UNIVERSITATEA "BABEŞ-BOLYAI" CLUJ-NAPOCA
FACULTATEA DE CHIMIE ŞI INGINERI CHIMICĂ

REZUMAT

TEZĂ DE DOCTORAT

STUDII DE SIMILARITATE TOPOLOGICĂ

CONDUCĂTOR ŞTIINŢIFIC
Prof.Dr. MIRCEA V. DIUDEA

DOCTORAND
CRISTINA DORINA MOLDOVAN

2012

CUPRINS

Cuvinte cheie: matrice, indice, descriptori topologici, activitate biologică, QSAR, analize de date, APC, statistica, coeficient de corelare, similaritate, androstan, benzimidazol, benzoxazol, SIMIL, TOPOCLUJ, HyperChem, DRAGON

INTRODUCERE

Matematizarea chimiei are o lungă şi diversificată istorie care începe acum aproape două secole. În 1874, Alexander Crum Brown (1838-1922), unul dintre fondatorii teoriei structurii chimice a prezis următoarele: "...chimia va deveni o ramură a matematicii aplicate; dar aceasta nu va înceta a fi o ştiinţă experimentală. Matematica ne va permite o privire retrospectivă pentru a justifica rezultatele obţinute de experiment, ce va fi utilă în cercetări şi câteodată chiar la prezicerea în întregime a noilor descoperiri. Noi nu ştim când schimbarea va avea loc sau dacă aceasta va fi treptată sau bruscă...". Această prezicere pare a fi în curând împlinită.

Proiectarea de structuri moleculare cu proprietăţi fizico-chimice ori biologice dorite reprezintă una din principalele obiective ale diferitelor ramuri ale industriei (chimice, farmaceutice, etc.), în special şi a cercetatorului în particular. Realizarea unui produs nou cu o anumită proprietate chimică sau activitate biologică, presupun mari cheltuieli atât material cât şi umane. Design-ul de medicamente este un proces iterativ care începe cu un compus care afişează un profil interesant biologic şi se termină cu optimizarea profilului de activitate pentru molecula de sinteză. Procesul este iniţiat atunci când chimistul concepe o ipoteză care leagă caracteristicile chimice ale moleculei (sau serie de molecule), de activitatea biologică. Fără o înţelegere detaliată a procesului biochimic, responsabil pentru activitatea biologică, în general, ipoteza este rafinată prin examinarea asemănărilor şi diferenţelor structurale pentru moleculele active şi inactive.

În acest sens, prezenta teză dezvoltă conceptele utilizate pentru a determina diferite modele QSPR/QSAR precum şi proiectarea unor noi structuri pe baza acestora. Aceasta este structurată în două parţi distincte. Prima parte se referă la prezentarea domeniul topologiei moleculare, a metodelor şi tehnicilor folosite, iar a doua parte se referă la contribuţiile personale în modelarea relaţiilor structură-proprietate şi structură-activitate biologică:

Capitolul I "Similaritate topologica" prezintă conceptul şi metodele de similaritate moleculară.

Capitolul II "Matrici topologice", prezentă pe lângă matricile clasice, noi matrici de tip Szeged şi Cluj.

Capitolul III "Indici topologici" tratează descriptorii topologici obţinuţi pe baza matricilor topologice. Sunt prezentaţi principalii indici topologici utilizaţi şi noi indici topologici propuşi, SP, indici de tip Szeged şi Cluj.

Capitolul IV "Relaţii cantitative structură– activitate biologică" tratează diferitele metode şi modele utilizate pentru descrierea proprietăţilor fizico-chimice şi biologice ale compuşilor.

Capitolului V "Analiza de date " se referă la metodele de prelucrare statistică utilizate în proiectarea modelelor QSPR/QSAR. De asemenea sunt prezentate câteva aplicaţii ale acestora pe diferite seturi de structuri.

În partea de contribuţii proprii sunt prezentate rezultatele obţinute ca urmarea a analizei unor seturi de structuri chimice, cu ajutorul metodelor de similaritate, precum şi prezicerea proprietaţilor unor compuşi noi cu potenţială activitate biologică.

CAPITOLUL 1. SIMILARITATEA TOPOLOGICĂ

Similaritatea structurilor moleculare exprimă existenţa unor trăsături comune într-un set de molecule. Similaritatea este definită pe baza unor variate criterii şi/sau proceduri şi ea generează clase de echivalenţă intermoleculară, în cadrul setului de molecule.

Similaritatea moleculară, la fel ca ramificarea, este o noţiune intuitivă, astfel că nu se poate defini o măsură unică şi non-ambiguă de similaritate. Multe dintre ariile ştiinţei cum sunt: chimia organică sintetică, biologia structurală, farmacologia şi toxicologia sunt necesare în procesul de dezvoltare a medicamentelor. Costurile foarte înalte şi timpul îndelungat, caracteristice acestui proces, accentuează necesitatea pentru investiţii în tehnologii care să accelereze procesul de proiectare a noilor structuri chimice, cu efectele biologice dorite, şi să scurteze timpul până când acestea sunt puse pe piaţă (acelaşi lucru este valabil şi pentru pesticide, fungicide, etc., chimicale pentru agricultură). Aceste investiţii au dus la dezvoltarea sistemelor foarte sofisticate de stocare, căutare şi procesare a unei variate game de informaţii chimice.

Descrierea moleculară, utilizată în analiza similarităţii moleculare, se realizează cu ajutorul descriptorilor moleculari. Orice descriere moleculară induce o partiţionare în clase de echivalenţă a setului de molecule. Este necesar să definim aici relaţia de echivalenţă:
Fie *S* un set de structuri moleculare şi *R* o relaţie binară pe *S* legând perechi de molecule. Dacă $x, y \in S$ sunt astfel legate, atunci se poate scrie *xRy*. Relaţia *R* este o relaţie de echivalenţă dacă satisface următoarele proprietăţi:

xRx , pentru oricare $x \in S$	(reflexivitate)	(1.1)
dacă *xRy*, atunci *yRx*	(simetrie)	(1.2)
dacă *xRy* şi *yRz*, atunci *xRz*	(tranzitivitate)	(1.3)

Subsetul elementelor y ∈ S, aflate în relaţia *xRy*, reprezintă clasa de echivalenţă a lui *x*. Impunând relaţia de echivalenţă *R* pe setul *S* rezultă o partiţionare a *S* în subseturi disjuncte, numite clase de echivalenţă sub *R*. Un astfel de subset, notat *S/R*, se mai scrie şi *S modulo R*.[1]

Fie *f* o funcţie *mapping* (de acoperire, potrivire, asociere, etc.) a elementelor setului *S* peste elementele unui set oarecare *Y*. Adică, pentru oricare $x \in S$, *f* atribuie o valoare corespunzătoare $y = f(x)$ în *Y*. Această corespondenţă poate fi scrisă ca $f: S \rightarrow Y$. Dacă Y este setul descrierilor, funcţia de suprapunere asociază o descriere moleculară fiecărei molecule în S. Acele molecule din S sunt echivalente care au aceeaşi descriere moleculară. O astfel de funcţie *f* poate fi o numerotare, un cod, ori un simplu proces de măsurare. Se poate demonstra că diferitele descrieri moleculare împreună cu reprezentarea lor algebrică formează un grup.[1,2]

O potrivire (*matching*) se poate realiză prin suprapunerea a două molecule. O astfel de operaţie poate indica aspecte comune ale celor două molecule sau ale descrierilor acestora.

O *ordonare parţială* se referă la o anumită ordonare locală, indusă de o acoperire (potrivire) parţială (i.e. substructure matching) între moleculele unui set. Matematic, *relaţia de ordonare* implică proprietatea de antisimetrie (1.4)
dacă *xRy* şi *yRx* atunci x = y, (1.4)
în locul celei de simetrie (1.2) (vezi mai sus). Randić[3] a raportat o ordonare parţială a isomerilor alcanilor indusă de numărul căilor (path numbers) p_2 şi p_3. Oricare alţi descriptori graf-teoretici (e.g., indici topologici, secvenţe ale distanţelor, etc.) pot fi utilizaţi în scopul caracterizării şi ordonării parţiale ca şi al grupării (*clustering*) structurilor moleculare.
Un întreg volum al revistei *MATCH* (*Communications în Mathematical and în Computer Chemistry*, **2000**, *42*) este dedicat ordonării parţiale în chimie.

