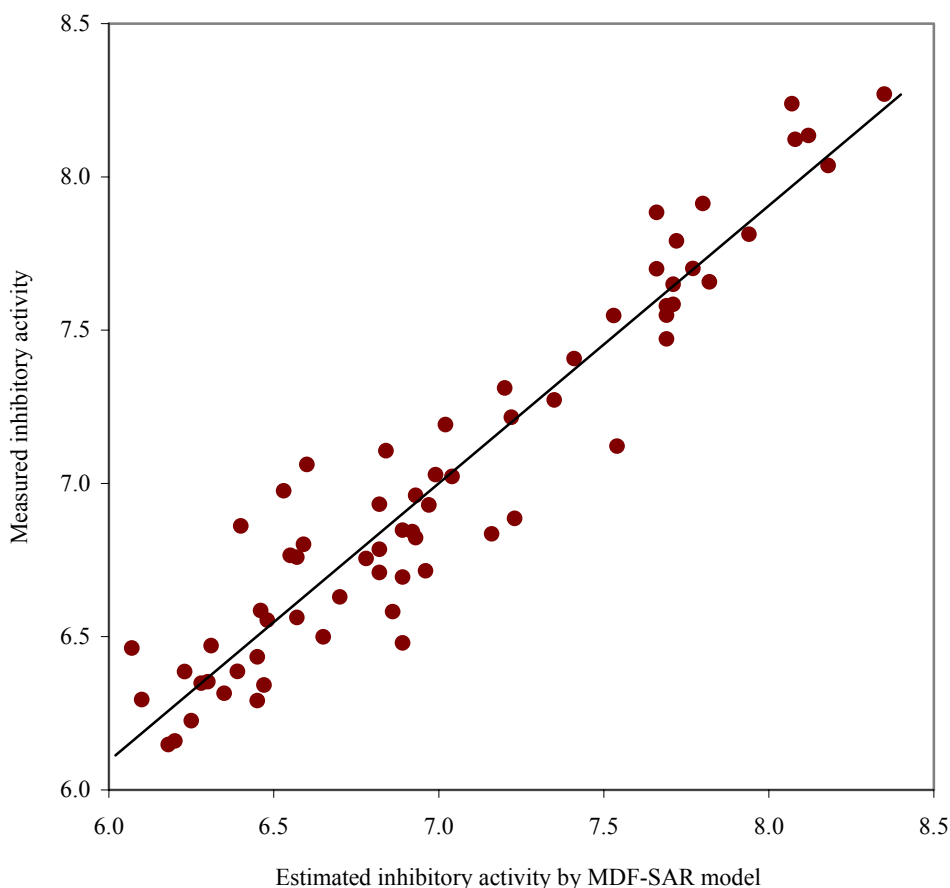


Figure 1. Plot of measured vs estimated by MDF-SAR inhibitory activity

Internal validation of the four-varied MDF SAR model with four descriptors was performed through splitting the whole set into training and test sets by applying of a randomization algorithm.

The coefficients for each model obtained in training sets, in conformity with the generic equation $Y_{est} = a_0 + a_1 \cdot ilmrKHt + a_2 \cdot liMDWHg + a_3 \cdot IsDrJQt - a_4 \cdot 10^{-2} \cdot LSPmEQg$, the number of compounds in training (N_{tr}) and test (N_{ts}) sets, the correlation coefficient for training (r_{tr}) and test (r_{ts}) sets with associated 95% confidence intervals (95%CI_{tr} and 95%CI_{ts}), the Fisher parameter associated with training (F_{tr}) and test (F_{ts}) sets, and the Fisher's Z parameter of correlation coefficients comparison ($Z_{trtr-rts}$) are in table 3.

Table 3. Statistics results on training versus test sets

a_0	a_1	a_2	a_3	a_4	N_{tr}	r_{tr}	95%CI _{tr}	F_{tr}	N_{ts}	r_{ts}	95%CI _{ts}	F_{ts}	$Z_{trtr-rts}$
3.93	1.61	2.43	6.56	$-9.67 \cdot 10^{-2}$	35	0.949	[0.899, 0.974]	67*	32	0.958	[0.916, 0.980]	59*	0.418 [†]
3.98	1.57	2.45	6.55	$-7.52 \cdot 10^{-2}$	36	0.951	[0.905, 0.975]	73*	31	0.951	[0.899, 0.976]	61*	0.000 [†]
3.84	1.55	2.15	9.08	$-9.12 \cdot 10^{-2}$	37	0.944	[0.893, 0.908]	66*	30	0.949	[0.895, 0.976]	55*	0.206 [†]
3.94	1.59	2.42	6.10	$-8.18 \cdot 10^{-2}$	38	0.951	[0.907, 0.974]	78*	29	0.947	[0.890, 0.975]	50*	0.144 [†]
3.91	1.56	2.25	8.22	$-1.04 \cdot 10^{-1}$	39	0.963	[0.931, 0.981]	110*	28	0.937	[0.867, 0.971]	39*	1.069 [†]

4.18	1.51	2.44	6.06	$-7.22 \cdot 10^{-2}$	40	0.956	[0.917, 0.975]	92*	27	0.936	[0.863, 0.971]	35*	0.721 [†]
3.76	1.63	2.32	7.35	$-1.02 \cdot 10^{-1}$	41	0.963	[0.931, 0.980]	116*	26	0.935	[0.858, 0.971]	34*	1.104 [†]
3.97	1.58	2.39	5.11	$-9.36 \cdot 10^{-2}$	42	0.956	[0.919, 0.976]	99*	25	0.954	[0.896, 0.980]	34*	0.115 [†]
3.64	1.64	2.30	7.00	$-8.15 \cdot 10^{-2}$	43	0.955	[0.917, 0.975]	98*	24	0.944	[0.873, 0.976]	37*	0.407 [†]
3.72	1.66	2.43	5.78	$-8.12 \cdot 10^{-2}$	44	0.938	[0.889, 0.966]	72*	23	0.964	[0.916, 0.985]	54*	1.030 [†]
3.59	1.64	2.25	4.94	$-9.98 \cdot 10^{-2}$	45	0.947	[0.904, 0.970]	86*	22	0.957	[0.898, 0.982]	37*	0.411 [†]
3.86	1.55	2.23	8.68	$-8.86 \cdot 10^{-2}$	46	0.940	[0.894, 0.967]	78*	21	0.983	[0.958, 0.993]	43*	2.290*
4.04	1.54	2.36	6.46	$-7.31 \cdot 10^{-2}$	47	0.949	[0.911, 0.972]	96*	20	0.963	[0.906, 0.985]	34*	0.538 [†]
3.63	1.63	2.24	4.27	$-8.93 \cdot 10^{-2}$	48	0.940	[0.895, 0.966]	82*	19	0.963	[0.904, 0.986]	44*	0.852 [†]
3.98	1.57	2.42	6.49	$-8.59 \cdot 10^{-2}$	49	0.946	[0.905, 0.969]	93*	18	0.960	[0.894, 0.985]	36*	0.535 [†]
3.77	1.61	2.32	6.37	$-8.46 \cdot 10^{-2}$	50	0.943	[0.902, 0.968]	91*	17	0.974	[0.927, 0.991]	52*	1.294 [†]
3.67	1.63	2.22	6.56	$-1.01 \cdot 10^{-1}$	51	0.954	[0.919, 0.973]	115*	16	0.950	[0.858, 0.983]	17*	0.126 [†]
3.81	1.61	2.39	6.87	$-7.70 \cdot 10^{-2}$	52	0.951	[0.916, 0.972]	112*	15	0.950	[0.853, 0.984]	22*	0.032 [†]
3.69	1.65	2.36	6.32	$-8.21 \cdot 10^{-2}$	53	0.953	[0.919, 0.972]	118*	14	0.956	[0.864, 0.986]	17*	0.128 [†]
3.97	1.56	2.40	6.16	$-7.51 \cdot 10^{-2}$	54	0.951	[0.916, 0.971]	115*	13	0.954	[0.851, 0.987]	17*	0.122 [†]

[†] p > 0.05; * p < 0.01

Definitions, Formulas, Interpretations, PHP functions, and Results

A number of add notations were used in the study, as follows:

- Pearson product-moment correlation coefficient (named after Karl Pearson (1857 - 1936), a major contributor to the early development of statistics):
 - r_{prs} = the *Pearson* correlation coefficient;
 - r_{prs}^2 = the squared *Pearson* correlation coefficient;
 - $t_{prs,df}$ = the Student test parameter, and its significance $p_{prs,df}$ at a significance level of 5% (where df = the degree of freedom);
- Spearman's rank correlation coefficient (named after Charles Spearman (1863 - 1945), English psychologist known for his work in statistics - *factor analysis*, and *Spearman's rank correlation coefficient*):
 - r_{spm} = the *Spearman* rank correlation coefficient
 - r_{spm}^2 = the squared of *Spearman* rank correlation coefficient;
 - $t_{prs,df}$ = the Student test parameter, and its significance $p_{spm,df}$;
 - r_{sQ}^2 = the squared of *Spearman semi-Quantitative* correlation coefficient;
 - t_{sQ} = the Student test parameter, and its significance p_{sQ} ;
- Kendall's tau correlation coefficients (named after Maurice George Kendall (1907 - 1983), a prominent British statistician; published in monograph *Rank Correlation* in 1948):
 - $\tau_{Ken,a}$ = the *Kendall tau-a* correlation coefficient;
 - $\tau_{Ken,a}^2$ = the squared of *Kendall tau-a* correlation coefficient;

ET108/2006 – Et. Unică/2006 – Lucrare in extenso

- Z_{Ken,τ_a} = the Z-test parameter of Kendall tau-a correlation coefficient, and its significance p_{Ken,τ_a} ;
- $\tau_{\text{Ken},b}$ = the *Kendall tau-b* correlation coefficient;
- $\tau_{\text{Ken},b}^2$ = the squared of *Kendall tau-b* correlation coefficient;
- Z_{Ken,τ_b} = the Z-test parameter of Kendall tau-b, and its significance p_{Ken,τ_b} ;
- $\tau_{\text{Ken},c}$ = the *Kendall tau-c* correlation coefficient;
- $\tau_{\text{Ken},c}^2$ = the squared of *Kendall tau-c* correlation coefficient;
- Z_{Ken,τ_c} = the Z-test parameter of *Kendall tau-c*, and its significance p_{Ken,τ_c} ;
- Gamma correlation coefficient (also known as Goodman and Kruskal's gamma):
 - Γ = the *Gamma* correlation coefficient;
 - Γ^2 = the squared of *Gamma* correlation coefficient;
 - Z_{Γ} = the Z-test parameter of *Gamma* correlation coefficient, and its significance p_{Γ} .

A series of *.php programs which to facilitate the calculation and to display of above-described correlation coefficients and their statistics (Student-test and Z-test parameters and associated significances) were implemented and was use in order to reach the objective of study [22].

Pearson correlation coefficient

Definition: a measure the strength and direction of the linear relationship between two variables, describing the direction and degree to which one variable is linearly related to another.

Assumptions: both variable (variables Y_m and Y_{est}) are interval or ratio variables and are well approximated by a normal distribution, and their joint distribution is bivariate normal [23].

[22] ***Rank, ©2005, Virtual Library of Free Software, available at: http://vl.academicdirect.org/molecular_topology/mdf_findings/rank/

[23] ***Pearson's Correlation Coefficient [online], Available at: <http://www.texasoft.com/winkpear.html>

Formula

$$r_{Prs} = \frac{\sum (Y_{m-i} - \bar{Y}_m)(Y_{est-i} - \bar{Y}_{est})}{\sqrt{(\sum (Y_{m-i} - \bar{Y}_m)^2)(\sum (Y_{est-i} - \bar{Y}_{est})^2)}}$$

where Y_{m-i} is the value of the measured inhibitory activity for compound i ($i = 1, 2, \dots, 67$)
 \bar{Y}_m is the average of the measured inhibitory activity, Y_{est-i} is the value of the estimated inhibitory activity for compound i , and \bar{Y}_{est} is the average of the estimated inhibitory activity.

Interpretation

The Pearson correlation coefficient can take values from -1 to +1. A value of +1 show that the variables are perfectly linear related by an increasing relationship, a value of -1 show that the variables are perfectly linear related by an decreasing relationship, and a value of 0 show that the variables are not linear related by each other. There is considered a strong correlation if the correlation coefficient is greater than 0.8 and a weak correlation if the correlation coefficient is less than 0.5.

The coefficient of determination (or r squared) gives information about the proportion of variation in the dependent variable which might be considered as being associated with the variation in the independent variable.

Related statistics

- The squared of Pearson correlation coefficient or Pearson coefficient of determination (r_{Prs}^2);
 - Describe the proportion of variance in Y_m that is related with linear variation of Y_{est} ;
 - Can take values from 0 to 1.

Statistical test

Student t-test was used to determine if the value of Pearson correlation coefficient is statistically significant, at a significance level of 5%.

The null hypothesis vs. the alternative hypothesis was:

$H_0: r_{Prs} = 0$ (there is no correlation between the variables)

$H_1: r_{Prs} <> 0$ (variables are correlated)

For a significance level equal with 5%, a p-value associated to $t_{Prs,df}$ less than 0.05 means that there is evidence to reject the null hypothesis in favor of the alternative hypothesis. In other words there is a statistically significant linear relationship between the variables.

PHP implementation

In order to compute the statistics associated with Pearson correlation coefficient, three functions were implemented:

```
function coef_rk(&$y1,&$y2){
    $my1=m1($y1);
    $dy2=m2($y1,$y1)-$my1*$my1;
    $mx1=m1($y2);
    $mxy=m2($y2,$y1);
    $m2x=$mx1*$mx1;
    $mx2=m2($y2,$y2);
    $dx2=$mx2-$m2x;
    $r2=pow($mxy-$mx1*$my1,2)/($dx2*$dy2);
    return $r2;
}
function t_p($n,$k,$r){
    return $r*pow($n-$k-1,0.5)/pow(1-pow($r,2),0.5);
}
function p_t($t,$df){
    $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);
}
```

The statistics of Pearson correlation coefficients are computed as follows:

- Pearson correlation coefficient:

$$r_{pe} = coef_rk(\$cmp[0],\$cmp[1]);$$

where $\$cmp[0]$ is the measured inhibitory activity (Y_m), and $\$cmp[1]$ is the estimated by MDF-SAR model with four descriptor inhibitory activity (Y_{est}).