Compuşii poziţionaţi mai aproape, în secvenţă, este de aşteptat să prezinte valori mai apropiate ale unor proprietaţi (i.e., proprietăţi similare).
Proximitatea exprimă în esenţă două categorii: *similaritatea* şi *disimilaritatea*.
Similaritatea exprimă asemănarea a două molecule, printr-un număr mare, dacă descrierile lor moleculare sunt apropiate şi respectiv printr-un număr tinzând spre zero în cazul în care aceste descrieri nu au nimic comun.[4] De exemplu, la suprapunerea a două molecule, raportul numărului de atomi şi legături care se suprapun şi numărul total al acestora în molecula întreagă, multiplicat cu acelaşi raport pentru molecula cu care se compară[5] a fost propus ca măsură a similarităţii între două molecule. O astfel de măsură are proprietatea corelaţiei (*zero* pentru non-corelaţie şi *unu* pentru corelaţie deplină).

Disimilaritatea exprimă asemănarea a două molecule, cu un număr apropiat de zero când descrierile lor moleculare sunt apropiate şi respectiv printr-un număr mare în cazul în care aceste descrieri sunt diferite.[4] În exemplul de supapunere de mai sus, numărul atomilor şi legăturilor care nu se suprapun poate fi luat ca măsură a disimilarităţii între două molecule. Acest caz particular de disimilaritate[6] este cunoscut în literatură ca *distanţă chimică*.[7,8, 9, 10]
Cercetări recente efectuate în domeniul proiectării substanţelor bioactive ("*drug design*"), în special medicamente dar şi insecticide, erbicide, etc., au relevat importanţa deosebită a *similarităţii*[1] compuşilor implicaţi, pe care-i vom denumi în continuare *efectori* (E), pentru interacţia lor cu un anumit *receptor biologic* (R), oricare ar fi natura lui. Deoarece în domeniul *drug design* se urmăreşte obţinerea unei relaţii cantitative între structura chimică a efectorilor E şi activitatea lor biologică - măsurată de obicei ca log(1/C), unde C reprezintă concentraţia molară (doza) care determină un răspuns biologic constant a devenit cuantificarea similarităţii, respectiv disimilarităţii compuşilor chimici. Aceasta presupune găsirea unor descriptori cantitativi adecvaţi care să poată fi utilizaţi în relaţii liniare (sau ne-liniare) structură chimică - activitate biologică (**QSAR** = **Q**uantitative **S**tructure - **A**ctivity **R**elationship) şi/sau pentru proiectarea seriilor QSAR.
Deşi ideea intuitivă de similaritate, aşa cum a fost ea utilizată de chimiştii organicieni - fiind "măsurată" prin numărul de caracteristici structurale comune pentru doi compuşi şi prin aranjamentul lor reciproc - pare simplă, în realitate există mai multe abordări datorate multiplelor reprezentări posibile ale compuşilor chimici organici (E): prin formula moleculară, topologic - cu ajutorul grafurilor moleculare constituţionale plane (2D), sau, mai exact, prin introducerea unei metrici euclidiene peste structurile moleculare reprezentate în spaţiul cartezian (3D) - formulele configuraţionale precum şi cele conformaţionale, sau pe baza formei moleculelor aşa cum este ea descrisă prin suprafeţele van der Waals (vdW) sau cu ajutorul unor câmpuri de forţe moleculare.
În consecinţă, noţiunea de similaritate, deşi intuitiv este simplă, depinde în mod esenţial de punctul de vedere din care este abordată. Din aceste motive, au fost propuse pentru fiecare din modalităţile de descriere a structurii moleculare prezentate mai sus *măsuri diferite de similaritate*, capabile să evalueze cantitativ această noţiune vagă (în sens matematic). Trebuie subliniat faptul că aceşti *indicatori cantitativi de similaritate* au fost utilizaţi, aproape în totalitate în domeniul proiectării substanţelor bioactive. În afara similarităţii, dar în strânsă legătură cu ea, există alte două noţiuni, totuşi distincte - *disimilaritatea* şi *complementaritatea* - de asemenea frecvent utilizate în QSAR. Fiecare din aceste noţiuni trebuie definită în mod univoc din punct de vedere matematic pentru a evita orice ambiguitate. Noţiunea de complementaritate a fost propusă la sfârşitul secolului trecut de către E. Fischer, care a pus bazele chimiei proteinelor şi carbohidraţilor, pentru a explica specificitatea acţiunii biologice a unor molecule E, prin teoria sa "*cheie în broască*". De atunci noţiunea de complementaritate, ca bază a interacţiunilor R-E, a fost continuu dezvoltată şi rafinată, pe baza ei fiind propuse o serie de metode QSAR (MSA[2], MSD[3], MTD[3], etc.) şi de modelare moleculară (CoMFA[4], etc).

Cele trei concepte menţionate mai sus - *similaritatea, disimilaritatea, şi complementaritatea* - ar putea fi încadrate pe o scară similară celei de corelare liniară simplă în care un coeficient de corelare r = +1 indică o corelare liniară perfectă y=f(x), toate punctele fiind dispuse strict pe o dreaptă cu panta pozitivă (similaritate totală între moleculele Y descrisă prin y_i, i=1,N şi X caracterizată prin x_i, i=1,N, x_i şi y_i fiind descriptori structurali), r = -1 corespunde unei corelări perfecte inverse (Y şi X sunt perfect complementare), dreapta având panta negativă, iar r =0 semnifică absenţa totală a oricărei corelări (Y şi X sunt în totalitate disimilare). Luând în considerare scara de mai sus, apare ca foarte probabil ca două molecule să fie disimilare în ceea ce priveşte, de exemplu, repartiţia atomilor în spaţiul 3D sau, topologic, în 2D, având un coeficient de corelare $r \to 0$ definit în cadrul unei metrici adecvate şi să prezinte o similaritate avansată, cu r apropiat de 1 ($r \to 1$), din punctul de vedere al formei, volumului sau suprafeţei, etc.

Similaritatea structurilor moleculare 2D şi 3D

Graful molecular este o reprezentare (în general 2D, dar acest fapt nu diminuează generalitatea metodei) a modului de legare a atomilor între ei (conectivitatea moleculară). Atomii unui sistem molecular dat formează nodurile (vârfurile) grafului molecular respectiv iar legăturile chimice dintre ei sunt asimilate muchiilor. Acest graf molecular poate fi descompus, succesiv, în fragmente din ce în ce mai mici, adică în subgrafuri (elemente de teorie legate de această problemă se găsesc în capitolul 1 al primei părţi). Odată definite subgrafurile de un anumit tip, compararea a două structuri moleculare pentru a stabili gradul de similaritate sau de disimilaritate poate fi realizată prin găsirea subgrafului comun de mărime maximă (**MCS**) sau prin luarea în considerare a tipului atomilor şi a conectivităţilor lor. Acest ultim mod de cuantificare poate asigura valorilor obţinute o mai mare semnificaţie fizică, fapt neglijat de multe ori în studiile QSAR.
Metoda **MCS** poate fi extinsă la spaţiul 3D prin luarea în considerare a matricei de distanţă. Astfel, dacă fiecărui fragment considerat i se asociază matricea de distanţe în limitele de încredere adecvate, a fost definită următoarea măsură de similaritate, $\mathbf{r}_{AB}$, între moleculele A şi B:

$$r_{AB} = \frac{MCS_{A,B}}{N_A + N_B - MCS_{A,B}} \qquad (1)$$

În relaţia (1) N_A şi N_B reprezintă numărul de atomi din, respectiv, moleculele A şi B. Principala utilizare a acestui indicator cantitativ este legată de procesul de căutare în bazele mari de date care conţin molecule bioactive, pentru a extrage moleculele similare din punct de vedere structural.

Concluzii privitoare la modalităţile de descriere a similarităţii

Găsirea unei metode de evaluare cantitativă a similarităţii este doar una dintre problemele care trebuie rezolvate de către *"proiectantul"* de substanţe bioactive. Odată stabilit modul de calcul este necesar un algoritm care să permită compararea celor două molecule, A şi B, a căror similaritate sau disimilaritate urmează a fi evaluată cantitativ. De obicei, una dintre molecule este menţinută fixă iar cealaltă este rotită pentru a suprapune anumite caracteristici moleculare cu scopul de a maximiza similaritatea structurilor A şi B.
O problemă dificilă este construirea unui algoritm de suprapunere. Una dintre metode se bazează pe rotaţia unei molecule, B, în jurul axelor Euler ale celeilalte molecule, A. Metoda a fost dezvoltată de Oxford Molecular Ltd. În cadrul programului "*ANACONDA*". Şi în acest

caz, metoda nu dă rezultate bune dacă formele structurilor moleculare sunt foarte disimilare. Principalele dificultăţi sunt legate de modul în care este fixat centrul de rotaţie al moleculei B. Astfel, centrul de inerţie este adecvat pentru o moleculă sferică dar nu şi pentru o moleculă cu structură elipsoidală. În plus, dacă molecula este flexibilă, dificultăţile cresc enorm deoarece centrul de inerţie trebuie continuu deplasat.
În concluzie, cea mai bună metodă de suprapunere a două molecule disimilare constă în dezvoltarea unor metode invariante la rotaţie şi translaţie. În acest sens, metodele care utilizează matricea distanţelor (topologice sau cu diferite metrici) sunt cele mai adecvate pentru a compara puncte distribuite în spaţiu deoarece, în acest caz, nu este necesar ca structurile moleculare care se compară să fie centrate identic şi, de asemenea, nu sunt necesare transformări ale punctelor în cursul procesului de superpoziţie.

CAPITOLUL 2. MATRICI TOPOLOGICE

Un graf molecular poate fi reprezentat prin: un număr, o secvenţă de numere, o matrice sau un polinom[11]. Aceste reprezentări se doresc a fi unice, pentru o structură dată. Randić consideră[12] că matricile topologice pot fi acceptate ca bază raţională pentru dezvoltarea de indici topologici, utili în studii corelaţionale ori de similaritate.

Matricea de adiacenţă

În 1874 Sylvester[3] a arătat că o moleculă organică adecvat numerotată, poate fi reprezentată printr-o matrice de adiacenţă, **A(G)**. Aceasta este o tabelă pătratică de dimensiuni **N∗N**, ale cărei elemente **$[A]_{ij}$** se definesc astfel:

$$[A]_{ij} = \begin{cases} 1 & \text{dacă} \quad i \neq j \quad \text{şi} \quad (i,j) \in E(G) \\ 0 & \text{dacă} \quad i = j \quad \text{sau} \quad (i,j) \notin E(G \end{cases} \tag{2}$$

iar matricea **A(G)**:

$$A(G) = \{[A]_{ij};\ \ i, j \in V(G)\} \tag{3}$$

A(G) caracterizează graful până la izomorfism, din ea putându-se reconstitui **G.** Matricea **A(G)** este simetrică faţă de diagonala principală, astfel că transpusa ei, **$A^T(G)$,** lasă matricea de adiacenţă neschimbată[13]:

$$\mathbf{A^T(G) = A(G)} \tag{4}$$

Matricea de conectivitate

Matricea **A** nu ia în considerare caracterul de multigraf (legătura multiplă). Pentru a indica tipul legăturii se utilizează matricea de conectivitate **C(G)**, definită prin relaţiile:

$$[C]_{ij} = \begin{cases} b_{ij} & \text{dacă} \quad i \neq j \quad şi \quad (i,j) \in E(G) \\ 0 & \text{dacă} \quad i = j \quad sau \quad (i,j) \notin E(G \end{cases} \tag{5}$$

unde **b_{ij}** reprezintă ordinul convenţional de legătură: 0; 1; 2; 3; 1.5 pentru nelegătură, legătură simplă, dublă, triplă şi respectiv aromatică. Matricea **C(G)** va fi:

$$C(G) = \{[C]_{ij};\ \ i, j \in V(G)\} \tag{6}$$

Matricea distanţelor

Matricea distantelor, **D**(G), este o tabelă pătratică de dimensiuni **N∗N**, ale cărei elemente, $\mathbf{[D]_{ij}}$, se definesc astfel:

$$[D]_{ij} = \begin{cases} numarul\ \ de\ \ arce\ \ pe\ \ drumul \\ cel\ \ mai\ \ scurt\ \ (i,j),\ \ dacă\ \ i \neq j \\ \\ 0 \quad dacă \quad i = j \end{cases} \tag{7}$$

iar matricea **D(G)** va fi:

$$\mathbf{D(G) = \{[D]_{ij};\ i,j \in V(G)\}} \tag{8}$$

CAPITOLUL 3. INDICI TOPOLOGICI

Un număr care reprezintă o structură chimică, în termeni graf-teoretici, se cheamă descriptor topologic. Fiind un invariant structural, el nu depinde de numerotarea atomilor ori de reprezentarea pictorială a grafului molecular. Cu toată pierderea considerabilă de informaţie prin "proiectarea" ca singur număr a structurii, astfel de invarianţi şi-au găsit largi aplicaţii în corelarea şi prezicerea a numeroase proprietăţi moleculare[14,15] şi de asemenea în testele de similaritate şi izomorfism[16,12]. Când un descriptor topologic se corelează cu o proprietate moleculară el poate fi numit indice molecular ori indice topologic (***IT***).

Indici bazaţi pe matricea de adiacenţă

Indicele RANDIC

Indicii care operează după catene (în particular arce) se denumesc indici de conectivitate. Indicele χ (de conectivitate, similar cu M_2) a fost introdus de Randic pentru caracterizarea legăturilor în graf[19]:

$$\chi = \sum_{(ij)\in E(G)} (k_i * k_j)^{-1/2} \tag{9}$$

Diudea[6] a definit indicele χ per vârf:

$$\chi_i = \sum_{j:(ij)\in E(G)} (k_i * k_j)^{-1/2} \quad \text{si} \quad \chi = \frac{1}{2}\sum_i \chi_i \tag{10}$$

Indici bazaţi pe matricea distanţelor

Indicele WIENER

Wiener[30] a definit numărul ***W*** ca "suma distanţelor (ca număr de legături carbon -carbon) dintre oricare doi atomi de carbon în moleculă". În structuri aciclice, autorul a calculat ***W*** ca sumă a contribuţiilor "per legătură" ("bond contributions", care se corelează cu proprietăţile termodinamice ale hidrocarburilor aciclice):

$$W = W(G) = \sum_e W_e = \sum_e N_{L,e} * N_{R,e} \tag{11}$$

unde:

$$\mathbf{N_{L,e} + N_{R,e} = N(G)} \tag{12}$$

$\mathbf{N_L}$, şi $\mathbf{N_R}$ fiind numărul de vârfuri la stânga şi la dreapta arcului **e**, însumarea făcându-se după toate vârfurile în **G**.

Indici de similaritate moleculară

Structura chimică a fiecărei molecule A este decodată într-un set de *n* descriptori structurali (SD) colectaţi într-un vector de tip X=X(A),

$$X(A) = \{SD_1, SD_2, SD_3, ...SD_n\} \quad (13)$$

Pentru un set M de molecule, M={A, B, C, ...} toţi descriptorii structurali sunt colectaţi într-o matrice de dimensiuni *mxn* unde fiecare rând corespunde unei molecule şi fiecare coloană corespunde unui descriptor structural particular. Pentru calcularea indicilor de similaritate, descriptorii structurali pot fi standardizaţi prin metoda Z_Score (autoscaling) care dă valorile ce au media zero şi sunt scalate la varianţă unitate.

Indicele de similaritate Cosine

Coeficientul Cosine, C_S pentru similaritatea între două molecule A şi B este dat de relaţia:

$$C_S(A,B) = \frac{\sum_{i=1}^{n} X(A)_i X(B)_i}{\left[\sum_{i=1}^{n} X(A)_i^{\,2} \sum_{i=1}^{n} X(B)_i^{\,2}\right]^{1/2}} \quad (14)$$

cu proprietatea ca C_S este cuprins în intervalul [-1,1]. Carbo[17] a utilizat o formă a indicelui de similaritate Cosine definit pe integrala de densitate electronică pe tot spaţiul.

Indicele de similaritate Dice

Coeficientul Dice, D_S pentru similaritatea dintre doi vectori a doi descriptori structurali X(A) şi X(B) este dat de:

$$D_S(A,B) = \frac{2\sum_{i=1}^{n} X(A)_i X(B)_i}{\sum_{i=1}^{n} X(A)_i^2 + \sum_{i=1}^{n} X(B)_i^2} \quad (15)$$

cu proprietatea că $-1 \le D_S \le 1$.