- t Student parameter:

$$t_{pe} = t_p(\$n,1,pow(\$r_{pe},0.5));$$

- Significance of t Student parameter

$$p_{pe} = p_t(t_{pe},\$n-2);$$

Results

$$r_{Prs}^2 = 0.9058$$

$$\begin{aligned}t_{Prs,1} &= 24.99 \\ p_{Prs,1} &= 4.74 \cdot 10^{-33} \%\end{aligned}\tag{1}$$

Spearman's rank correlation coefficient

Definition

A non-parametric measure of correlation between variable which assess how well an arbitrary monotonic function could describe the relationship between two variables, without making any assumptions about the frequency distribution of the variables. Frequently the Greek letter ρ (rho) is use to abbreviate the Spearman correlation coefficient.

Spearman's rank correlation is satisfactory for testing the null hypothesis of no relationship, but is difficult to interpret as a measure of the strength of the relationship [24].

Assumptions

- Does not required any assumptions about the frequency distribution of the variables;
- Does not required the assumption that the relationship between variable is linear;
- Does not required the variable to be measured on interval or ration scale.

Formula

In order to compute the Spearman rank correlation coefficient, the two variables (Y_m , respectively Y_{est}) were converted to ranks (see table 4 for exemplification). For each measured and estimated inhibitory activity a rank was assigned ($RankY_m$ - for measure inhibitory activity, $RankY_{est}$ - for estimated by MDF-SAR model inhibitory activity) according with the position of value into a sort serried of values.

In assignment of rank process, the lowest value had the lowest rank. When there are two equal values for two different compounds (for measured and/or estimated inhibitory activity), the associated rank had equal values and was calculated as means of corresponding ranks. For example, the compounds abbreviated as c_52 and c_59 have the same measured inhibitory activity (6.45, see table 4). The rank associated with these values is equal with 13.5 (is the average between the rank for c_52 - 13 and the rank of c_59 - 14).

[24] Methods based on rank order. In: Bland M., *An Introduction to Medical Statistics*, Oxford University Press; Oxford, New York, Tokyo, p. 205-225, 1995.

Table 4. Compounds abbreviation, measured and estimated activity and associated ranks

Abb.	Y _m	RankY _m		Y _{est}	RankY _{est}	Abb.	Y _m	RankY _m		Y _{est}	RankY _{est}
c_64	6.07	1	0	6.4626	13	c_32	6.92	35	0	6.8423	32
c_65	6.10	2	0	6.2948	5	c_66	6.93	36.5	5	6.8225	30
c_67	6.18	3	0	6.1479	1	c_36	6.93	36.5		6.9609	38
c_54	6.20	4	0	6.1595	2	c_40	6.96	38	0	6.7150	24
c_37	6.23	5	0	6.3859	10	c_17	6.97	39	0	6.9298	36
c_48	6.25	6	0	6.2254	3	c_45	6.99	40	0	7.0283	41
c_31	6.28	7	0	6.3483	8	c_41	7.02	41	0	7.1919	45
c_49	6.30	8	0	6.3528	9	c_15	7.04	42	0	7.0225	40
c_10	6.31	9	0	6.4703	14	c_28	7.16	43	0	6.8355	31
c_56	6.35	10	0	6.3149	6	c_09	7.20	44	0	7.3115	48
c_47	6.39	11	0	6.3866	11	c_18	7.22	45	0	7.2156	46
c_53	6.40	12	0	6.8614	34	c_43	7.23	46	0	6.8855	35
c_52	6.45	13.5	1	6.2913	4	c_29	7.35	47	0	7.2724	47
c_59	6.45	13.5		6.4336	12	c_14	7.41	48	0	7.4072	49
c_16	6.46	15	0	6.5851	20	c_24	7.53	49	0	7.5476	51
c_34	6.47	16	0	6.3422	7	c_22	7.54	50	0	7.1218	44
c_58	6.48	17	0	6.5536	17	c_26	7.66	51.5	6	7.7002	57
c_35	6.53	18	0	6.9755	39	c_08	7.66	51.5		7.8841	61
c_42	6.55	19	0	6.7654	27	c_27	7.69	54	7	7.4715	50
c_30	6.57	20.5	2	6.5625	18	c_13	7.69	54		7.5489	52
c_61	6.57	20.5		6.7594	26	c_12	7.69	54		7.5793	53
c_33	6.59	22	0	6.8010	29	c_04	7.71	56.5	8	7.5841	54
c_51	6.60	23	0	7.0616	42	c_11	7.71	56.5		7.6497	55
c_39	6.65	24	0	6.4993	16	c_19	7.72	58	0	7.7915	59
c_38	6.70	25	0	6.6297	21	c_23	7.77	59	0	7.7014	58
c_57	6.78	26	0	6.7552	25	c_25	7.80	60	0	7.9130	62
c_60	6.82	28	3	6.7091	23	c_01	7.82	61	0	7.6576	56
c_44	6.82	28		6.7847	28	c_21	7.94	62	0	7.8130	60
c_55	6.82	28		6.9318	37	c_06	8.07	63	0	8.2391	66
c_20	6.84	30	0	7.1067	43	c_03	8.08	64	0	8.1224	64
c_46	6.86	31	0	6.5813	19	c_07	8.12	65	0	8.1353	65
c_50	6.89	33	4	6.4794	15	c_05	8.18	66	0	8.0372	63
c_62	6.89	33		6.6942	22	c_02	8.35	67	0	8.2702	67
c_63	6.89	33		6.8475	33						

The method of rank assignment for more than two equal values of measured and/or estimated inhibitory activity is the same as for two equal values. If there are an odd number of compounds which have the same measured value (see compounds c_60, c_44, and c_55 from table 2) then the rank will be an integer $((27+28+29)/3 = 28$, see the rank for c_60, c_44, and c_55).

In studied example, there are equal values for measured activity: five situations of two equal values (c_52-c_59, c_30-c_61, c_66-c_36, c_26-c_08, and c_04-c_11), and three situations of three equal values (c_60-c_44-c_55, c_50-c_62-c_63, and c_27-c_13-c_12).

By conversion of the measured and estimated inhibitory activity to ranks, the distribution of ranks does not depend on the distribution of measured, respectively estimated inhibitory activity.

The formula for calculation of the Spearman rank correlation coefficient is:

$$r_{\text{Spm}} = \frac{\sum (R_{Y_{m-i}} - \bar{R}_{Y_m})(R_{Y_{\text{est}-i}} - \bar{R}_{Y_{\text{est}}})}{\sqrt{(\sum (R_{Y_{m-i}} - \bar{R}_{Y_m})^2)(\sum (R_{Y_{\text{est}-i}} - \bar{R}_{Y_{\text{est}}})^2)}}$$

where $R_{Y_{m-i}}$ is the rank of the measured inhibitory activity for compound i , $\bar{R}_{Y_{m-i}}$ is the average of the measured inhibitory activity, $R_{Y_{\text{est}-i}}$ is the rank of the estimated by MDF-SAR inhibitory activity for compound i , and $\bar{R}_{Y_{\text{est}-i}}$ is the average of the estimated inhibitory activity.

The simple formula for r_{Spm} is based on the difference between each pairs of ranks:

$$r_{\text{Spm}} = 1 - \frac{6 \sum D^2}{n(n^2 - 1)}$$

where D is the differences between each pair of ranks (e.g. $D = R_{Y_{m-1}} - R_{Y_{\text{est}-1}}$) and n is the volume of the sample.

The formula of the *Spearman semi-quantitative method* is:

$$r_{\text{sQ}} = \sqrt{\frac{\sum (Y_{m-i} - \bar{Y}_m)(Y_{\text{est}-i} - \bar{Y}_{\text{est}})}{\sqrt{(\sum (Y_{m-i} - \bar{Y}_m)^2)(\sum (Y_{\text{est}-i} - \bar{Y}_{\text{est}})^2)}}} \cdot \frac{\sum (R_{Y_{m-i}} - \bar{R}_{Y_m})(R_{Y_{\text{est}-i}} - \bar{R}_{Y_{\text{est}}})}{\sqrt{(\sum (R_{Y_{m-i}} - \bar{R}_{Y_m})^2)(\sum (R_{Y_{\text{est}-i}} - \bar{R}_{Y_{\text{est}}})^2)}}$$

Interpretation

- Identical with Pearson correlation coefficient.

Related statistics

- r_{Spm}^2 = the squared of *Spearman* rank correlation coefficient;
- r_{sQ}^2 = the squared of semi-quantitative correlation coefficient.

Statistical significance

- Compute by the use of a permutation test (a statistical test in which the reference distribution is obtained by permuting the observed data points across all possible outcomes, given a set of conditions consistent with the null hypothesis);
- Comparing the observed r_{Spm} with published tables for different levels of significance (eg. 0.05, 0.01...). It is a simple solution when the researchers want to know the significance within a certain range or less than a certain value;
- Tested by applying the Student t-test (for sample sizes > 20): the method used in this study.

The null hypothesis vs. the alternative hypothesis for Spearman rank correlation coefficient was:

$$H_0: r_{\text{Spm}} = 0 \text{ (there is no correlation between the ranked pairs)}$$

$$H_1: r_{\text{Spm}} < > 0 \text{ (ranked pairs are correlated)}$$

The null hypothesis vs. the alternative hypothesis for semi-quantitative correlation coefficient was:

$$H_0: r_{\text{sQ}} = 0 \text{ (there is no correlation between the ranked pairs)}$$

$$H_1: r_{\text{sQ}} < > 0 \text{ (ranked pairs are correlated)}$$

PHP implementation

The formulas for Spearman and respectively semi-quantitative correlation coefficients used two defined above functions (t_p and respectively p_t). The Spearman rank correlation coefficient used the *coef_rk* function defined as:

```
function coef_rk(&$y1,&$y2){
    $my1=m1($y1);
    $dy2=m2($y1,$y1)-$my1*$my1;
    $mx1=m1($y2);
    $mxy=m2($y2,$y1);
    $m2x=$mx1*$mx1;
    $mx2=m2($y2,$y2);
    $dx2=$mx2-$m2x;
    $r2=pow($mxy-$mx1*$my1,2)/($dx2*$dy2);
    return $r2;
}
```

where

```
function m1(&$v){
    $rez=0;
    $n=count($v);
    for($i=1;$i<$n;$i++)
        $rez+=$v[$i];
}
```

```
    return $rez/($n-1);  
  }  
  function m2(&$v,&$u){  
    $rez=0;  
    $n=count($v);  
    for($i=1;$i<$n;$i++)  
      $rez+=$v[$i]*$u[$i];  
    return $rez/($n-1);  
  }  
}
```

Spearman correlation coefficient

The statistics of Spearman rank correlation coefficients are computed as follows:

- Spearman correlation coefficient:

$$r_{sp} = coef_rk(\$poz[0],\$poz[1]);$$

where $\$poz[0]$ is the position on sort series of measured inhibitory activity, and $\$poz[1]$ is the position on sort series of estimated inhibitory activity by MDF-SAR model with four descriptor.

- t Student parameter:

$$t_{sp} = t_p(\$n,1,pow(\$r_{sp},0.5));$$

- Significance of t Student parameter

$$p_{sp} = p_t(t_{sp},\$n-2);$$

Semi-quantitative correlation coefficient

The statistics of semi-quantitative correlation coefficients are computed as follows:

- Semi-quantitative correlation coefficient:

$$r_{sq} = pow(\$r_{pe}*\$r_{sp},0.5);$$

- t Student parameter:

$$t_{sq} = t_p(\$n,1,pow(\$r_{sq},0.5));$$

- Significance of t Student parameter

$$p_{sq} = p_t(t_{sq},\$n-2);$$

Results

$$r_{Spr}^2 = 0.8606$$

$$t_{Spm,1} = 20.03$$

(2)

$$\begin{aligned} p_{\text{Spm},1} &= 1.62 \cdot 10^{-29} \\ r_{\text{sQ}}^2 &= 0.8829 \\ t_{\text{sQ}} &= 22.14 \\ p_{\text{sQ}} &= 5.57 \cdot 10^{-32} \end{aligned} \tag{3}$$

Kendall's rank correlation coefficients

Definition

Kendall-tau is a non-parametric correlation coefficient that can be used to assess and test correlations between non-interval scaled ordinal variables. Frequently the Greek letter τ (tau), is use to abbreviate the Kendall tau correlation coefficient.

The Kendall tau correlation coefficient is considered to be equivalent to the Spearman rank correlation coefficient. While Spearman rank correlation coefficient is like the Pearson correlation coefficient but computed from ranks, the Kendall tau correlation rather represents a probability.

There are three Kendall's tau correlation coefficient known as tau-a, tau-b, and tau-c.

Formula

Let $(Y_{m-i}, Y_{\text{est}-i})$ and $(Y_{m-j}, Y_{\text{est}-j})$ be the pair of measured and estimated inhibitory activity. If $Y_{m-j} - Y_{m-i}$ and $Y_{\text{est}-j} - Y_{\text{est}-i}$, where $i < j$ have the same sign the pair is *concordant*, if have opposite signs the pair is *discordant*.

In a sample of n observations it can be found $n(n-1)/2$ pairs corresponding to choices $1 \leq i < j \leq n$.

The formulas of Kendall's tau correlation coefficients are as follows:

- Kendall tau-a correlation coefficient ($\tau_{\text{Ken},a}$):

$$\tau_{\text{Ken},a} = (C-D)/[n(n-1)/2]$$

- Kendall tau-b correlation coefficient ($\tau_{\text{Ken},b}$):

$$\tau_{\text{Ken},b} = (C-D)/\sqrt{[(n(n-1)/2-t)(n(n-1)/2-u)]}$$

where t is the number of tied Y_m values and u is the number of tied Y_{est} values.

- Kendall tau-c correlation coefficient ($\tau_{\text{Ken},c}$):

$$\tau_{\text{Ken},c} = 2(C-D)/n^2$$

Interpretation

- If the agreement between the two rankings is perfect and the two rankings are the same, the coefficient has value 1.
- If the disagreement between the two rankings is perfect and one ranking is the reverse of the other, the coefficient has value -1.
- For all other arrangements the value lies between -1 and 1, and increasing values imply increasing agreement between the rankings.
- If the rankings are independent, the coefficient has value 0.

Related statistics

- $\tau_{\text{Ken},a}^2$ = the squared of Kendall tau-a correlation coefficient;
- $\tau_{\text{Ken},b}^2$ = the squared of Kendall tau-b correlation coefficient;
- $\tau_{\text{Ken},c}^2$ = the squared of Kendall tau-c correlation coefficient.