Indicele de similaritate Richards

Indicele de similaritate Richards este definit prin ecuaţia:

$$R_S(A,B) = \frac{1}{n}\sum_{i=1}^{n}\left(1 - \frac{|X(A)_i - X(B)_i|}{\max(|X(A)_i|, |X(B)_i|)}\right) \quad (16)$$

CAPITOLUL 4. RELAŢII CANTITATIVE STRUCTURĂ– ACTIVITATE BIOLOGICĂ

Metodele moderne utilizate în scopul "proiectării" unor structuri moleculare cu activitate biologică specificată, de exemplu medicamente, insecticide, ierbicide şi fungicide, se bazează pe

cuantificarea bioactivităţii ca o funcţie de structura moleculară[18]. Acest mod de abordare îşi are originea în lucrările lui Meyer şi Overton[19] de la sfârşitul secolului trecut şi începutul secolului nostru. Astfel, ei au demonstrat cu succes pentru prima oară o dependenţă a bioactivităţii de un parametru fizico-chimic, coeficientul de partiţie, care este o funcţie de structura moleculară.

Introducerea conceptului de situs al receptorului biologic a constituit un element vital, de o importanţă excepţională pentru acest domeniu al cercetării ştiinţifice; el a fost intuit de Langley,[20] dar fundamentat şi dezvoltat de către Ehrlich,[21] părintele chimioterapiei. Conform acestui model, activitatea biologică depinde de recunoaşterea substratului bioactiv (efector) de către situsul receptorului; această etapă este urmată de legarea efectorului în situs.

Descoperirea dependenţei dintre bioactivitate şi configuraţie[22] a condus la recunoaşterea faptului că efectele sterice, indiferent care este tipul sau natura lor, joacă un rol esenţial în cadrul interacţiunilor receptor-efector, condiţionând şi modelând potenţa biologică a efectorului.

Modele avansate în QSAR

Un compus bioactiv, introdus într-un organism viu, induce un răspuns biologic, o reacţie specifică din partea organismului. Răspunsul este condiţionat de structura şi identitatea chimică a compusului bioactiv.

Interacţia compusului bioactiv cu organismul se face la nivelul molecular, în aşa-numiţii receptori biologici. Aceştia sunt situs-uri active localizate în macromolecule proteice, în interiorul celulei vii sau pe membranele celulare.

Receptorii biologici[23,24,25,26] au următoarele caracteristici :

(i) specificitate : receptorii recunosc enantiomerii activi ("eutomeri" spre deosebire de "distomeri" care nu sunt activi biologic) sau diastereoizomerii agoniştilor sau antagoniştilor;

(ii) saturabilitate : numărul situs-urilor active într-o formaţiune celulară este finit;

(iii) în general, receptorii se află în celula care generează răspunsul biologic.

E. FISCHER (1894) a formulat prima teorie a interacţiei compusului bioactiv (efector) cu receptorul. Conform acesteia, receptorul este privit ca o cavitate rigidă, în care efectorul trebuie să se "potrivească" asemănător "cheii în broască" ("kee in lock"). Ulterior s-a admis că receptorii sunt semi-rigizi, ei căutând optimizarea mutuală (deşi limitată) cu efectorii săi. Ambii parteneri ai complexului efector-receptor se abat de la starea conformaţională de minim energetic, pentru a realiza complexul cu cea mai mare stabilitate (ex. care induce răspunsul biologic). Despre alosterism şi alte aspecte legate de interacţiunea efector-receptor vezi[148,149,150,151].

Răspunsul biologic provocat de un compus chimic pătruns în organism este proporţional cu concentraţia complexului **E...R** (efector-receptor), rezultat în reacţia:

$$E + R \underset{}{\overset{K}{\longleftrightarrow}} E...R \xrightarrow{k} P \qquad (17)$$

Cât timp complexul **E...R** există, se manifestă o reacţie specifică a organismului, numită răspuns biologic. Complexul poate disocia în componente (echilibrul caracterizat prin constanta **K**) sau poate forma (cu viteza caracterizată prin constanta **k**) produsul **P**. Concentraţia acestuia variază în timp cumform relaţiei:

$$\frac{d[P]}{dt} = k[E...R] \qquad (18)$$

concentraţia complexului putând fi aproximată din relaţia de echilibru:

$$[E...R] = [E][R]K = [E][R]\exp(-\Delta G / RT) \qquad (19)$$

CAPITOLUL 5. ANALIZA DE DATE

Coeficientul de corelaţie. Coeficientul de corelaţie teoretic ρ_{xy} a două variabile aleatoare **x** şi **y** este covarianţa variabilelor normate corespunzătoare:

$$\rho_{xy} = \mathrm{cov}\left(\frac{x-\bar{x}}{\sigma_x}, \frac{y-\bar{y}}{\sigma_y}\right) = \mathrm{Med}\left[\left(\frac{x-\bar{x}}{\sigma_x}\right)\left(\frac{y-\bar{y}}{\sigma_y}\right)\right] = \frac{Med\left[(x-\bar{x})(y-\bar{y})\right]}{\sigma_x\sigma_y} = \frac{\mathrm{cov}(x,y)}{\sigma_x\sigma_y} \qquad (20)$$

Valorile lui fiind cuprinse în domeniul: $-1 \le \rho_{xy} \le 1$. În cazul unei selecţii, relaţia (20) se va scrie:

$$\rho_{xy} = \frac{\mathrm{cov}(x,y)}{S_x S_y} = r_{xy} \qquad (21)$$

r_{xy} fiind coeficientul de corelaţie empiric. Când variabilele x şi y sunt independente, cov(x,y) = 0 şi $r_{xy} = 0$. *Reciproca nu este adevărată*. Subliniem că ρ (coeficientul de corelaţie teoretic) se referă la întreaga populaţie iar r (coeficientul de corelaţie empiric) se referă la o selecţie. În cazul în care între x_i şi y_i este definită o *funcţie de regresie* $\mathbf{y(x_i) = f(x_i)}$, se calculează un *coeficient de corelaţie (regresie) parţial,* r_{xiy} *, al variabilei* x_i .

Coeficientul de corelaţie (regresie) global, R^2, este dat de raportul dispersiei valorilor calculate de la medie cu dispersia valorilor empirice de la medie:

$$R^2 = 1 - \frac{\sum_i (y_i - y_{i,calc})^2}{\sum_i (y_i - \bar{y})^2} = \frac{\sum_i (y_{i,calc} - \bar{y})^2}{\sum_i (y_i - \bar{y})^2} \qquad (22)$$

R^2 ia valori în domeniul [0, 1]. Valori apropiate de 1 arată că dependenţa liniară este potrivită pentru descrierea relaţiei dintre y şi x.

Regresii liniare

Fie setul de structuri chimice $C_1, C_2, \ldots, C_n$ şi valorile observate ale unei proprietăţi moleculare $y_1, y_2, \ldots, y_n$ (y_i corespunde compusului C_i). Estimarea valorilor y_i cu ajutorul variabilelor independente x_{ij} (numite şi variabile predictor sau explicative) se poate face conform relaţiei:

$$y_{i,calc} = b_0 + \sum_{j=1}^{m} b_j x_{ij} \qquad (23)$$

Variabilele x_{ij} codifică numeric caracteristica structurală (topologică) sau fizico-chimică j - prezentă în compusul C_i.

Statistica Fischer F, indică nivelul de semnificaţie al ecuaţiei de regresie. Estimatorul F se calculează cu relaţia:

$$F = \frac{SS_{reg}}{SS_e/(n-k)} \qquad (24)$$

unde SS_{reg} este suma pătratelor erorilor atribuite regresiei:

$$SS_{reg} = \sum_i (y_{i,calc} - \bar{y})^2 \qquad (25)$$

Statistica t - Student. Estimatorul t indică nivelul de semnificaţie al coeficienţilor b_j; el se calculează cu formula:

$$\mathbf{t_j = \frac{|b_j|}{\sigma_{bj}}} \qquad (26)$$

unde σ_{bj} este eroarea standard a coeficientului de regresie b_j. Din compararea valorii t_j calculate cu valori tabelate (pentru un prag de semnificaţie impus, care este funcţie de gradele de libertate ale regresiei - vezi şi estimatorul F) se validează sau nu aportul unei variabile x_{ij} la corelaţia globală.[27,28].

Analiza clusterilor

Obiectivele analizelor de clusteri

Noţiunea **analiză de clusteri (CA)** se încadrează într-o familie de metode care este folosită în principal pentru găsirea şi scoaterea în evidenţă a structurilor din interiorul datelor. Din acest punct de vedere este mai degrabă o optimizare a datelor decât o manipulare a lor. În această direcţie analiza clusterilor poate fi văzută ca şi o metodă numită **model de cunoaştere.** Este folosită pentru aceasta o denumire sinonimă cu numere taxonomice şi clasificare automată. Dacă scopul analizelor de date sau întrebările ce se pun despre ele sunt fixate clar şi duc la obţinerea unor rezultate, următoarea întrebare va fi: care este cauza structurii găsite?