Statistical significance

Statistical significance of the Kendall's tau correlation coefficient is tested by the Z-test, at a significance level of 5%. The null hypothesis vs. the alternative hypothesis for Kendall's tau correlation coefficients was:

- Kendall tau-a correlation coefficient:

$H_0: \tau_{\text{Ken},a} = 0$ (there is no correlation between the two variables)

$H_1: \tau_{\text{Ken},a} < > 0$ (the two variables are correlated)

- Kendall tau-b correlation coefficient:

$H_0: \tau_{\text{Ken},b} = 0$ (there is no correlation between the two variables)

$H_1: \tau_{\text{Ken},b} < > 0$ (the two variables are correlated)

- Kendall tau-c correlation coefficient:

$H_0: \tau_{\text{Ken},c} = 0$ (there is no correlation between the two variables)

$H_1: \tau_{\text{Ken},c} < > 0$ (the two variables are correlated)

PHP implementation

Kendall function was implemented in order to calculate the Kendall's tau correlation coefficients:

```
function Kendall(&$cmp){
```

```

$n = count($cmp[0]);
$pz = 0;
if(!is_numeric($cmp[0][0])) $pz = 1;
$C = 0;
$D = 0;
$E = 0;
for($i=$pz;$i<$n-1;$i++){
    for($j=$i+1;$j<$n;$j++){
        $sgx = 0;
        $sgy = 0;
        if($cmp[0][$i]>$cmp[0][$j]) $sgx = 1;
        if($cmp[0][$i]<$cmp[0][$j]) $sgx = -1;
        if($cmp[1][$i]>$cmp[1][$j]) $sgy = 1;
        if($cmp[1][$i]<$cmp[1][$j]) $sgy = -1;
        if($sgx*$sgy>0) $C++;
        if($sgx*$sgy<0) $D++;
        if($sgx*$sgy==0) $E++;
        if($sgx==0)$tied_x[$i][]=$j;
        if($sgy==0)$tied_y[$i][]=$j;
    }
}
$t1 = 0;    $u1 = 0;
$vt = 0;    $vu = 0;
$v2t = 0;   $v2u = 0;
if(isset($tied_x))
if(is_array($tied_x)){
    foreach($tied_x as $vx){
        $nt = count($vx)+1;
        $t1 += $nt*(nt-1);
        $vt += $nt*(nt-1)*(2*$nt+5);
        $v2t += $nt*(nt-1)*(nt-2);
    }
}
if(isset($tied_y))
if(is_array($tied_y)){
    foreach($tied_y as $vy){
        $nu = count($vy)+1;
        $u1 += $nu*(nu-1);
        $vu += $nu*(nu-1)*(2*$nu+5);
        $v2u += $nu*(nu-1)*(nu-2);
    }
}
}
$v1 = $t1*$u1;
$t1 /= 2;
$u1 /= 2;
$v2 = $v2t*$v2u;
$S = $C - $D;
$n = $n - $pz;
$scn2 = $n*(n-1)/2;

```

```

$tau_a2 = pow($S,2)/pow($cn2,2);
$v_tau_a = $cn2*(2*$n+5)/9;
$z_tau_a = $S/pow($v_tau_a,0.5);
$T = ($cn2-$t1)*($cn2-$u1);
$tau_b2 = pow($S,2)/$T;
$vT0 = $v_tau_a - ($vt + $vu)/18;
$vT1 = $v1/(4*$cn2);
$vT2 = $v2/(18*$cn2*($n-2));
$v_tau_b = pow($vT0 + $vT1 + $vT2 , 0.5);
$z_tau_b = $S/$v_tau_b;
$gamma = pow(($C - $D)/($C + $D),2);
$v_gamma = (2*$n+5)/9.0/$cn2;
$z_gamma = $gamma/pow($v_gamma,0.5);
$tau_c2 = 4*pow($S,2)/pow($n,4);
$z_tau_c = $z_tau_b*($n-1)/$n;
return array( $tau_a2, $z_tau_a, $tau_b2, $z_tau_b,
$tau_c2, $z_tau_c, $gamma, $z_gamma );
}

```

where C is the number of concordant pairs (C = (<, <) or (>, >)), D is the number of discordant pairs (D = (<, >) or (>, <)), and E is the number of equal pairs (E = (=, .) or (., =)).

Results

- Kendall's τ_a correlation coefficient and associated statistics:

$$\begin{aligned} \tau_{\text{Ken},a}^2 &= 0.6129 \\ Z_{\text{Ken},\tau_a} &= 9.37 \\ p_{\text{Ken},\tau_a} &= 7.44 \cdot 10^{-21} \end{aligned} \tag{4}$$

- Kendall's τ_b correlation coefficient and associated statistics:

$$\begin{aligned} \tau_{\text{Ken},b}^2 &= 0.6177 \\ Z_{\text{Ken},\tau_b} &= 9.37 \\ p_{\text{Ken},\tau_b} &= 7.26 \cdot 10^{-21} \end{aligned} \tag{5}$$

- Kendall's τ_c correlation coefficient and associated statistics:

$$\begin{aligned} \tau_{\text{Ken},c}^2 &= 0.5948 \\ Z_{\text{Ken},\tau_c} &= 9.23 \\ p_{\text{Ken},\tau_c} &= 2.70 \cdot 10^{-20} \end{aligned} \tag{6}$$

Gamma correlation coefficient

Definition

The Gamma correlation coefficient (Γ , gamma) is a measure of association between variables that comparing with Kendall's tau correlation coefficients is more resistant

to tied data [25], being preferable to Spearman rank or Kendall tau when data contain many tied observations [26].

Formula

The formula for Gamma correlation coefficient is:

$$\Gamma = (C-D)/(C+D)$$

where the significance of C and D were described above.

Interpretation

- In the same manner as the Kendall tau correlation coefficient.

Related statistics

- Γ^2 = the squared of Gamma correlation coefficient.

Statistical significance

Statistical significance of Gamma correlation coefficient was tested by the Z-test, at a significance level of 5%. The null hypothesis vs. the alternative hypothesis for Gamma correlation coefficients was:

$H_0: \Gamma = 0$ (there is no correlation between the two variables)

$H_1: \Gamma < > 0$ (the two variables are correlated).

PHP implementation

The function which computes the Gamma correlation coefficient was presented at Kendall's tau correlation coefficient, in PHP implementation section.

Results

$$\Gamma^2 = 0.6208$$

$$Z_{\Gamma} = 7.43$$

$$p_{\Gamma} = 1.11 \cdot 10^{-13}$$

(7)

[25] Goodman L. A., Kruskal W.H., *Measures of association for cross-classifications III: Approximate sampling theory*, J. Amer. Statistical Assoc. 58, 1963, p. 310-364.

[26] Siegel S., Castellan N. J., *Nonparametric Statistics for the Behavioural Sciences*, 2nd Edition, McGraw-Hill, 1988.

Conclusions

All seven computational methods used to evaluate the correlation between measured and estimated by MDF-SAR model inhibitory activity are statistically significant (p-value always less than 0.0001, correlation coefficients always greater than 0.5).

More research on other classes of biologic active compounds may reveal whether it is appropriate to analyze the MDF-SAR models using the Pearson correlation coefficient or other correlation coefficients (Spearman rank, Kendall's tau, or Gamma correlation coefficient).

Antiallergic Activity of Substituted Benzamides: Characterization, Estimation and Prediction

Abstract

Antiallergic activity of twenty-three substituted N 4-methoxyphenyl benzamides was model by the use of an original methodology. After sketching out the compounds structure and creating the file with the observed activities, strictly based on compounds structure, the molecular descriptors family was generated and descriptors entered into a multiple linear regression analysis. The multi-varied model with four descriptors proved to render higher ability in estimation (squared correlation coefficient, $r^2 = 0.9986$) as well as in prediction (cross-validation leave-one-out score, $r^2_{cv-100} = 0.9956$) of antiallergic activity of compounds, obtained significantly greater correlation coefficient compared with the previously reported model ($p < 0.01$). Characterization of antiallergic activity of substituted N 4-methoxyphenyl benzamides by integration of complex structure information provides a stable and efficient multi-varied model with four descriptors. According with the multi-varied model with four descriptors the antiallergic activity of substituted N 4-methoxyphenyl benzamides is like to be of geometry nature, depending by the number of directly bonded hydrogen's, and the atomic relative mass, being in relation with the partial charge of compounds

Keywords

Molecular Descriptors Family on Structure-Activity Relationships (MDF-SAR), Substituted N 4-methoxyphenyl benzamides, Antiallergic activity, Multiple linear regression (MLR)

Background

Benzamide derivatives, known for their anti-inflammatory and immunomodulatory [1,2], anti-tumoral [3], antipsychotic [4], and antiallergic [5] activities, are drugs widely used in medicine [6].

Twenty-three derivatives of N 4-methoxyphenyl benzamide were previously synthesized and their antiallergic activity was tested using dinitrochlorobenzene by inducing delayed allergy of rat skin in vivo [5]. The inhibitory tumor swell rate (IR) was model by the use of molecular connectivity indices, and the following equation was obtained:

ET108/2006 – Et. Unică/2006 – Lucrare in extenso

$$\log IR = 3.002 + 0.8909^4 x_p - 1.3465^5 x_p - 13.8234^6 x_{cg} \quad (1)$$

where $\log IR$ is inhibitory rate expressed in logarithm scale, and 4x_p , 5x_p , $^6x_{cg}$ are molecular connectivity indices.

The statistical characteristics of previously reported model [5] are:

$$r = 0.8865, F = 23, s = 0.572, n = 23 \quad (2)$$

where r = correlation coefficient, F = parameter of Fisher-test, s = standard deviation, and n = sample size.

A CoMFA analysis was also applied by Yu-xin Zhou et al. [5] and the following results were obtained:

$$r = 0.990, r_{cv}^2 = 0.830 \quad (3)$$

where r_{cv}^2 = cross-validated correlation coefficient.

The relationship between antiallergic activities of some substituted N 4-methoxyphenyl benzamides and the information obtained from their structure was studied by the use of an original MDF-SAR methodology. The aim of the research was to analyze the performances of the MDF-SAR methodology in estimation and prediction of antiallergic activities of twenty-three substituted N 4-methoxyphenyl benzamides.

Materials and Methods

N 4-methoxyphenyl benzamides Pharmacology

A number of twenty-three substituted N 4-methoxyphenyl benzamides were included into the study. The generic structure of the substituted N 4-methoxyphenyl benzamides, corresponding substituent(s) of compounds, and inhibitory activity expressed on logarithmical scale are in table 1. In the default cases, $X = Y = Z = T = H$. The inhibitory rate, was calculated by Yu-xin Zhou & all [5] based on the following formula: $(V_1 - V_2)/V$ where V_1 is the swell value of tumor of the reference set of rats (treated with hydrocortisone 20 mg) and V_2 the swell value of tumor of the test set of rats (treated with hydrocortisone 100 mg).

Table 1. Characteristics of studied substituted N 4-methoxyphenyl benzamides and their inhibitory activity

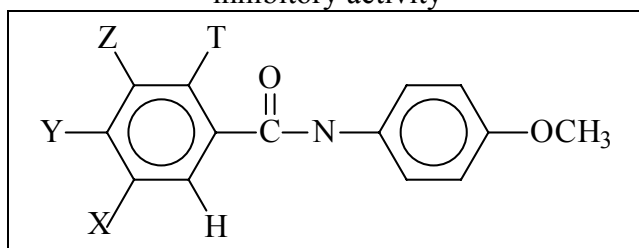


Abb.	Substituent	logIR
com_01	Y = Z = OH	0.13
com_02	X = T = OH	0.07
com_03	T = F	-0.06
com_04	Y = Cl	-0.12
com_05	Y = NO ₂	0.05
com_06	Y = CH ₂ CH ₃	0.03
com_07	Y = <i>t</i> -Bu	0.07
com_08	Y = <i>t</i> -Pro	0.00
com_09	T = OCH ₂ COOH	-0.09
com_10	Y = Z = CH ₂ -mophlinyl; X = OH	-4.00
com_11	Y = OH; Z = CH ₂ -mophlinyl	-4.00
com_12	X = Z = NH ₂	0.03
com_13	Y = OCH ₃	0.05
com_14	T = OCH ₃	0.03
com_15	X = Y = Z = OCH ₃	-0.06
com_16	T = OCH ₂ Ph	-0.80
com_17	T = OCH ₂ CHCH ₂	-0.04
com_18	Z = OC ₄ H ₉	-0.39
com_19	Y = OC ₈ H ₁₇	0.11
com_20	Y = OCH ₂ CHCH ₂	-0.34
com_21	T = SH	-0.18
com_22	Z = SH	-0.25
com_23	Y = SH	0.07

MDF-SAR methodology

The MDF-SAR methodology applied on substituted N 4-methoxyphenyl benzamides consisted of the following steps [7]:

- 1: Sketch out 3D structure of each substituted N 4-methoxyphenyl benzamides compounds by the use of HyperChem software [8];
- 2: Create of the file with measured Antiallergic Activity (logIR) of substituted N 4-methoxyphenyl benzamides compounds;
- 3: Generate the molecular descriptors family (MDF) members for substituted N 4-methoxyphenyl benzamides compounds [9,10]. All twenty-three compounds were used in generation of the molecular descriptors family. The algorithm of generation the molecular descriptors family was strictly based on compounds structure. The process of molecular descriptors family generation was followed by a filtration in which there were deleted from databases identical descriptors by imposing a significance selector equal with 10⁻⁹ (the redundant information was clear away). The name of each molecular descriptor refers its calculation mode and includes: compound geometry or topology (the 7th letter), atomic property (cardinality, number of directly bonded hydrogen's, atomic relative mass, atomic

ET108/2006 – Et. Unică/2006 – Lucrare in extenso

electronegativity, group electronegativity, partial charge - the 6th letter), the atomic interaction descriptor (the 5th letter), the overlapping interaction model (the 4th letter), the fragmentation criterion (the 3rd letter), the molecular selector (the 2nd letter), and the linearization function applied in molecular descriptor generation (the 1st letter).

4: Find and identify the MDF-SAR models.

5: Validation of the obtained MDF-SAR models were performed through computing the cross-validation leave-one-out correlation score (r_{cv}^2) [11]. The cross-validation leave-one-out correlation score was compute by exclusion one time and applying to all sample one compound from dataset, rebuilding the MDF-SAR model and estimation of excluded compound activity based on MDF-SAR model.

6: Analyze the selected MDF-SAR models through: squared correlation coefficients, statistical parameters of estimation and prediction analysis, model stability analysis (the differences between squared correlation coefficient and cross-validation leave-one-out score - the lowest value correspond to the most stable model), and correlated correlation analysis [12] by comparing the results of the MDF-SAR models with previously reported models.

Results

The best performing multi-varied MDF-SAR models (one with two-descriptors and one with four descriptors) are:

- MDF-SAR model with two-descriptors: $\hat{Y}_{2d} = -8.8 \cdot 10^{-3} - 5.1 \cdot 10^{-5} \cdot isDRtHg + 0.13 \cdot iHMMtHg$ (4)

- MDF-SAR model with four-descriptors: $\hat{Y}_{4d} = -0.15 + 9 \cdot 10^{-4} \cdot imMRkMg - 0.32 \cdot imMDVQg - 5.2 \cdot 10^{-5} \cdot isDRtHg + 0.14 \cdot iHMMtHg$ (5)

where \hat{Y}_{2d} respectively \hat{Y}_{4d} are estimated log IR by the MDF-SAR model with two, respectively with four molecular descriptors, and *isDRtHg*, *iHMMtHg*, *imMRkMg*, *imMDVQg* are molecular descriptors. Statistical characteristics of the MDF-SAR models are presented in table 2 and 3.

Table 2. Statistical characteristics of MDF-SAR models for antiallergic activity of substituted N 4-methoxyphenyl benzamides

Characteristic (notation)	Value	
Number of variable (v)	2	4
Correlation coefficient (r)	0.9942	0.9986
Squared correlation coefficient (r^2)	0.9884	0.9973
Adjusted squared correlation coefficient (r_{adj}^2)	0.9872	0.9967

Standard error of estimated (s_{est})	0.1300	0.0664
Fisher parameter (F_{est})	848*	1638*
Cross-validation leave-one-out (loo) score (r^2_{cv-loo})	0.9864	0.9956
Fisher parameter for loo analysis (F_{pred})	725*	1007*
Standard error for leave-one-out analysis (s_{loo})	0.1405	0.0847
Model stability ($r^2 - r^2_{cv(loo)}$)	0.0019	0.0016

* $p < 0.001$

The experimental values (logIR) and values predicted by MDF-SAR models with two (\hat{Y}_{2d}), respectively with four-descriptors (\hat{Y}_{4d}) and previously reported model (\hat{Y}_{CoMFA}) are in figure 1.

The absolute differences between estimated by models (\hat{Y}_{2d} , \hat{Y}_{4d} , and \hat{Y}_{CoMFA}) and measured (logIR) antiallergic activities of substituted N 4-methoxyphenyl benzamides were used in order to obtain the best estimation (figure 2).

Table 3. Regression analysis of the MDF-SAR models

	<i>StdError</i>	<i>t Stat</i>	<i>95%CI_{coefficient}</i>
<i>MDF-SAR model with four descriptors</i>			
Intercept	0.037	-3.985*	[-0.224, -0.0693]
imMRkMg	0.000	6.917*	[0.0006, 0.0012]
imMDVQg	0.042	-7.628*	[-0.411, -0.2337]
isDRtHg	$7 \cdot 10^{-7}$	-77.12*	$[-5.4 \cdot 10^{-5}, -5 \cdot 10^{-5}]$
iHMMtHg	0.002	64.03*	[0.1345, 0.1437]
<i>MDF-SAR model with two descriptors</i>			
Intercept	0.031	-0.288*	[-0.073, 0.0552]
isDRtHg	$1 \cdot 10^{-6}$	-40.719*	$[-5.4 \cdot 10^{-5}, -5 \cdot 10^{-5}]$
iHMMtHg	0.003	36.678*	[0.1226, 0.1374]

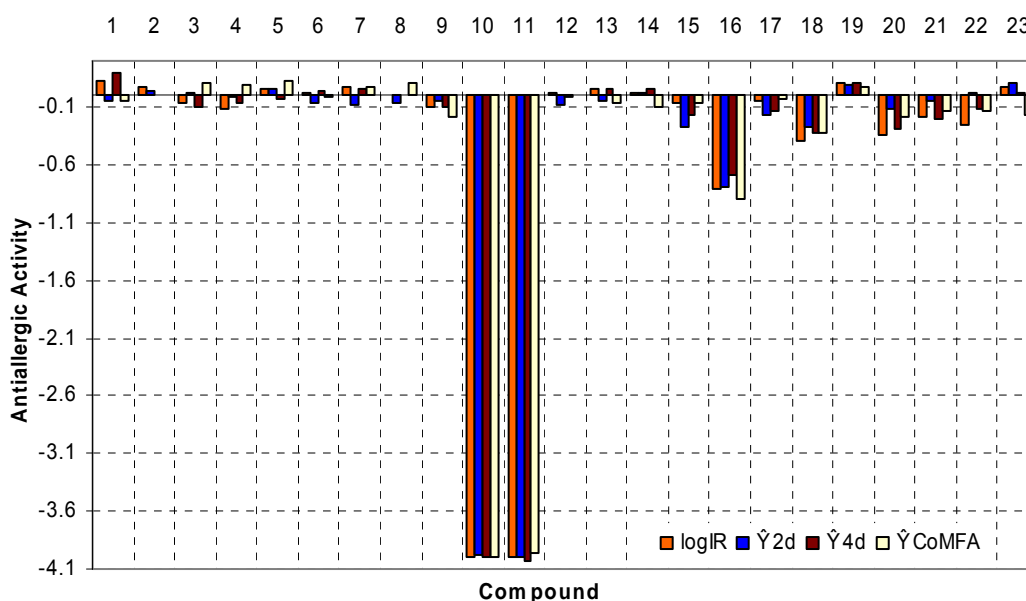


Figure 1. Measured antiallergic activity (logIR) and estimated by MDF-SAR, respectively CoMFA models

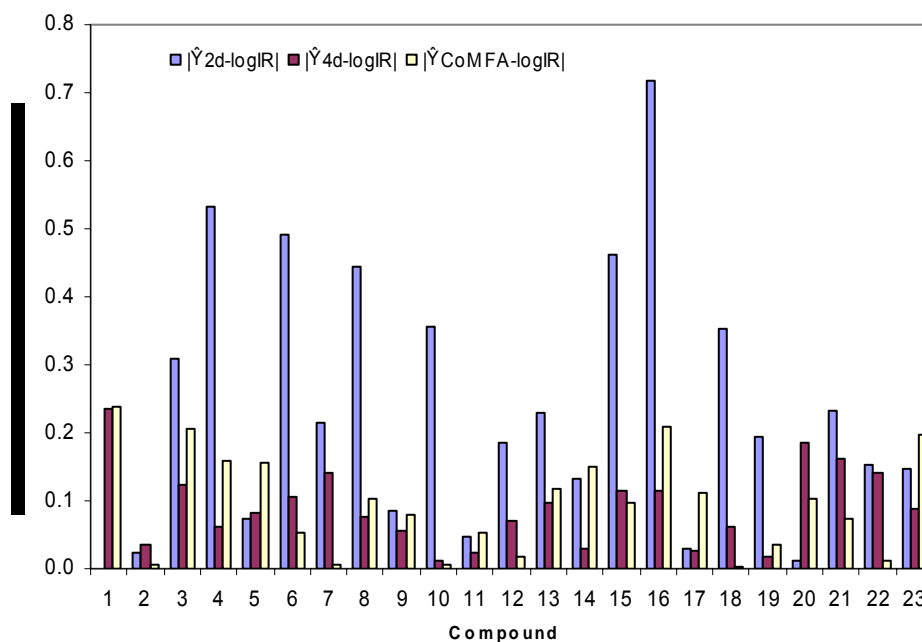


Figure 2. Measured (logIR) and estimated activities of compounds using MDF-SAR models and previously reported CoMFA model

In eleven out of twenty-three cases, the best estimation is obtained by multi-varied MDF-SAR model with four descriptors (see figure 3).

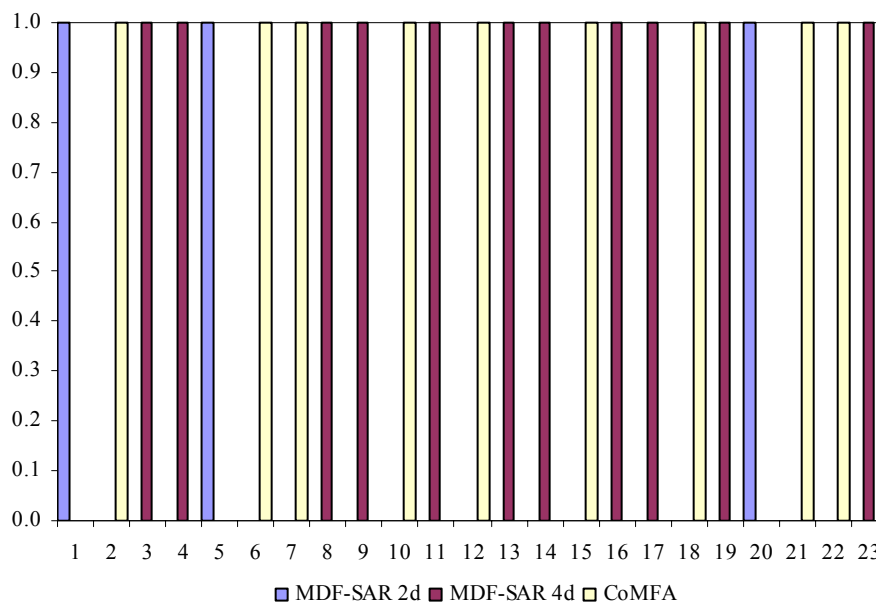


Figure 3. Best estimation antiallergic activity by MDF-SAR and CoMFA models

The comparison of MDF-SAR models with previously reported CoMFA model was performed by applying a correlated correlation analysis and the results are presented in table 4.

Table 4. The results of comparison between MDF-SAR model with four descriptors and previously reported CoMFA model

Characteristic	Value	
Number of descriptors used in MDF SAR model	2	4
$r(\log IR, \hat{Y}_{MDF-SAR})$	0.9941	0.9986
$r(\log IR, \hat{Y}_{CoMFA})$	0.9952	0.9952
$r(\hat{Y}_{MDF-SAR}, \hat{Y}_{CoMFA})$	0.9945	0.9943
Steiger's Z parameter	-0.5113	2.7974*

* $p < 0.05$

Discussions

Antiallergic activity of twenty-three substituted N 4-methoxyphenyl benzamides was characterized by the use of an original methodology, based on complex structure information of the compounds in order to explain associated biological activity.

Two multi-varied MDF-SAR models, one with two descriptors and other with four descriptors, proved to obtained performances in antiallergic activity estimation and prediction. The MDF-SAR models are statistically significant at a significance level less than 0.001 (see table 2).

The MDF-SAR model with two descriptors uses two molecular descriptors which take into consideration the geometry (**g**) and the number of directly bonded hydrogen's (**H**) of compounds (see Eq. 4). Almost ninety-nine percent of variation in antiallergic activity of substituted N 4-methoxyphenyl benzamides can be explainable by its linear relation with *isDRtHg* and *iHMMtHg* descriptors. The correlation coefficient obtained by the MDF-SAR model with two descriptors is not statistical significant different by the previously reported CoMFA model (see table 4) at a significance level of 5%. The performance of the MDF-SAR model with two descriptors is sustained by the correlation coefficient and the squared of the correlation coefficient ($r = 0.9942$, $r^2 = 0.9942$, table 2); the stability of the model is proved by the very lower value of the differences between squared correlation coefficient and cross-validation leave-on-out squared correlation coefficient. The cross-validation leave-one-out score ($r^2_{cv-loo} = 0.0019$) sustain the stability of the MDF-SAR model with two descriptors and its prediction abilities.

Looking at the MDF-SAR model with two descriptors it can be say that the antiallergic activity of studied compounds is of molecular geometry and is strongly depend on the number of directly bonded hydrogen's.

Analyzing the cross-validation leave-one-out scores, it can be said that multi-varied MDF-SAR model with four descriptors is the best performing MDF-SAR model. Almost one

hundred percent of variation in antiallergic activity of substituted N 4-methoxyphenyl benzamides can be explainable by its linear relation with four molecular descriptors. Both descriptors used in MDF-SAR model with two descriptors can be found again on model with four descriptors; the other two descriptors consider the geometry of the molecule (**g**), atomic relative mass (**M**) and the partial charge (**Q**) as atomic property with role in antiallergic activities.

Looking at the multi-varied MDF-SAR model with four descriptors it can be observed that the antiallergic activities of studied compounds is positive correlated with *imMRkMg* and *iHMMtHg* descriptors and negative correlated with *imMDVQg* and *isDRtHg* descriptors. The values of squared correlation coefficient ($r^2 = 0.9973$) demonstrate the goodness of fit of the multi-varied MDF-SAR model with four descriptors (see tables 2 and 3, figures 2 and 3). The power of the MDF-SAR model with four descriptors in prediction of antiallergic activity of substituted N 4-methoxyphenyl benzamides compounds is demonstrate by the cross-validation leave-one-out correlation score ($r^2_{cv(100)} = 0.9956$). The stability of the MDF-SAR model with four descriptors is give by the difference between the squared correlation coefficient and the cross-validation leave-one-out correlation score ($r^2 - r^2_{cv(100)} = 0.0016$). Analyzing multi-varied MDF-SAR model with four descriptors it can be said that antiallergic activities of substituted N 4-methoxyphenyl benzamides strongly depend on the geometry of the compounds and is in relation with number of directly bonded hydrogen's, atomic relative mass and partial charge of compounds.

Correlated correlations analysis results (see table 4) demonstrate that the multi-varied MDF-SAR model with four descriptors obtained a significantly greater correlation coefficient compared with the previously reported CoMFA.

Starting with knowledge learned from the studied set of substituted N 4-methoxyphenyl benzamides, antiallergic activity of new compound from the same class can be predict by the use of an original software [13]. After the user draw the chemical structure of the new compound and saved it as *.hin file, the software is able to predict the antiallergic activity of new substituted N 4-methoxyphenyl benzamides compound in real time, without any experiments.

Conclusions

Modeling the antiallergic activity of substituted N 4-methoxyphenyl benzamides by integration of complex structural information provide a stable and performing MDF-SAR model with four variables, allowing to make remarks about relation between structure of compounds and their activities.

The antiallergic activity of substituted N 4-methoxyphenyl benzamides is like to be of geometry nature, depending by the number of directly bonded hydrogen's, and the atomic relative mass, being in relation with the partial charge of compounds.