Să presupunem că avem un număr rezonabil de obiecte, *n*, şi proprietăţile lor, ce au fost selectate şi aranjate în matricea X ca şi în exemplul de mai jos:

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{pmatrix} \qquad (27)$$

şi să începem extragerea de informaţii din această matrice.

Unităţi asemănătoare şi preprocesarea informaţiilor

În ordinea de găsire a structurilor în grupul de date sau descoperirea similarităţilor probelor, organismelor,…care mai jos vor fi denumite **obiecte**, mai înainte de toate avem nevoie de **unităţi asemănătoare**. Cele mai simple unităţi asemănătoare pot deriva din geometrie. Fără a demonstra că conceptele intuitive de similaritate şi distanţă sunt complementare în natură şi amintindu-ne legea lui PITAGORA, distanţa *d* dintre două puncte O_1şi O_2 într-un sistem rectangular cu două axe *x* şi *y* este:

$$d(O_1, O_2) = \sqrt{(y_1 - y_2)^2 + (x_1 - x_2)^2} \qquad (28)$$

Aceasta este prezentată în figura 5.1.

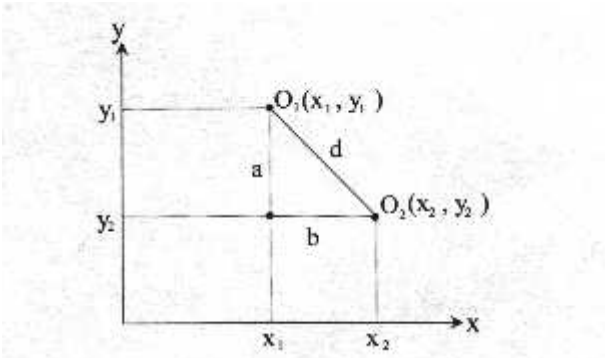


Figura 1. Distanţa dintre două obiecte în spaţiul plan (legea lui Pitagora)

Întinderea acestei legi peste mai mult de două dimensiuni a "spaţiului PITAGORA" conduce la distanţa euclidiană a oricăror două obiecte O_i şi O_k care mai jos vor fi simplu scrise ca şi d(I,k):

$$d(i,k) = \sqrt{\sum_{j=1}^{m} \left(x_{ij} x_{kj}\right)^2} \qquad (29)$$

unde: *m*-numărul trăsăturilor.

Analiza factorială

Analiza factorială a apărut ca o metodă de reducere a efectivului de variabile propuse pentru descrierea unui domeniu, prin construcţia unor noi variabile (numite *factori*), în număr mult mai mic, şi prin determinarea unor relaţii matematice care să precizeze legătura dintre variabilele iniţiale şi factori, astfel încât prin aceste noi variabile să se reproducă, în cea mai mare măsură, informaţia dată de variabilele iniţiale. Metoda datează încă de la începutul secolului trecut şi este datorată lui K. Pearson (1901) şi C. Spearman (1904).
Tehnicile statistice cunoscute sub această denumire au ca obiectiv comun reducerea numărului de variabile ce caracterizează o mulţime de obiecte la un număr mai mic de variabile, de obicei diferite de cele iniţiale.

Ecuaţia de bază a analizei factoriale
Din matricea de variabilelor standardizate poate fi dedusă matricea corelaţiilor dintre variabilele iniţiale:

$$R = \frac{1}{n} Z \cdot Z^T, \; r_{ij} = \frac{1}{n}\left(z_{i1} z_{j1} + z_{i2} z_{j2} + \ldots + z_{in} z_{jn}\right) \qquad (30)$$

Numărul factorilor de extras

O modalitate empirică a modului suficient de descompunere cu k factori a fost definită de Malinowki (1977) prin funcţia IND. Valoarea minimă IND indică numărul probabil al factorilor relevanţi.

$$IND = \frac{RE}{\left(m-k\right)^2} \qquad (31)$$

$$RE^2 = \frac{1}{n\left(m-k\right)} \sum_{j=k+1}^{m} \lambda_i \qquad (32)$$

unde m – numărul variabilelor
n – numărul obiectelor
k – numărul extras de factori
RE – „eroarea reală"

CONTRIBUTII PERSONALE

Tehnicile relaţiilor cantitative structură-activitate (QSAR) devin indispensabile în toate aspectele cercetării privind interpretarea moleculară a proprietăţilor biologice.[29] Este de înţeles ca proprietăţile fizice, chimice sau biologice ale compuşilor depind de aranjamentul 3D (tri-dimensional) al atomilor din moleculă. Abilitatea de a produce corelaţii cantitative intre structura moleculară 3D şi activitatea biologica este importantă în alegerea căilor de sinteza a substanţelor biologic active.[30]

Activitatea biologică a steroizilor variază considerabil cu modificarea foarte uşoară a structurii. Aceste familii importante de molecule prezintă caracteristici foarte schimbătoare

pentru orice metodă de predicţie, datorită flexibilităţii relative scăzute a scheletului de cyclopentanoperhidrofenantrenă. Din acest motiv, foarte multe modele QSAR bazate pe proprietăţile 2D, cum ar fi descriptorii topologici, au o calitate comparabilă cu modelele provenite de la metodele complexe 3D.[31,32]

În acesta parte sunt identificate si prezentate aspectele legate de structura moleculare care sunt relevante în particular pentru activitatea biologică (afinitatea de legare cu receptorul) pentru diferite clase de substante cu activitate biologica.

Compuşi steroidici

Setul de 31 de structuri steroidice (androstan) AS (Figura 2, Tabel 1), cu afinitatea de legare a corticosteroizilor de globulină (CBG) a fost luat din publicaţiile lui Dunn *et al.*[33] şi Tuppurainen, *et al.*[34] Acest set de structuri a fost de multe ori utilizat pentru evaluarea performanţelor noilor metode de analiză QSAR. Oricum, menţionând alţi autori putem spune că foarte multe publicaţii ce au utilizat acest set au inclus erori în structura steroizilor. Structurile utilizate în acest studiu au fost verificate foarte atent cu scopul de a evita o viitoare propagare de erori. Calitativ, moleculele cu substituenţi cum ar fi oxigen sau hidroxil în poziţia 17 pe structura steroidică, au activitate CBG crescută, în timp ce prezenţa unei catene voluminoase cum ar fi -$COCH_2OH$, conduce la scăderea acestei activităţi.

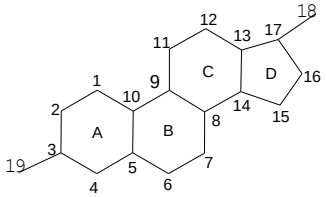


Figura 2. Structura Androstanului

Tabelul 1. Setul de structuri AS:

	Compuşi	Activitate		Compuşi	Activitate
1	aldosterone	6.279	17	pregnenolone	5.255
2	androstanediol	5.000	18	17-hydroxypregnenolone	5.000
3	androstenediol	5.000	19	progesterone	7.380
4	androstenedione	5.763	20	17-hydroxyprogesterone	7.740
5	androsterone	5.613	21	testosterone	6.724
6	corticosterone	7.881	22	prednisolone	7.512
7	cortisol	7.881	23	cortisol 21-acetate	7.553
8	cortisone	6.892	24	4-pregnene-3,11,20-trione	6.779
9	dehydroepiandrosterone	5.000	25	epicorticosterone	7.200
10	deoxycorticosterone	7.653	26	19-nortestosterone	6.144
11	deoxycortisol	7.881	27	16R,17-dihydroxy-4-pregnene-3,20-dione	6.247
12	dihydrotestosterone	5.919	28	16-methyl-4-pregnene-3,20-dione	7.120
13	estradiol	5.000	29	19-norprogesterone	6.817
14	estriol	5.000	30	11β,17,21-trihydroxy-2R-methyl-4-pregnene-3,20-dione	7.688
15	estrone	5.000	31	11β,17,21-trihydroxy-2R-methyl-9R-fluoro-4-pregnene-3,20-dione	5.797
16	etiocholanolone	5.255			

Subsetul parametrilor electronici include descriptori moleculari derivaţi din sarcinile atomice parţiale. Cu programul TOPOCLUJ, sarcinile parţiale Ch_i au fost calculate astfel:

$$Ch_{i,j} = \log(S_j / S_i)^{1/(d_{i,j})^2} \quad (33)$$

$$Ch_i = \sum_j ch_{i,j} \quad (34)$$

În ambele cazuri S_i, S_j reprezintă electronegativităţile Sanderson de grup calculate pentru grupuri hibride (de exemplu pentru atomi grei care sunt înconjuraţi de atomi de hidrogen) în molecule, în timp ce d_{ij} este distanţa Euclidiană ce separă atomii *i* şi *j* din structura chimică optimizată cu energie minimă (HyperChem).[137] $Ch_{i,j}$ este perturbaţia electronegativităţii atomului *i* produsă de orice atom *j* din moleculă în timp de Ch_i este rezultatul acestor perturbaţii pe atomul *i.* Pentru alte calcule topologice ale sarcinii parţiale se pot consulta referinţele.[35,36]

Orice compus steroidic poate fi descris prin aceste sarcini parţiale care caracterizează poziţiile substituite sau nesubstituite şi heteroatomii. Bazându-ne pe acest lucru am definit un nou descriptor global flexibil (***CD***) care poate fi definit ca o funcţie aditivă de autocorelare ponderată cu sarcinile parţiale corespunzătoare ale atomului *j* considerat:

$$CD = \sum_j c_j \cdot Ch_j \quad (35)$$

unde c_j este coeficientul de regresie dat de regresia multivariată $\mathbf{\log(A_i obs) = f(Ch_j)}$. Această "ad-hoc" ponderare depinde de setul de molecule luate în considerare şi de asemenea de descriptorii locali utilizaţi. Sarcinile parţiale (Ch_j) corespund următoarelor poziţii de pe structura de bază: 3, 10, 11, 13, 17, 18, 19 (Figura 2).

Softul *Dragon 2.1*[136] a fost folosit pentru a calcula 1600 de descriptori moleculari pentru compuşii studiaţi. Cei mai relevanţi dintre aceşti descriptori utilizaţi aici sunt cei cu funcţii de distribuţie radială (RDF), indici de autocorelare şi descriptori geometrici.

Descriptorii aparţinând clasei cu funcţie de distribuţie radială sunt bazaţi pe distribuţia distanţei în reprezentarea geometrică a moleculei. În plus faţă de distanţele interatomice în întreaga moleculă, RDS aduc informaţii despre distanţele legăturii, tipul inelului, sisteme planare sau neplanare, tipuri de atomi şi alte modificări structurale importante. Prin utilizarea diferitelor scheme de ponderare, care includ tipuri de atomi, electronegativităţi, masa atomică (*RDF090m*) sau distanţe van der Walls, RDF poate fi ajustat să dea evoluţii importante ale descriptorilor în studiile QSAR.

Al doilea grup de descriptori este obţinut aplicând funcţii de autocorelare bidimensionale pe graful molecular. Astfel de descriptori exprimă corelaţia dintre valorile numerice, care pot fi ponderate statistic utilizând proprietăţi atomice, la intervale egale de valoare dată.[37] De exemplu, *MATS1p*- auto-corelaţie Moran-lag 1/ponderată cu polarizabilităţile atomice; *MATS4e*- auto-corelaţie Moran-lag 4/ponderată cu electronegativităţile Sanderson atomice. Aplicaţiile electronegativităţilor Sanderson ca şi coeficienţi de ponderare, luate în acest context în unele cazuri schimbă distribuţia în interiorul moleculei.

Descriptorii geometrici indică mărimea moleculei, ei sunt derivaţi din coordonatele tridimensionale ale nucleelor atomice şi a maselor atomice şi / sau distanţa atomică din moleculă.

Prelucrarea datelor

Ţinând cont de complexitatea interacţiilor dintre molecula receptor şi moleculele cu potenţial inhibitor, este destul de dificil să modelezi setul de structuri folosind doar modele simple de regresie liniară. In vederea analizei datelor obtinute s-a propus urmatorul model:

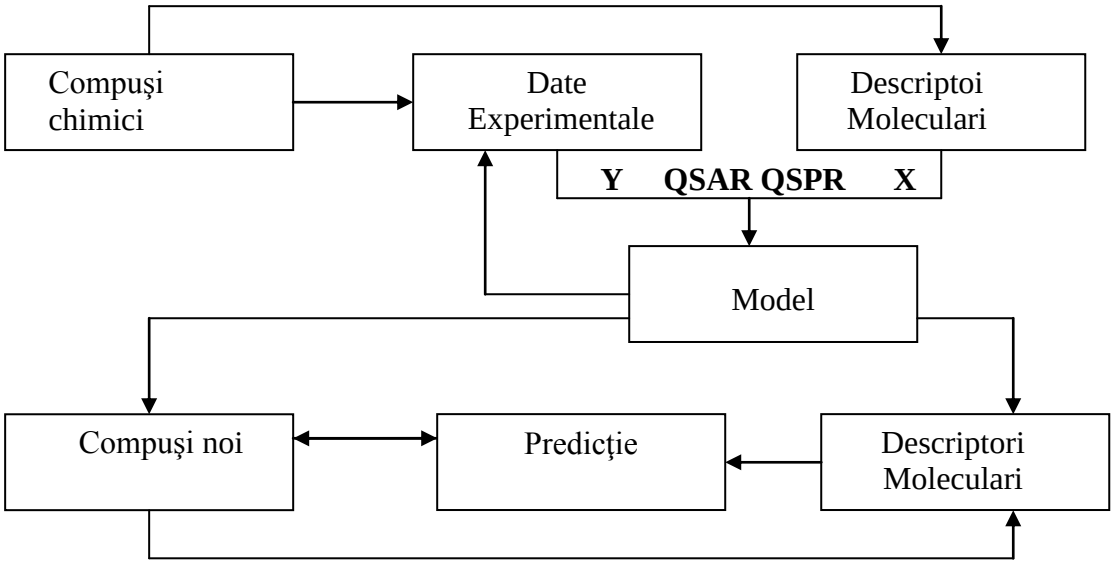


Figura 3. Reprezentarea schematică a construcţiei modelului.

Analiza QSAR constă în următorii paşi:

(i) optimizarea structurii folosind metoda semiempirică PM3;
(ii) calcularea descriptorilor moleculari;
(iii) regresie multivariabilă pentru găsirea coeficienţilor de autocorelare;
(iv) împărţirea setului de date în unul şcoală (calibrarea regresiei) şi unul de predicţie (pentru validarea modelului);
(v) testarea capacităţii de a prezice modelul;
(vi) analiza componentelor principale (PCA);
(vii) găsirea unei funcţii de regresie pentru model;
(viii) testarea capacităţii predictive a acestui model;
(ix) interpretarea modelului.

În ambele regresii monovariate şi bivariate (tabelul 2) compusul 13 apare ca fiind *outlier*. Acest compus nu a fost inclus în analizele viitoare. Cu acest *outlier* exclus, am observat o îmbunătăţire a corelaţiei cu descriptorii prezentaţi mai jos.

Tabelul 2. Modelele utilizate şi rezultatele procedurii de *cross-validare*

Setul de date	**Numărul obser-vaţiilor (*n*)**	**Modelul**	**Coeficientul de corelaţie (r^2) (înainte de LOO)**	**Structura Outlier**	**Coeficientul de corelaţie (r^2) (după LOO)**
AS	31	$\log(P_i\text{calc}) = f(CD)$	0.891	13	0.920
		$\log(P_i\text{calc}) = f(CD, L/Bw)$	0.931	13	0.939

Regresia trivariabilă a dat rezultate asemănătoare dar nu a adus îmbunătăţiri esenţiale. Rezultatele obţinute vor fi prezentate mai departe.

Setul AS de structuri

Am împărţit setul de structuri AS în doua seturi: setul scoală (n = 20) şi setul de predicţie (validare) (n = 11) aşa cum este arătat în Tabelul 3.

(a) Setul Scoală (n = 20)

Descriptorul electronic ***CD***, actualul ***CDP***, a fost calculat *de nuovo* pe setul şcoală ţinând cont de ecuaţia 6.7, aceasta datorită greutăţilor de corelare ale lui c_j potrivit doar cu

proprietatea selectată şi cu setul dat (in acest caz, setul şcoală având 20 de structuri). În Tabelul 3 sunt prezentaţi cei mai relevanţi descriptori pentru setul de structuri date. Cele mai bune modele pentru setul de structuri AS sunt prezentate mai jos.