References

- [1] Hatzelmann A, Schudt C. Anti-inflammatory and immunomodulatory potential of the novel PDE4 inhibitor roflumilast in vitro. *J Pharmacol Exp Ther* 2001;297(1):267-79.
- [2] Carbonnelle D, Ebstein F, Rabu C, Petit JY, Gregoire M, Lang F. A new carboxamide compound exerts immunosuppressive activity by inhibiting dendritic cell maturation. *Eur J Immunol* 2005;35(2):546-56.
- [3] Suzuki K, Nagasawa H, Uto Y, Sugimoto Y, Noguchi K, Wakida M & all. Naphthalimidobenzamide DB-51630: A novel DNA binding agent inducing p300 gene expression and exerting a potent anti-cancer activity. *Bioorg Med Chem* 2005;13(12):4014-21.
- [4] Simonini MV, Camargo LM, Dong E, Maloku E, Veldic M, Costa E, Guidotti A. The benzamide MS-275 is a potent, long-lasting brain region-selective inhibitor of histone deacetylases. *Proc Natl Acad Sci U S A* 2006; 31;103(5):1587-92.
- [5] Yu-xin Zhou, Lu Xu, Ya-ping Wu, Bai-li Liu. A QSAR study of the antiallergic activities of substituted benzamides and their structures. *Chemometrics and Intelligent Laboratory Systems* 1999;45:95-100.
- [6] Malík I, Sedlářová E, Andriamainty F, Csöllei J. Relationship between structure and biological activity of benzamide derivatives. *Farmaceuticky Obzor* 2006; 75(1):3-9.
- [7] Jäntschi L., *Molecular Descriptors Family on Structure Activity Relationships 1. The review of Methodology*, Leonardo Electronic Journal of Practices and Technologies 2005;6:76-98.
- [8] ***, HyperChem , Molecular Modelling System [Internet page]; ©2003, Hypercube, Inc. [cited 2005 May]. Available from: URL: <http://hyper.com/products/>

- [9] Diudea M, Gutman I, Jäntschi L. *Molecular Topology*, Nova Science, Huntington, New York 2001.
- [10] Jäntschi L, Katona G, Diudea M. Modeling Molecular Properties by Cluj Indices, *Commun Math Comput Chem* 2000;41:151-88.
- [11] ***, Leave-one-out Analysis. ©2005, Virtual Library of Free Software [cited 2005 Nov]. Available from: URL: http://vl.academicdirect.org/molecular_topology/mdf_findings/loo/
- [12] Steiger JH. Tests for comparing elements of a correlation matrix. *Psychol Bull* 1980;87:245-51.
- [13] ***, MDF SAR Predictor, © 2005, Virtual Library of Free Software. [cited 2006 February]. Available from: URL: http://vl.academicdirect.org/molecular_topology/mdf_findings/sar

Activitatea Anti-Alergică a Derivaților de Benzamide: Caracterizare, Estimare și Predicție

Rezumat

Activitatea antialergică a unui eșantion de douăzeci și trei de derivați N 4-methoxyphenyl benzamide a fost caracterizată, estimată și prezisă prin folosirea unei metode originale. Familia descriptorilor moleculari care a stat la baza obținerii modelelor a fost generată strict pe baza structurii compușilor, după crearea structurii tri-dimensionale a acestora și a fișierului cu activitatea anti-alergică măsurată. Modelul multi-variabil cu patru descriptori s-a dovedit a abilități atât în estimarea ($r^2 = 0.9986$) cât și în prezicerea ($r^2_{cv-100} = 0.9956$) activității anti-alergice a compușilor studiați, obținând un coeficient de corelație semnificativ mai mare în comparație cu modelul raportat în literatura de specialitate ($p < 0.01$). Caracterizarea activității anti-alergice a derivaților N 4-methoxyphenyl benzamide prin integrarea informațiilor structurale complexe, oferă un model cu patru variabile stabil și eficient. Activitatea anti-alergică a derivaților N 4-methoxyphenyl benzamide este de natură geometrică, depinde de numărul legăturilor directe de hidrogen și masa atomică relativă, fiind în relație cu sarcina parțială.

Cuvinte cheie: Familia de descriptori moleculari, Relații structură-activitate (MDF-SAR), derivați N 4-methoxyphenyl benzamide, Activitate anti-alergică, Regresie liniară multiplă

Mobile Phase Optimization in Three Solvents High Performance Thin-Layer Chromatography: Methodology and Evaluation

Synopsis

A mobile phase optimization program, based on an original mathematical approach, was developed in order to optimize mobile phase composition of high performance thin-layer chromatography with a mixture of three solvents. The mathematical approach, implement two equations with six and respectively seven parameter, taking into consideration the solvents molar fraction. Three chromatographic parameters were included into optimization process: objective function, resolution and retention factor. Starting with the mathematical optimization model, a program was developed and its abilities in optimization of mobile phase were analyzing on two classes of compound, two sets of steroids and one set of N-alkyl phenothiazine sulfones. The obtained results sustain the accuracy of the optimization procedure, proving to be a precise and reproducible method, opening a new pathway in analyzing and characterization of HPTLC parameters.

Keywords

High Performance Thin-Layer Chromatography (HPTLC), Three Solvents, Mobile Phase Optimization

Introduction

Chromatographic analysis, define as techniques used for the separation of a mixture of compounds by their distribution between two phases, was invented in 1901 by Russian botanist Mikhail Semyonovich Tsvet, during his research on plant pigments [1]. He used liquid-adsorption column chromatography with calcium carbonate as adsorbent and petrol ether/ethanol mixtures as eluent to separate chlorophylls and carotenoids. He described his method at the XI Congress of Naturalists and Doctors in St. Petersburg in 1901 and he used the term of chromatography for the first time in 1906 in a paper published in *Berichte der Deutschen Botanischen Gesellschaft* journal [2].

There are many types of chromatography as Column Chromatography, Thin Layer Chromatography (TLC), and Gas Chromatography (GC) [3]. The chromatography is used in chemistry [4,5], biology [6,7], and medicine [8,9] fields as an analytical techniques.

An important task in separation of compounds from a mixture by chromatography is choosing of the proper mobile phase [10], this task being time consuming. Some researchers studied optimization methods for column liquid chromatography [11] and for high performance liquid chromatography [12,13].

Starting with results previously obtained in optimization of the mobile phase of chromatography separation [14,15], the aim of the research was to develop and to assess an original mathematical model for mobile phase optimization with applicability on High Performance Thin-Layer Chromatography.

Method

Mathematical Model

Into a mixture of three solvents, the quantitative measure of choused chromatographic parameter depends on the composition of mobile phase through a dependence equation, which can be one of two forms:

$$M6(x_1, x_2, x_3) = a_1x_1 + a_2x_2 + a_3x_3 + a_4x_1x_2 + a_5x_1x_3 + a_6x_2x_3 \quad \text{Eq.(1)}$$

$$M7(x_1, x_2, x_3) = a_1x_1 + a_2x_2 + a_3x_3 + a_4x_1x_2 + a_5x_1x_3 + a_6x_2x_3 + a_7x_1x_2x_3 \quad \text{Eq.(2)}$$

where x_1, x_2, x_3 are molar fraction of the three solvents ($x_1 + x_2 + x_3 = 1$), M6 and M7 are estimators and then predictors of choused chromatographic parameter, and $a_1, a_2, a_3, a_4, a_5, a_6, a_7$ are coefficients first determined based on the best estimation of choused chromatographic parameter and then used in prediction of used parameter for any composition of mobile phase. Starting from the above presented equations (Eq.(1) and Eq.(2)) the following chromatographic parameters were modeled:

- $RF(i, e) = l(i)/l(e) \quad \text{Eq.(3)}$

where i is one of the separation compounds, e is the eluent used as mobile phase, $l(i)$ is the coordinate at which was migrated in e eluent, $l(e)$ is the coordinate at which the eluent was migrated, and RF is the series of retention factor of separation compounds for eluent e .

- $RS(i, j, e) = 2*(l(i)-l(j))/(w(i)+w(j)) \quad \text{Eq.(4)}$

where i, j are two separation compounds, $w(i)$ and $w(j)$ the width of the compound's spots, and RS is the matrix of calculated resolution for separation of i compound by j compound.

ET108/2006 – Et. Unică/2006 – Lucrare in extenso

- $RSO(i,e) = 2*(lo(i)-lo(i+1))/(w(i)+w(i+1))$ Eq.(5)

where $lo(i)$ is the i migration coordinate in the ordered list of migration length, and RSO is the matrix of resolutions ordered for separation of consecutive compounds.

- $Sm(e) = \sqrt{(\sum_j(\Delta RFT-\Delta RF(j,e)))/\sqrt{(n+1)}}$ Eq.(6)

where n is the total number of compounds which must be separate, ΔRFT is the retention ideal value in $1/n$ separation, $\Delta RF(j,e)$ is the differences of i 's between two consecutive retention factors from the retention factors ordered list, and Sm is the separation mean recorded for the eluent e .

- $RSA(e) = \sum_j RSO(j,e)/n$ Eq.(7)

where RSA is the mean of resolution of separation using the e eluent.

- $RRP(e) = \prod_j RSO(j,e)/RSA(e)$ Eq.(8)

where RRP is the ponderate product with the mean of resolutions used in separation by the use of e eluent.

- $Inf(e,m) = \sum_k (n_k/n) \log_2(n_k/n)$ Eq.(9)

where n_k is the number of compounds which migrate into at k out of m equidistant interval (the total length of migration $l(e)$ was split into m equidistant intervals), and Inf is the quality factor of separation calculated by Logit method, quality factor which is null for an ideal separation.

- $FOB(e,m) = \sum_j a_j F_j(Sm(e), Inf(e,m), RSA(e), RRP(e))$ Eq.(10)

where $1 \leq j \leq 4$, F_j are functions which each conceive one expression of four parameters, a_j are coefficients choused arbitrary or through of ponderate relation mathematic defined, and FOB is an objective function which characterized the separation with e eluent in report with selection of coefficients a_j , of F_j functions and respectively of number of equidistant intervals m .

By application of one of the above describe equations (Eq.(3)-Eq.(10)) on a series of p experiments, result a M_{ob} matrix with one (Eq.(6)-Eq.(10)) ore more than one (Eq.(3)-Eq(5)) rows, one for each experiment. The elements of M_{ob} matrix represent the values of chromatographic parameter which is modeled by using Eq.(1) or Eq.(2). The optimization algorithm has a unique determine solution for $p \geq 6$ for Eq.(1) and respectively for $p \geq 6$ for Eq.(2).

Optimization procedure

For each row of M_{ob} matrix is build a system with p linear equations with six or seven terms (Eq.(1), Eq.(2)) in a_j coefficients as following example:

$$M_{ob}(j) = a_1x_1 + a_2x_2 + a_3x_3 + a_4x_1x_2 + \dots \quad \text{Eq.(11)}$$

where x_i are molar fraction of each solvent ($I = 1, 2, 3$) which enter into the composition of the e_j eluent ($j = 1, 2, \dots, p$).

To the above describe system (Eq.(11)) the least squared method is applied for construction of the system with unique determine solution MMCP, which is obtained by applying the following formula:

$$MMCP(k,0) = M2(MOB,A(k)), MMCP(k,1) = M2(A(k),A(l)) \quad \text{Eq.(12)}$$

where $(k,0) = 1, 2, \dots, 6$ for Eq.(1) and $(k,0) = 1, 2, \dots, 7$ for Eq.(2), $A(k)$ is the series of terms known from Eq.(11), $M2$ calculate the mean for the product of MOB series and $A(k)$, and $MMCP$ is the extended matrix of system of linear equations which is used in determination of a_k coefficients.

For determination of the solution for Eq.(12) is applied the Gaussian method. The solution for the system from Eq.(1) and Eq.(2) are:

$$\begin{aligned} A0 &= (a_{01}, a_{02}, \dots, a_{06}) - \text{for Eq.(1)} \\ A0 &= (a_{01}, a_{02}, \dots, a_{07}) - \text{for Eq.(2)} \end{aligned} \quad \text{Eq.(13)}$$

At one time as the coefficients $A0$ are determined, theirs values are used for prediction of the chromatographic parameter of interest by using one of the equations Eq.(1) or Eq.(2), and it is used the equation for which the coefficients were determined.

For example if Y is the choused chromatographic parameter, the MOB matrix (the predictor of Y) has more than one row as well as the estimator of Y . If z is the number of MOB matrix rows (and implicit the number of predictors) then it can be state the estimator of choused chromatographic parameter \hat{Y} as followings:

$$\hat{Y} = (\hat{y}_1, \dots, \hat{y}_z) \quad \text{Eq.(14)}$$

The optimum is obtained by application of a maximization or minimization function (as is for example the characterization of a separation of many compounds through the worst separation of two compounds):

$$\hat{y}_o = \text{opt}(\hat{Y}), \text{ where opt} = \text{"max"} \text{ or opt} = \text{"min"} \quad \text{Eq.(15)}$$

Moving through all domains of possible values for the composition of the mobile phase, the optimum point is identified, this being the optimum composition of mobile phase (x_1, x_2, x_3):

$$(\cdot, \cdot, \cdot) | \hat{Y}(\cdot, \cdot, \cdot) = \text{opt} \{ \hat{Y}(i/100, j/100, k/100) | i=0..100, j=0..100-i, k=100-i-j \} \text{ Eq.(16)}$$

Software implementation

Starting with the experience acquired with implementation of a fitting statistical regressions application [16], the above describe mathematical model and optimization procedure were integrated into an online application by the use of the PHP (Pre Hypertext Processed) language [17]. PHP has the unique distinction of being an open-source scripting language, which proved its usefulness in creation and development of interactive and dynamic applications [18,19].

Materials

A number of three sets of compounds (two sets steroids and one set of N-alkyl phenothiazine sulfones) previous studied were included into analysis (see table 1). The optimization procedure described above was applied on these three classes of compounds.

The abbreviation of the class, the compounds of class, the solvents and the optimum mobile phase previously obtained are in table 1.

Table 1. Characteristic of sets included into analysis

Abb.	Compounds	Solvents	Optimum Mobile Phase	Ref.
steroids_01	metazepan, napoton nitrazepan, oxazepan diazepam	Chloroform Iso-Propanol Acetone	73: 26: 1	[15]
steroids_02	metazepan, napoton nitrazepan, oxazepan diazepam	Chloroform Methyl-Ethyl-Cetone Cyclohexan	52.5: 25.5: 22	[20]
thiazine_01	N-CH ₃ -PhT-S, N-C ₂ H ₅ -PhT-S, N-C ₃ H ₇ -PhT-S, N-C ₄ H ₉ -PhT-S, N-C ₅ H ₁₁ -PhT-S, N-iC ₅ H ₁₁ -PhT-S, N-C ₇ H ₁₅ -PhT-S	Toluene Ethyl Ether Chloroform	30:50:20	[21]

Results

Mobile Phase Optimization Program

The application of optimization of mobile phase for chromatographic separation which used mixture's of three solvents was created, and can run on any computer connected to the Internet, being available at the following URL:

http://vl.academicdirect.org/molecular_dynamics/mobile_phase_opt/

The main features of the application are:

- Mobile phase optimization on high performance thin-layer chromatography with mixture of three solvents;
- Assisting users in choosing the solvents and/or the compounds;
- Assisting users in choosing one out of three optimized chromatographic parameter with best performance in separation.

Optimization of HPTLC chromatographic parameters on three sets of compound

The retention factor was considered as an important parameter into chromatography and the results of optimization refer this factor. The optimum mobile phase previous reported was took into consideration in choosing the best performing optimization model for the three sets of compounds included into analysis.

In all three sets of compounds, the optimum mobile phase was obtained with the following generic equation:

$$\Delta rf = a_1x_1 + a_2x_2 + a_3x_3 + a_4x_1x_2 + a_5x_1x_3 + a_6x_2x_3 + a_7x_1x_2x_3 \quad \text{Eq.(17)}$$

The optimum phases obtained through optimization procedure are in Table 2.

Table 2. Characteristics of the optimum mobile phase

Abb.	Optimum Mobile Phase
steroids_01	91: 0: 9
steroids_02	90: 10: 0
thiazine_01	41: 23: 36

The values of retention factor obtained experimental (Exp.) and by the used of the optimization method (Est.) for all three sets of compounds are in Table 3-5.

Table 3. Experimental and optimized retention factor for data set steroids_01

Type	Solvent's mixture	Metazepan	Napoton	Nitrazepan	Oxazepan	Diazepan
Exp.	0	0.8169	0.8648	0.8895	0.9186	0.9390
Exp.	1	0.7151	0.7878	0.8706	0.8081	0.8706
Exp.	2	0.8441	0.7654	0.8584	0.8455	0.8798
Exp.	3	0.4849	0.5509	0.6700	0.7877	0.8278
Exp.	4	0.3769	0.5855	0.7254	0.7733	0.8225
Exp.	5	0.8214	0.8525	0.9188	0.9296	0.9513
Exp.	6	0.7736	0.8469	0.8691	0.9149	0.9241
Est.	0	0.8169	0.8648	0.8895	0.9186	0.9390
Est.	1	0.7151	0.7878	0.8706	0.8081	0.8706
Est.	2	0.8441	0.7654	0.8584	0.8455	0.8798

Est.	3	0.4849	0.5509	0.6700	0.7877	0.8278
Est.	4	0.3769	0.5855	0.7254	0.7733	0.8225
Est.	5	0.8214	0.8525	0.9188	0.9296	0.9513
Est.	6	0.7736	0.8469	0.8691	0.9149	0.9241

Table 4. Experimental and optimized retention factor for steroids_02 set

Data type	Solvent's mixture	Metazepan	Napoton	Nitrazepan	Oxazepan	Diazepan
Exp.	0	0.7020	0.7980	0.7801	0.3146	0.4258
Exp.	1	0.2121	0.3156	0.2926	0.0230	0.0564
Exp.	2	0.8395	0.8852	0.8829	0.6945	0.7213
Exp.	3	0.3969	0.5730	0.5162	0.0669	0.1550
Exp.	4	0.0663	0.1127	0.0862	0.0188	0.0254
Exp.	5	0.7010	0.7783	0.7559	0.4502	0.5241
Exp.	6	0.7570	0.7979	0.7886	0.4182	0.4708
Exp.	7	0.0000	0.0000	0.0000	0.0000	0.0000
Exp.	8	0.9309	0.9365	0.9342	0.8071	0.7770
Exp.	9	0.1066	0.2211	0.1757	0.0261	0.0317
Est.	0	0.7003	0.8189	0.7967	0.3005	0.4168
Est.	1	0.2313	0.2829	0.2704	0.0732	0.1077
Est.	2	0.9101	0.9539	0.9472	0.6738	0.7079
Est.	3	0.3149	0.4403	0.3975	0.1024	0.1590
Est.	4	0.0843	0.1536	0.1213	0.0000	0.0123
Est.	5	0.7046	0.7908	0.7668	0.4497	0.5291
Est.	6	0.7323	0.7824	0.7725	0.4136	0.4625
Est.	7	0.0000	0.0129	0.0081	0.0000	0.0000
Est.	8	0.8918	0.8926	0.8940	0.8224	0.7869
Est.	9	0.1529	0.2900	0.2379	0.0098	0.0318

Table 5. Experimental and optimized retention factor for thiazine_01 set

Data type	Solvent's mixture	N-CH3 -PhT-S	N-C2H5 -PhT-S	N-C3H7 -PhT-S	N-C4H9 -PhT-S	N-C5H11 -PhT-S	N-iC5H11 -PhT-S	N-C7H15 -PhT-S
Exp.	0	0.0340	0.0470	0.0610	0.0590	0.0810	0.0830	0.0770
Exp.	1	0.2670	0.2700	0.3320	0.3760	0.4250	0.4260	0.4480
Exp.	2	0.4830	0.5940	0.8120	0.8760	0.9140	0.9080	0.9170
Exp.	3	0.5960	0.6720	0.7680	0.8180	0.8370	0.8360	0.8690
Exp.	4	0.2240	0.2200	0.2730	0.2620	0.2620	0.2380	0.2330
Exp.	5	0.4300	0.5170	0.6420	0.7070	0.7460	0.7390	0.7980
Exp.	6	0.5420	0.5850	0.7000	0.7380	0.7800	0.7710	0.8470
Est.	0	0.0340	0.0470	0.0610	0.0590	0.0810	0.0830	0.0770
Est.	1	0.2670	0.2700	0.3320	0.3760	0.4250	0.4260	0.4480
Est.	2	0.4830	0.5940	0.8120	0.8760	0.9140	0.9080	0.9170
Est.	3	0.5960	0.6720	0.7680	0.8180	0.8370	0.8360	0.8690
Est.	4	0.2240	0.2200	0.2730	0.2620	0.2620	0.2380	0.2330
Est.	5	0.4300	0.5170	0.6420	0.7070	0.7460	0.7390	0.7980
Est.	6	0.5420	0.5850	0.7000	0.7380	0.7800	0.7710	0.8470

ET108/2006 – Et. Unică/2006 – Lucrare in extenso

For all five compounds there were not identify a statistical significant differences ($p > 0.05$) between experimental (Exp.) and optimized (Est.) retention factor for the set steroids_01.

There were not identified any statistically significant differences between experimental retention factor and the values obtained by the used of proposed optimization method for the steroids_02 set ($p > 0.05$).

The representation of the estimated through optimization procedure versus experimental data for Metazepan, steroids_02 set, is in figure 1.

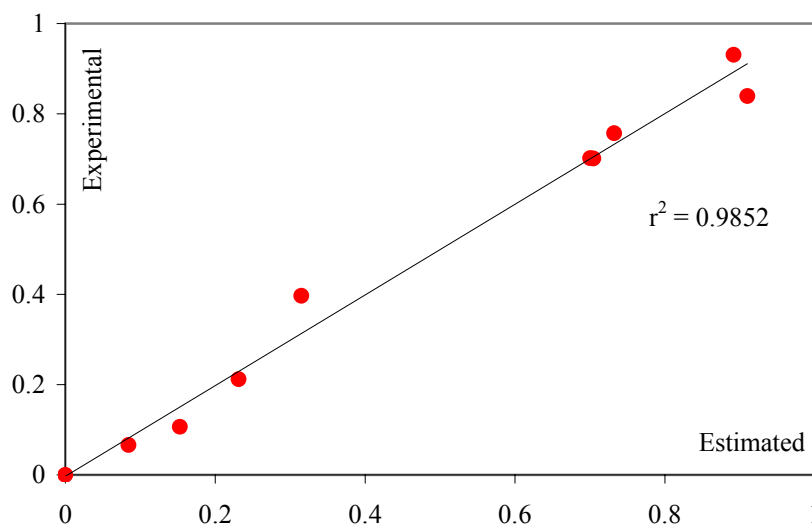


Figure 1. Estimated versus experimental retention factor for Metazepan, steroids_02 set

There were not identified any statistically significant differences between experimental retention factor (Exp.) and the values obtained by the used of proposed optimization method (Est.), for thiazine_01 set ($p > 0.05$).

The plot generate by the application, for the thiazine_01 set obtained with the Eq.(17) for a Z_{\min} equal with 0.001 created by the use of 25 colors is in figure 2.

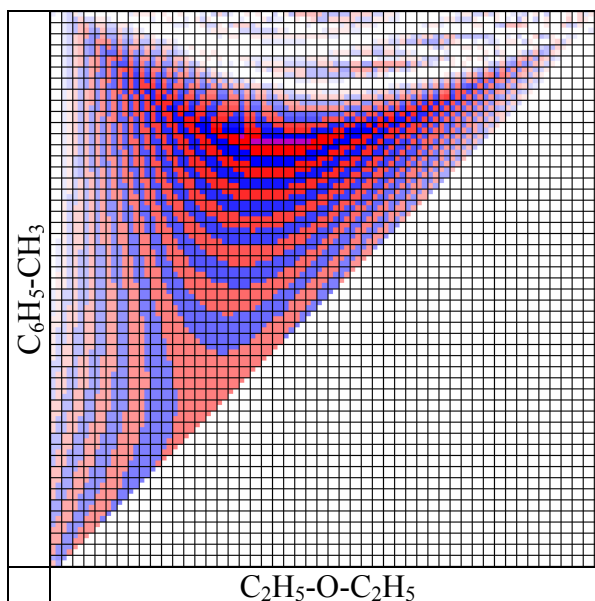


Figure 2. Plot of the optimized retention factor for thiazine_01 set

Discussions

The characteristics of the Mobile Phase Optimization Program are:

⊕ Interaction: the user has possibility of choosing the solvents of interest and also the compounds which want to be separated. This information's are used just for the presentation of the optimization results.

⊕ Control of experiment(s): the user's can choose one or more than one compositions of experimental mobile phase. Taking this option into consideration the users' has the control on the optimization, choosing the experiment or experiments of interest.

⊕ Control of optimization: the mathematical model and the chromatographic parameter(s) can be chosen by the user. The optimization process is based on a model with six and/or seven parameters (Eq.(1), Eq.(2)) and is applied one or more than one of the following chromatographic parameters: objective function, resolution and retention factor.

⊕ Optimization results: the results of optimization are clear and easy to find. The program display:

- The name of the optimized chromatographic parameter;
- The equation of model, the values of chromatographic parameter as they results from the experiment as well as the values obtained by applying of the equation results from the modeling;

ET108/2006 – Et. Unică/2006 – Lucrare in extenso

- The list of the whole possible estimated by the optimized model values for the chromatographic parameter of interest for each possible mobile phase with an one percent unit increment;
- The optimum mobile phase calculated by the model and the values of the computed chromatographic parameter of interest in this point;
- Access to a graphic interface: represent graphically as surface level the dependence of predicted chromatographic parameter on mobile phase composition by the use of four specification choused by the user: two out of three solvents choused in order to be presented on 0x and 0y axis, the minimum value of the chromatographic parameter of interest which to appear on the graph, and the intermediary hue of two colors used for graphical representation of surface level. The program create a graphical representation or *.png (Portable Network Graphics) type;
- The possibility of saving the results as *.txt file.

The application is multi-user; allow being use simultaneously by more than one user. The program is accessible to anyone who is interested in, its use being restricted just by the existence of a computer connected to the Internet. The Mobile Phase Optimization Program is a flexible program allowing repeating the optimization process as many time as it is necessary, in a short time comparing with a chromatographic separation, and ensuring of obtaining the same results.

In order to become a useful tool in separation of compounds by HPTLC with three solvents, the abilities of the program must be analyzed on more compounds sets, the objective which is in our future plan of program development.

Searching the available specialty literature it can be observed that there are just a few articles which present and described the optimization process of HPTLC [19,22]. Present paper described the results obtained for optimization of three chromatographic parameters, the objective function, resolution and retention factor. The optimization results were presented just for the retention factor because it was consider that this chromatographic parameter is the one which is most important in compounds' separation on HPTLC.

Analyzing the results obtained for optimization of the retention factors three observations can be perform. First observation refers the generic equation of optimization of retention factor which in all three sets of compound is an equation with seven parameters (see Eq.(17)). Second observation refers the optimum mobile phase. In all three sets of compounds it can be

ET108/2006 – Et. Unică/2006 – Lucrare in extenso

observed that there are not statistically significant differences between retention factor obtained through experiments and those obtained by after applying of the optimization procedure ($p > 0.05$, see Table 3-5). Third observation refers the composition of the optimum mobile phase. Even if the optimum mobile phase obtained experimentally (see Table 1) are not the same with those obtained through optimization (see Table 2) the experimental and estimated values of the retention factors are not statistical different ($p < 0.05$). It is well known that the retention factor is a constant from one experiment to the next one only if the chromatography conditions represented by the solvent system, adsorbent and its thickness, amount of material spotted and temperature are constant. Comparing with the experimental procedure, the optimization procedure produces the same results of a choused set of compounds at every application.

Conclusions

The proposed optimization procedure opens a new pathway in analyzing and characterization of three chromatographic parameters of HPTLC analysis which used a mixture of three solvents.

Mobile Phase Optimization Program, proved to assure accurate results regarding the retention factors analyzed on three sets on compounds. The program can become a useful instrument in characterization of HPTLC parameters, opening the possibility of development of online library of optimized HPTLC parameters.

References

- [1] Senchenkova, E. M.; In: Gillispie (Ed.). Dictionary of scientific biography; Charles Scribner Sons: New York, 1976, 13, 486-488.
- [2] Tswett, M. Berichte der Deutschen botanischen Gesellschaft 1906, 24, 316-323.
- [3] Fishbein, L. Journal of Chromatography 1974; 98(1): 177-251
- [4] Duarte, A. C.; Capelo, S. J Liq Chromatogr Related Technol 2006, 29(7-8), 1143-1176.
- [5] Papageorgiou, V. P.; Assimopoulou, A. N., Samanidou, V. F., Papadoyannis, I. N. Curr Org Chem 2006, 10(5), 583-622.
- [6] Miyake, T.; Yafuso, M. Plant Species Biol 2005, 20(3), 201-208.
- [7] De Abreu, I. N.; Sawaya, A. C. H. F.; Eberlin, M. N.; Mazzafera, P. In Vitro Cell Dev Biol Plant 2005, 41(6), 806-811.