Regresie monovariată

$\log P_i calc = 7.236 + 1.033 \cdot CDP_i$ (36)

$n = 20 \quad R^2 = 0.903 \quad s = 0.10 \quad F = 159.99$

Regresie bivariată

$\log P_i calc = 5.737 + 0.914 \cdot CDP_i + 0.194 \cdot L/Bw_i$ (37)

$n = 20 \quad R^2 = 0.931 \quad s = 0.31 \quad F = 114.12$

Regresie multiplă

$\log P_i calc = 6.268 + 0.17 \cdot L/Bw_i - 0.166 \cdot RDF090m_i + 0.904 \cdot CDP_i$ (38)

$n = 20 \quad R^2 = 0.962 \quad s = 0.24 \quad F = 129.73$

Tabelul 3. Descriptorii topologici şi coeficientul de partiţie observat log*P* pentru setul de structuri AS

Structura	L/Bw	RDF090m	CDP	log *P* obs.
Setul Şcoala				
1	7.3	3.17	-1.722	5
2	6.5	3.127	-2.101	5
4	6.5	3.505	-2.062	5
5	6.9	1.755	-2.456	5
7	6.3	2.803	-2.188	5
8	4.3	0.191	-1.950	5.255
9	7.1	1.91	-2.087	5.255
12	7	4.01	-0.880	5.797
13	6.1	3.497	-0.187	5.919
14	7.2	0.61	-1.212	6.144
17	6.9	2.729	-0.336	6.724
18	6	3.288	0.492	6.779
20	7.1	3.976	-0.243	6.892
22	6.6	2.525	-0.046	7.2
23	7.1	0.942	0.085	7.38
25	9.5	1.704	0.205	7.553
26	9.1	0.673	0.006	7.653
28	7.8	1.122	0.317	7.74
29	8.7	2.31	0.040	7.881
30	9	1.077	0.309	7.881
Setul de predicţie (validare)				
3	6.7	2.669	-2.350	5
6	6.6	0.811	-2.117	5
10	6.3	1.338	-2.184	5.613
11	6.2	1.272	-0.479	5.763
15	6.9	1.986	0.400	6.247
16	6.4	1.759	-0.716	6.279
19	9.1	0.406	-0.810	6.817
21	6.9	2.014	0.046	7.12
24	5.9	2.777	0.110	7.512
27	6.9	3.512	0.675	7.688
31	8.7	1.423	0.225	7.881

(b) Setul de predicţie (validare) (n = 11)

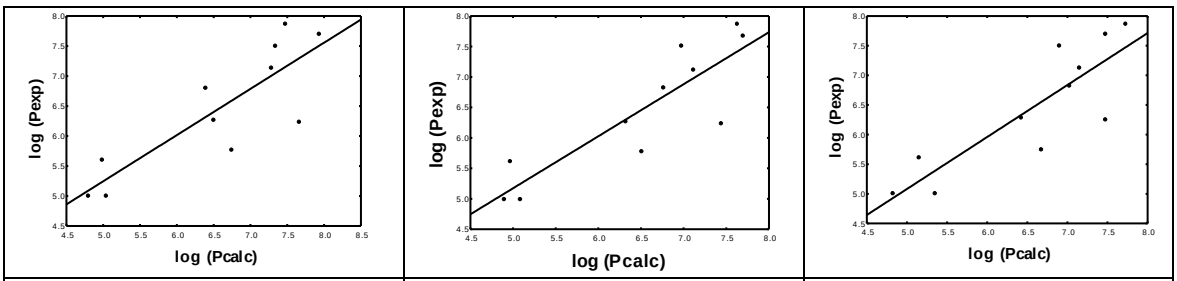
Orice model QSAR trebuie să fie validat cu un set de predicţie extern. Calcularea descriptorului *CDP* în setul de predicţie s-a realizat pe baza sarcinilor parţiale şi a parametrilor $c_j^{'}$ generaţi pentru setul şcoală (vezi tabelul 4.). Noi presupunem că activitatea biologică a setului de predicţie este necunoscută.

Tabelul 4. Setul de predicţie pentru structurile AS (n = 11)

Structuri	**log $P_i$obs**	**log $P_i$calc (ecuaţia 6)**	**log $P_i$calc (ecuaţia 7)**	**log $P_i$calc (ecuaţia 8)**
3	5.000	4.808	4.890	4.836
6	5.000	5.049	5.083	5.339
10	5.613	4.980	4.964	5.140
11	5.763	6.741	6.504	6.676
15	6.247	7.650	7.444	7.472
16	6.279	6.496	6.326	6.415
19	6.817	6.400	6.765	7.015
21	7.120	7.284	7.121	7.147
24	7.512	7.350	6.984	6.907
27	7.688	7.934	7.696	7.466
31	7.881	7.468	7.633	7.713
R^2		**0.716**	**0.766**	**0.728**
CV%		7.089	5.010	6.586

Pentru setul de structuri AS abilitatea de prezicere pare a fi mult mai bună în cazul regresiei bivariate (tabelul 4).

Figura 4. a-c prezintă valorile experimentale *vs* valorile calculate pentru afinitatea de legare a receptorului pentru structurile din AS : (a) valorile calculate conform ecuaţiei 36; (b) valorile calculate conform ecuaţiei 37, (c) valorile calculate conform ecuaţiei 38.



a	b	c

(a) log ($P_i$obs) = 1.3932 + 0.77043 × log ($P_i$calc); R = 0.84636;
(b) log ($P_i$obs) = 0.90106 + .85433 × log ($P_i$calc);R = 0.87565;
(c) log ($P_i$obs) = 0.69650 + 0.87704 × log ($P_i$calc); R = 0.85352

Figura 4. Reprezentarea grafică a valorilor experimentale versus valorile calculate ale afinităţii de legare a receptorului pentru setul de structuri AS

În scopul explicării contribuţiei fiecărui substituent al grafului din setul de structuri AS la afinitatea de legare a receptorului (CBG), am generat un descriptor electronic simplu în funcţie de sarcinile atomice parţiale, corelat parţial cu proprietatea studiată.

Modelele QSAR descrise în acest studiu indică faptul că acest descriptor global este unul dintre cei mai semnificativi în predicţia activităţii compuşilor noştri. El poate indica poziţiile celor mai importanţi substituenţi. Astfel, **CD** calculate pentru atomii din poziţiile mai sus

menţionate, fără substituenţi în poziţia 17 pe structura steroidică, duc la o varianţa de 25% a activităţii CBG în acest set de structuri **AS** în timp ce incluzând poziţia 17 creste la 89%.
Modelul a fost validat cu un set extern de predicţie. Modelul derivat pentru activitatea moleculară utilizând descriptorul **CD** şi descriptorii obţinuţi prin *factor loading* al **PCA**-ului, este comparabil cu celelalte modele descrise în literatura de specialitate, având o abilitate bună de predicţie. De notat faptul că modelul simplu 2D, la fel ca acesta pe care noi l-am dezvoltat, este comparabil cu rezultatele obţinute cu ajutorul modelelor complexe 3D (CoMFA, COMSA, GRIND, EEVA etc)[27,38,39,] care cer mult mai multe resurse computaţionale.

Compuşi cu acţiune antibacteriană

Setul de 38 de derivaţi de 2-furiletilenă, cu activitate antibacteriană a fost luat din publicaţiile lui Miguel Angel Cabrera Pérez[40] precum şi Yovani Marrero Ponce[41]. Derivaţii de 2-furiletilenă sunt substanţe biologic active cu spectru larg antimicrobial, antispasmic, citotoxic dar în anumite cazuri având activităţi carcinogenice şi mutagenetice (Yahagi *et al*, 1974; Dore şi Viel, 1975[42]; Miyaji, 1976; McCalla, 1979; McCalla, 1983; Kelloval *et al*, 1984; Estrada, 1998). Acest interes în studierea derivaţilor de 2-furiletilenă a crescut mult în ultimii ani ca şi consecinţă a descoperirilor a unor noi compuşi cu potenţial microcidal cu acestă structură chimică (McCoy and Thornburgh, 1992; Castañedo *et al*, 1994[43]; Blondeau *et al*, 1999[44]).
Modelul utilizat în acest caz este prezentat schematic în figura 5.