ET108/2006 – Et. Unică/2006 – Lucrare in extenso

- [8] Sharma, L.; Desai, A.; Sharma, A. *Biochem Mol Biol Educ* 2006, 34(1), 44-48.
- [9] Reddy, B. S.; Rozati, R., Reddy, B. V. R.; Raman, N. V. V. S. S. *BJOG* 2006, 113(5), 515-520.
- [10] Mulja, M.; Indrayanto, G. In: Cazes, J. (Ed.), *Encyclopedia of Chromatography*; Marcel Dekker: New York, 2001, 794-797.
- [11] Loeser, E.; Babiak, S.; Zhu, P.; Yowell, G.; Konigsberger, M.; Drumm, P. *Chromatographia* 2006, 63(7-8), 345-351.
- [12] Zhang, Y. P.; Zhang, Y. J.; Gong, W. J.; Gopalan, A. I.; Lee, K.-P. *J Chromatogr A* 2005, 1098(1-2), 183-187.
- [13] Dharmadi, Y.; Gonzalez, R. *AIChE Annual Meeting, Conference Proceedings* 2005, 9056.
- [14] Cimpoiu, C.; Jäntschi, L.; Hodisan, T. *J Planar Chromatogr - Mod TLC* 1998, 11, 191-194.
- [15] Cimpoiu, C.; Jäntschi, L.; Hodisan, T. *J Liq Chromatogr Related Technol* 1999, 22(10), 1429-1441.
- [16] Jäntschi, L. *LJS* 2002, 1, 31-52.
- [17] ZEND. The PHP Company [homepage on the Internet]. © 1999-2005 <http://zend.com/zend/docs.php>, 2006 March
- [18] Bolboaca, S. D.; Jäntschi, L.; Deneş, C., Achimaş Cadariu, A. *Roentgenologia & Radiologia* 2005, XLIV(3), 189-193.
- [19] Drugan, T.; Bolboacă, S. D.; Jäntschi, L.; Achimaş Cadariu, A. *LEJPT* 2003, 3, 45-74.
- [20] Jäntschi, L.; Hodişan, S.; Cimpoiu, C.; Ceteraş, I. *Acta Universitatis Cibiniensis Seria F Chemia* 2005, 8, 67-76.
- [21] Cimpoiu, C.; Hodişan, S.; Toşa, M.; Paizs, C.; Majdik, C.; Irimie, F.-D. *J. Pharm Biomed Anal* 2002, 28(2), 385-389.
- [22] Coran, S. A.; Giannellini, V.; Bambagiotti-Alberti, M. *J Chromatogr A* 2004, 1045(1-2), 217-222.

Cromatografia planară (TLC)

O variantă mai simplă a LC

Cromatografia planară (PC) întâlnită adesea sub denumirea de cromatografie în strat subțire este cea mai *simplă și ieftină* dintre toate metodele cromatografice cunoscute. Mai este denumită și *cromatografia de lichide a săracului*.

În literatura de specialitate se utilizează mai multe denumiri (cu prescurtările asociate acestora). Pe lângă termenul consacrat - cromatografia planară (planar chromatography - PC [27]) se mai întâlnesc denumiri ca cromatografia în strat subțire de înaltă performanță (high performance thin layer chromatography - HPTLC) sau cromatografia de lichide planară (planar liquid chromatography - PLC).

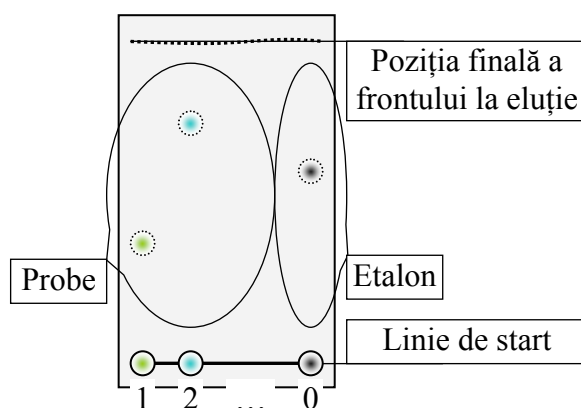


Fig. 1. Placa cromatografică și pozițiile relative ale spoturilor

În această variantă a *cromatografie de lichide*, separarea nu mai are loc într-o coloană închisă ci pe o fază staționară similară, granulară (poroasă) dispusă într-un *strat subțire*, formând un plan (fig. 1). Acest strat denumit *subțire*, se realizează dintr-un adsorbent cu grosimi cuprinse între 100-250 μ m și poate fi simplu sau legat adeziv de un plan rigid, fiind pe tot parcursul separării în contact cu o fază gazoasă - mai mult sau mai puțin saturată cu vapori de eluent. Primul caz a fost mult utilizat în trecut în cazul așa-numitei cromatografii pe hârtie unde faza staționară consta dintr-o bandă de hârtie de filtru, confecționată din celuloză pură, sau, extrem de rar, prin folosirea unor plăci din materiale ceramice sinterizate poroase.

[27] Din l. engleză

ET108/2006 – Et. Unică/2006 – Lucrare în extenso

Celălalt caz mult mai utilizat (chiar în zilele noastre) face apel la straturi subțiri realizate dintr-un adsorbent pulverulent (silicagel, celuloză, alumina, poliamidă sau derivate ale acestora) dispuse în straturi subțiri pe plăci rigide din sticlă sau, pe folii flexibile din aluminiu, poliester sau alte materiale inerte față de sistemul pe care are loc separarea. Se mai poate recurge și la straturi formate pe baghete de sticlă sau tuburi din sticlă.

Obținerea straturilor subțiri se realiza la început în laborator, pornindu-se de la o suspensie apoasă a adsorbentului pulverulent (10-40 μ m) împreună cu un liant anorganic (de exemplu ghips, SiO₂-coloidal) sau organic (amidon, carboximetilceluloză), iar pentru aplicare se foloseau niște dispozitive mecanice simple. Straturile subțiri mai pot include și indicatori de fluorescență care în lumină UV fac posibilă vizualizarea spoturilor substanțelor care absorb în acest domeniu prin stingerea fluorescenței, adică prin apariția unor spoturi întunecate pe fond luminos. Straturile subțiri pot fi achiziționate gata preparate de la firme producătoare specializate (Camag, Cole-Parmer etc.).

Aplicarea probei

Pentru a se putea demara procesul de eluție, în prealabil pe placa cu strat subțire se aplică proba. Acest lucru se realizează cu seringi sau micro-pipete, dar și alte dispozitive specializate (fig. 2), astfel încât să se obțină aliniată, pe linia de start (fig. 1), mai multe spoturi de probe, respectiv de amestecuri etalon - supuse simultan separării. Întrucât probele se aplică din soluții diluate (1-2%), în anumiți solvenți, pentru a se evita interferența acestora în procesul de eluție, plăcile se usucă înainte de introducerea în amestecul de solvenți. Astfel se pot supune separării probe care se concentrează pe zone înguste (de lungime și lățime preselectate) asigurându-se o eficiență mărită separării.

Migrarea eluentului

Are loc prin coloana deschisă, care acționează cu totul analog celei închise (vezi LC), are loc sub acțiunea forțelor capilare și provoacă migrarea diferențiată a componentelor amestecului de separat. Acest lucru se realizează în urma simplei scufundări (manuale) a plăcii cromatografice în eluentul potrivit. Din acest moment, eluentul irigând prin capilaritate stratul poros migrează ascendent prin stratul subțire, provocând separarea. Timpul de separare variază între 3 și 60 min.

Nu este totuși exclusă utilizarea unor dispozitive mai sofisticate de alimentare cu solvenți (minipompe) - făcându-se uneori apel chiar la gradienti de concentrație (v. par. 16.5 în cazul LC). Se poate de asemenea practica migrarea eluentului pe orizontală sau descendent. Se poate elua o placă chiar de mai multe ori sau se poate evapora solventul în timpul migrării, în acest fel mărindu-se eficiența pe seama timpului de separare.

Pentru realizarea separării (eluției) se utilizează *camere de dezvoltare*. Câteva dintre cele mai utilizate dintre acestea se prezintă în fig. 3. Formele preferate sunt cele paralelipipedice sau cilindrice (pahare), prevăzute cu un capac și eventual cu un dispozitiv de fixare a plăcii plane (hârtie sau strat subțire) și pot fi saturate cu amestecul de solvenți din tanc, pentru a se mări viteza de eluție. Camerele paralelipipedice (fig. 3N) se numesc camere de tip N (normale) iar cele subțiri (fig. 3S) poartă numele de camere de tip S (sandwich). Ultimele prezintă avantajul unui volum mai redus al fazei gazoase având o viteză de saturare mai mare și o durată a procesului de separare cu ceva mai mică.

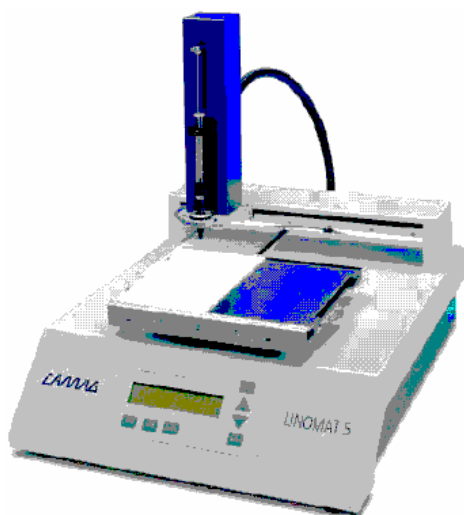


Fig. 2. Aparatul Linomat: aplică probele în formă de benzi pe straturile subțiri

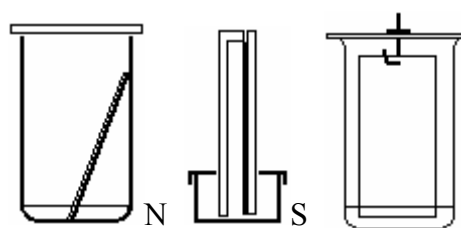


Fig. 3. Câteva dintre cele mai uzuale camere cromatografice în PC

Eluția are loc după introducerea plăcii în camera cromatografică, până când amestecul de solvenți atinge o înălțime finală, fixată de obicei între 5 și 18 cm, sau un anumit timp stabilit, în prealabil, prin încercări preliminare. Solvenții utilizați se aleg în funcție de proprietățile de eluție ale substanțelor supuse separării (analizei). Natura acestora depinde nu numai de substanțele implicate dar și de mecanismul de separare propus, respectiv de faza staționară avută la dispoziție.

După eluție, placa se scoate, se usucă și dacă spoturile nu se văd, se trece la vizualizarea acestora (operație numită uneori revelare). Pentru aceasta, placa fie se scufundă într-un reactiv, fie se pulverizează cu acesta sau se introduce într-o atmosferă conținând gaze reactive și chiar într-o etuvă, la cald, când spoturile devin vizibile în urma unor reacții chimice (având un aspect apropiat celor din fig. 4).



Doar apoi se poate trece la etapa analizei propriu-zise. Reactivii de culoare pot fi generali, ca de exemplu acidul sulfuric la cald (120°C), care determină carbonizarea majorității substanțelor organice, sau specifici, când aceștia reacționează doar anumite substanțe sau funcțiuni organice.

O altă variantă de a face spoturile vizibile, preferată tot mai mult în ultimul timp, este folosirea unor *straturi subțiri fluorescente* (spre exemplu materialul pulverulent conține ZnS - fluorescent). Prin examinarea cromatogramelor eluate și uscate în lumină UV, se vor observa spoturi închise la culoare sau colorate, pe fond fluorescent - luminos. Atunci când chiar substanțele separate sunt fluorescente nu mai este nevoie de fondul fluorescent și este suficientă o placă obișnuită, observată în lumină UV, spoturile devenind luminoase pe un fond întunecat (fig. 4).

O variantă și mai modernă, foarte eficientă de vizualizare, folosind tot lumină UV, constă în utilizarea unor plăci cu straturi subțiri conținând *amestecuri de luminofori*. În acest caz plăcile au o culoare compusă (de regulă lumina emanând din trei substanțe luminescente având culori diferite). Cum fiecare luminofor emite la o altă lungime de undă, iar substanțele separate absorb diferit lumina, fiecare component de pe placă va avea, în consecință, o altă culoare. În acest caz detecția este mai sigură pentru că nu diferă doar poziția relativă a spotului respectiv pe placă ci și culoarea.

Faze mobile și faze staționare

Analog cu cromatografia de lichide, respectiv cu varianta HPLC [28], în PC sunt posibile mai multe mecanisme de separare: (1) cromatografia de adsorbție, (2) cromatografia de repartiție (cu faze directe și cu faze inversate), (3) cromatografia de schimb ionic și (4) cromatografia de excluziune sterică. Aceleași faze staționare (silicagel, alumină, celuloză sau derivați ai acestora) se utilizează în granulații fine (<40μm) dar fără părțile extra fine (<1μm). În ultimul timp câștigă tot mai mult teren fazele chimice legate, pentru separări în cromatografia de repartiție cu faze inversate (solvenți polari și faze staționare nepolare). Ca și în HPLC, se utilizează tot silicagelul silanizat având grefate grupări alchilice conținând 8 sau 18 atomi de carbon - echivalente cu o peliculă subțire de fază nepolară depusă pe granula-suport.

Fazele mobile sunt adesea amestecuri de 2 până la 5 solvenți cât mai diferiți ca natură chimică, aleși prin încercări preliminare. Pentru straturi subțiri din silicagel (cele mai utilizate) prezentăm o listă cu 10 solvenți extrași din 8 grupe de selectivitate (proapse de Snyder, un cercetător american care a clasificat solvenții organici în funcție de mai mulți parametri structurali și fizici), care diferă între ele prin structura chimică: eter etilic (grupa I), izopropanol și etanol (II), tetrahidrofuran (III), acid acetic (IV), diclormetan (V), acetat de etil și dioxan (VI), toluen (VII) și cloroform (VIII). Hexanul este un solvent considerat complet nepolar și este recomandat a fi introdus în amestecuri pentru aducerea spoturilor în domeniul considerat optim, de $R_f = 0.2-0.8$ dacă mai este necesar.

Analiza chimică prin PC

Analiza calitativă

Fiecare compus separat prin PC este caracterizat (calitativ) de parametrul de retenție denumit R_f (o prescurtare de la termenul din l. engl.: *retardation factor*). Acesta se calculează astfel:

$R_f = (\text{Distanța parcursă de componentul dat, } i) / (\text{Distanța parcursă de frontul solventului})$
sau

$$R_f = \frac{x_i}{x_0}$$

unde cu x_i , respectiv cu x_0 , s-au notat distanțele parcurse de componentul dat, respectiv de frontul solventului, până la oprirea cromatogramei. Toate mărimile calculate pentru o *coloană*

[28] HPLC = High Performance Liquid Chromatography

ET108/2006 – Et. Unică/2006 – Lucrare in extenso

închisă au o mărime corespondentă și în PC. De exemplu parametrul de retenție R devine R_f (f de la front):

$$R_f = \frac{x}{x_0} = \frac{\bar{v}}{v_0} = \frac{t_0}{t} = \frac{1}{k+1}$$

unde v reprezintă viteze de migrare, t - timpi de migrare iar k factorul de capacitate. De aici, implicit:

$$k = \frac{1}{R_f} - 1$$

Comparându-se valorile R_f ale spoturilor etaloanelor cu cele ale substanțelor din proba de analizat (lucru realizat de cele mai multe ori vizual) se spune că se face analiza calitativă. Aceasta este cea mai utilă aplicație a metodei PC. Se pot identifica pesticide din ape, sol dar și alte substanțe pentru scopuri științifice și tehnice cu o sensibilitate mult mai bună decât în eprubetă. Desigur că sensibilitatea atinsă este inferioară celei din *HPLC*. De aceea, în aplicațiile analitice de performanță, cu toată simplitatea și prețul de cost scăzut, PC nu poate înlocui întotdeauna *HPLC*.

Tot în scopul realizării analizei calitative se poate practica și *cromatografia bidimensională*. În această variantă (fig. 5), se utilizează o placă pătrată, iar spotul (unul singur) se aplică într-unul dintre colțuri, de exemplu: în dreapta-jos. Se eluează cu un amestec de solvenți 1, având loc o primă separare. Apoi, după oprire se usucă placa. Se rotește placa cu 90° și se începe o nouă irigare a acesteia folosind un alt amestec de solvenți, 2. Doar apoi placa se vizualizează și se compară cu o placă etalon - pe care avem un amestec similar dar cunoscut.

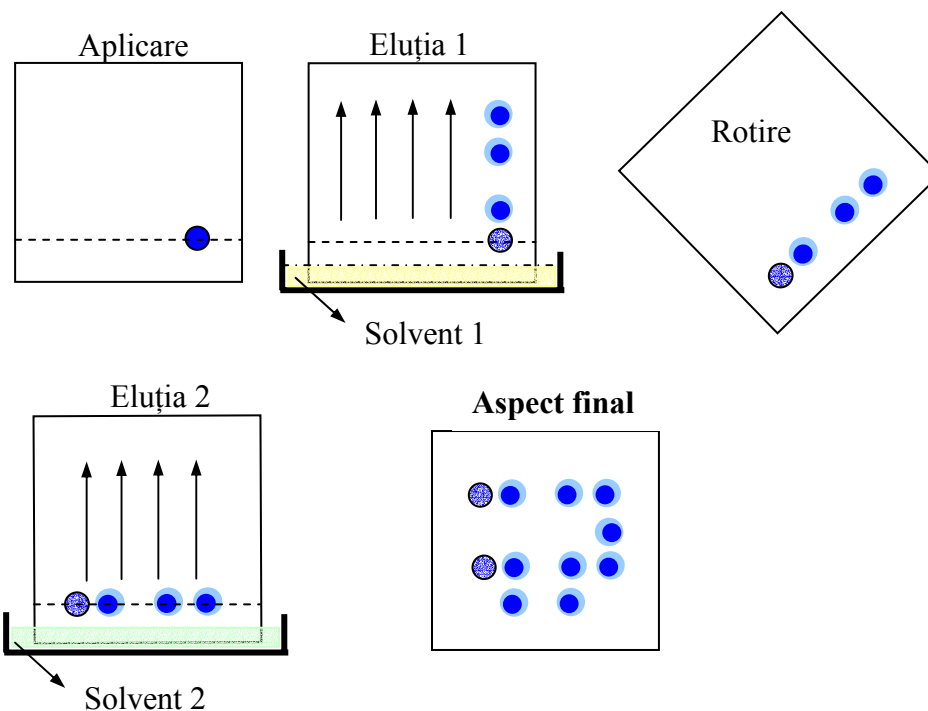
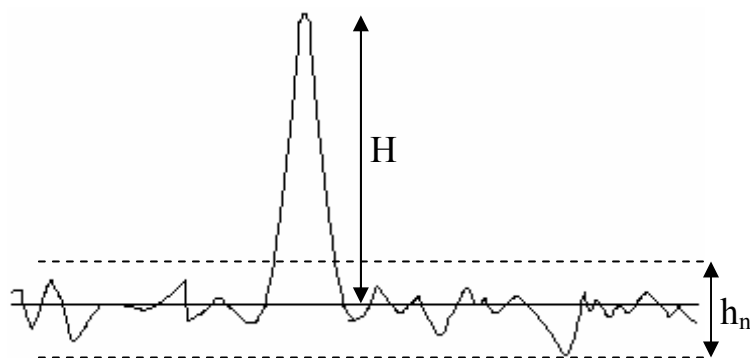


Fig. 5. Modul de executare a cromatografiei în strat subțire bidimensional

Analiza cantitativă poate începe după efectuarea separării și vizualizării sau prin executarea unei măsurători asupra unui parametru fizic (radioactivitate) proporțional cu concentrația substanței din spotul de interes analitic. Metodele practice pe scară largă sunt *densitometria* prin transmisie sau reflexie a spotului (spoturilor), respectiv mai recent scanarea. Densitometria se poate realiza în lumină vizibilă sau UV, în ultimul caz fiind posibilă și măsurarea radiației de fluorescență. Semnalul măsurat servește la construirea unei curbe de etalonare din semnalele măsurate pentru diferitele spoturi, conținând cantități crescătoare de component, aplicate, alături de proba necunoscută, pe aceeași placă. Se preferă *metoda curbei de etalonare* în reflexie (adică prin măsurarea reflectanței) deoarece domeniul liniar al metodei este destul de îngust și adesea curba de etalonare este neliniară. Motivul principal pentru care rezultatele analizei cantitative prin PC sunt inferioare HPLC sunt neuniformitățile stratului subțire care contribuie în mod hotărâtor la mărirea raportului semnal/zgomot, notat S/N. Astfel, conform fig. 6, unde H este înălțimea picului (semnalului) analitului iar h_n - lățimea domeniului în care variază zgomotul de fond (intervalul $\pm 2\sigma$).



$$S/N = \frac{H}{h_n/2} = \frac{2H}{h_n}$$

Fig. 6. Ilustrarea mărimilor care afectează raportul semnal/zgomot

O variantă mai simplă dar mai laborioasă de analiză cantitativă constă în *spălarea* zonei corespunzătoare analitului de pe suport într-un pahar, folosind un solvent adecvat și apoi determinarea, cu o altă metodă instrumentală, a concentrației soluției rezultate.

Evaluarea cu ajutorul *foto-densitometrelor* este până în prezent cea mai utilizată metodă. În aceste instrumente, placa cromatografică cu stratul pe ea, se deplasează odată cu suportul (de regulă în direcția dezvoltării), prin fața fantelor sistemului optic de iluminare, respectiv de măsurare a intensității luminii reflectate. Schema unui astfel de dispozitiv se poate observa pe fig. 7.

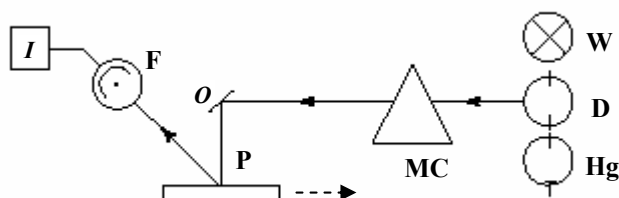


Fig. 7. Reprezentarea schematică a unui densitometru

Pe această figură se pot remarca posibilitatea de utilizare a unei lămpi cu incandescență - cu halogen - (*W*), pentru determinări în domeniul vizibil (400-800nm), a unei lămpi cu deuteriu (*D*) pentru spoturile care absorb în UV (190-400nm) sau a unei lămpi UV cu intensitate mai ridicată (cu xenon sau cu vapori de mercur), notată *Hg*, pentru spoturile fluorescente. Lumina monocromatică care părăsește monocromatorul *MC* se reflectă pe oglinda *O*, se reflectă pe placa cu strat subțire *P* iar lumina reflectată este receptată pe fotomultiplicatorul *F*. Semnalul obținut este înregistrat de înregistratorul *I*, care poate fi chiar un calculator.

După înregistrare, *densitogramele* se evaluează cantitativ, fie prin măsurarea înălțimii picurilor, fie a suprafeței acestora. Oricare dintre mărimile măsurate poate constitui semnalul analitic. Trasarea graficului semnal analitic în funcție de cantitatea de substanță din spot, duce la o *curbă de etalonare*. În cazul determinărilor de fluorescență acest grafic este liniar.

Un exemplu de determinare cantitativă din domeniul controlului poluării mediului îl constituie determinarea seleniului din ape. În acest caz înainte de aplicarea probei pe placă, aceasta se supune unei reacții chimice numită derivatizare. Derivatizarea seleniului se poate realiza cu 2,3-diaminonaftalină (DAN). Compusul rezultat este fluorescent. Pe fig. 8a se prezintă densitograma unei probe de apă conținând seleniu derivatizat. Se poate observa că reactivul de derivatizare, DAN, apare separat pe aceeași placă alături de compusul rezultat cu seleniul din apă. Repetându-se separarea cu mai multe probe cunoscute se obțin picuri cu înălțimi diferite (fig. 8b). Limita de detecție este pentru această metodă de 250fg Se. Reprezentarea grafică a înălțimilor duce la curba de etalonare (fig. 8c).

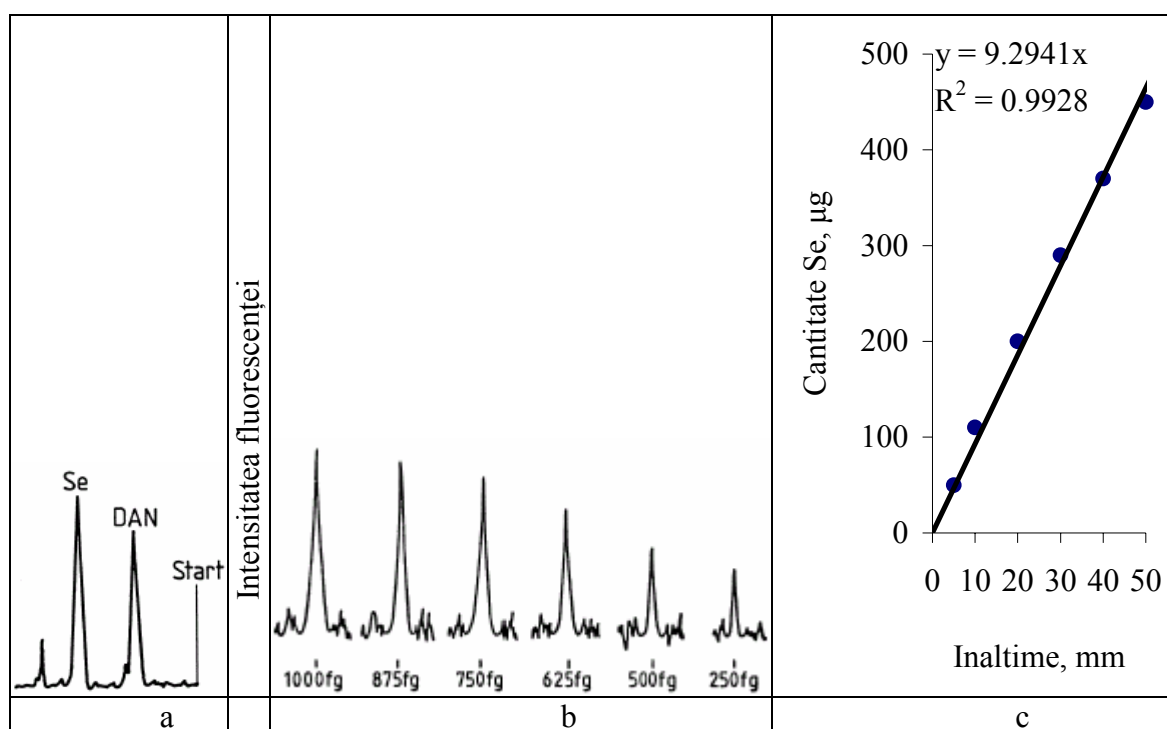


Fig. 8. (a) Aspectul unei densitograme a compusului fluorescent al seleniului DAN (1,2-diaminonaftalină); (b) Densitograme rezultate din măsurătorile de fluorescență ale produsului derivatizării seleniului din probe de apă cu 1,2-diaminonaftalină, în cantități crescătoare ($1\text{fg}=10^{-12}\text{g}$); (c) Aspectul unei curbe de etalonare în PC

Metoda PC este simplă și ieftină dar, ca precizie și exactitate, este adesea inferioară HPLC. În calitate de *metodă semicantitativă*, PC este o metodă competitivă și în lipsa unei

ET108/2006 – Et. Unică/2006 – Lucrare in extenso

aparaturi performante (de exemplu HPLC) rămâne uneori singura alternativă. O analiză *semicantitativă* poate răspunde la întrebarea: *Este prezentă specia X într-o concentrație mai mare decât o concentrație limită, C?* De foarte multe ori răspunsul la această întrebare este suficient în practica curentă. Dar pentru monitorizarea automată a conținutului de poluanți din ape sau sol, pentru expertize sau pentru controlul alimentelor, doar metoda PC nu este suficientă.

Concluzii

Integrarea informațiilor complexe provenite din experiment este permise efectuarea de predicții foarte bune în ceea ce privește comportamentul substanțelor chimice cu structură chimică bine definită.

Pentru fenomenele care se petrec la interfața fazelor, așa cum este cazul la cromatografia pe strat subțire, integrarea mai multor modele, așa cum face funcția obiectiv, aduce un supliment de certitudine în predicțiile efectuate.

Următoarele aplicații realizate și-au dovedit capabilitatea predictivă:

- Optimizarea fazei mobile la amestecuri de 3 solvenți în cromatografia planară și de lichide de înaltă performanță:

http://vl.academicdirect.org/molecular_dynamics/mobile_phase_opt/

- Cinetica reacțiilor simple și complexe:

http://vl.academicdirect.org/molecular_dynamics/reaction_kinetics/

- Modelarea activității biologice prin descriptori de natură cuantică semiempirici:

http://vl.academicdirect.org/molecular_topology/mdf_findings/

Cluj-Napoca,
la 28.10.2006

Director temă ET108/2006,
Șef. L. Dr., Ing. Lorentz JĂNTSCHI
<http://lori.academicdirect.org>