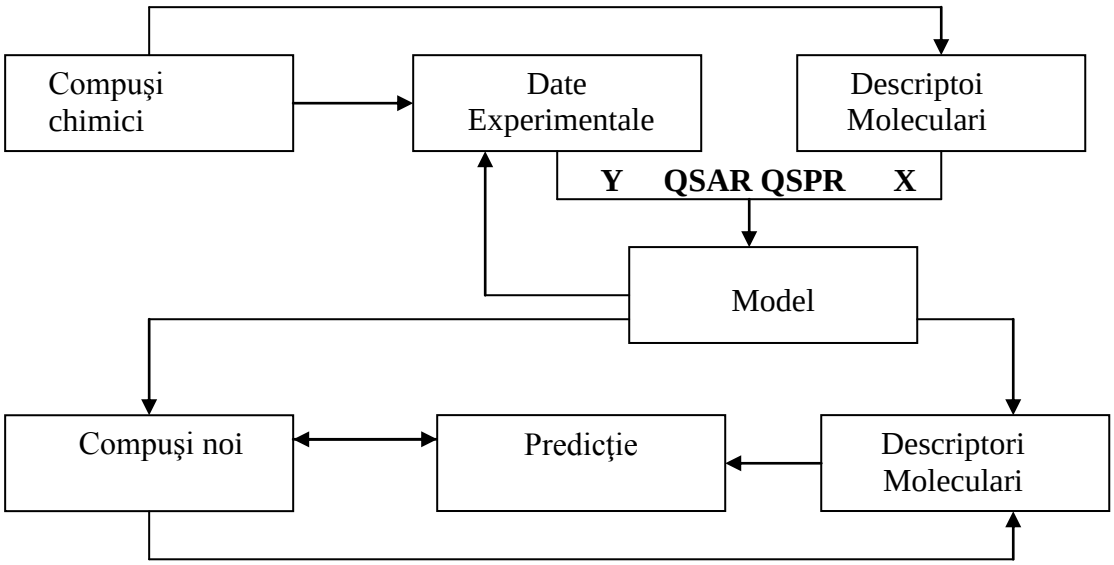
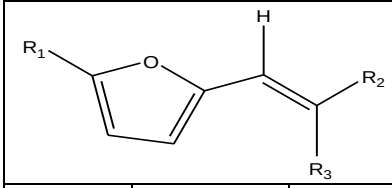


Figura 5. Reprezentarea schematică a construcţiei modelului.

Softul *Dragon 2.1* a fost folosit pentru a calcula pentru compuşii studiaţi a 1600 de descriptori moleculari. Ce mai relevanţi dintre aceşti descriptori utilizaţi aici sunt cei constituţionali (*MW* – masă moleculară), de conectivitate (*X4v* – indicele de conectivitate al valenţei *chi-4*) şi geometrici (G(N..O) – suma distanţelor geometrice între N şi O).
Descriptorii topologici calculaţi de programul TOPOCLUJ sunt derivaţi din matricele topologice sau polinoamele care descriu grafurile moleculare. Algoritmii de calcul includ câteva scheme de ponderare a grafurilor moleculare cu sarcini atomice parţiale, electronegativităţi şi mase ale fragmentelor moleculare. Matricele topologice calculate includ pe cele de bază: adiacenţă, distanţă, detour, conectivitate, Wiener, etc. precum şi matrici dezvoltate de grupul TOPO Cluj care includ: Cluj şi Cluj Fragmental, matricea W, LM, SM, operatorul matricial $W_{(M1,M2,M3)}$. Cei mai relevanţi dintre aceşti descriptori sunt :

- *IE[CfMax[Density]]*,
- *VAA1*,
- *VAD1*,
- *CS[LM[Electronegativity]]*,
- *CS[Sh[W4[Charge_Adjacency]]]*.

Tabelul 5. Setul de structuri a derivaţilor de 2-furiletilenă.



	R_1	**R_2**	**R_3**	**log P**
1	H	NO_2	$COOCH_3$	1.879
2	CH_3	NO_2	$COOCH_3$	2.439
3	Br	NO_2	$COOCH_3$	2.739
4	$COOCH_3$	NO_2	$COOCH_3$	1.869
5	NO_2	NO_2	$COOCH_3$	1.599
6	NO_2	$COOC_2H_5$	$COOC_2H_5$	2.504
7	NO_2	H	NO_2	1.303
8	H	H	NO_2	1.583
9	NO_2	H	$CONHC_2H_5$	1.386
10	NO_2	H	$CONH(CH_2)_2CH_3$	1.86
11	NO_2	H	$CONHCH(CH_3)_2$	1.803
12	NO_2	H	$CONH(CH_2)_3CH_3$	2.356
13	NO_2	H	$CONHCH_2CH(CH_3)_2$	2.225
14	NO_2	H	$CONHCH(CH_3)C_2H_5$	2.284
15	NO_2	H	$CONHC(CH_3)_3$	2.333
16	NO_2	H	$CONHCH_2C(CH_3)_3$	2.605
17	NO_2	H	$COOCH_3$	1.652
18	NO_2	H	$COOC_2H_5$	2.098
19	NO_2	H	$COO(CH_2)_2CH_3$	2.673
20	NO_2	H	$COOCH(CH_3)_2$	2.641
21	NO_2	H	$COO(CH_2)_3CH_3$	2.827
22	NO_2	H	$COOCH_2CH(CH_3)_2$	3.135
23	NO_2	H	$COOCH(CH_3)C_2H_5$	3.091
24	NO_2	H	$COOC(CH_3)_3$	3.06
25	NO_2	H	$COO(CH_2)_4CH_3$	3.404
26	NO_2	H	Br	2.447
27	NO_2	H	CN	1.05
28	NO_2	H	OCH_3	1.591
29	NO_2	H	H	1.611
30	NO_2	CN	$COOCH_3$	1.488
31	I	NO_2	$COOCH_3$	2.999
32	NO_2	H	$CONH_2$	0.649
33	NO_2	H	$CONHCH_3$	0.984
34	NO_2	H	$CON(CH3)_2$	0.819
35	Br	NO_2	Br	2.820
36	Br	NO_2	CH_3	2.730
37	H	NO_2	H	1.290
38	H	NO_2	CH_3	1.940

Analiza şi prelucrarea datelor

Coeficientul de partiţie *n*-octanol/apă (log P) joacă un rol important în înţelegerea comportamentului biologic al acestor derivaţi de 2-furiletilenă.

Ţinând cont de complexitatea interacţiilor dintre molecula receptor şi moleculele cu potenţial inhibitor, este destul de dificil să modelezi setul de structuri folosind doar simple modele de regresie liniară.

În pachetul software SIMIL sunt implementate proceduri de filtrare folosind secvenţele valenţelor vârfurilor utilizate la clusterarea sistemelor informatice chimice.

Analiza QSAR constă în următorii paşi:

(x) optimizarea structurii folosind metoda semiempirică PM3;

(xi) calcularea descriptorilor moleculari;

(xii) împărţirea setului de date în unul şcoală (calibrarea regresiei) şi unul de predicţie (pentru validarea modelului) pe baza similarităţii moleculare;

(xiii) analiza componentelor principale (PCA);

(xiv) testarea capacităţii de a prezice modelul;

(xv) găsirea unei funcţii de regresie pentru model;

(xvi) testarea capacităţii predictive a acestui model;

(xvii) interpretarea modelului.

Tabelul 6. Similariatea pentru setul de derivaţi de 2-furiletenă faţă de structurile lider alese în setul de predicţie şi coeficientul lor de partiţie n-octanol/apă (log P).

Nr. structurii	**35**	**36**	**37**	**38**	**log P**
1	0.59524	0.72024	0.71429	0.78571	1.879
2	0.55556	0.67222	0.66667	0.73333	2.439
3	0.67222	0.8	0.66667	0.73333	2.739
4	0.462963	0.56019	0.55556	0.61111	1.869
5	0.490196	0.59314	0.58824	0.64706	1.599
6	0.459375	0.48151	0.45125	0.50114	2.504
7	0.64103	0.64103	0.76923	0.6993	1.303
8	0.83333	0.83333	1	0.90909	1.583
9	0.50139	0.6125	0.60167	0.66818	1.386
10	0.47005	0.57422	0.56406	0.62642	1.86
11	0.47005	0.57422	0.56406	0.62642	1.803
12	0.490196	0.54044	0.53088	0.58957	2.356
13	0.490196	0.54044	0.53088	0.58957	2.225
14	0.490196	0.54044	0.53088	0.58957	2.284
15	0.490196	0.54044	0.53088	0.58957	2.333
16	0.462963	0.51042	0.50139	0.55682	2.605
17	0.5372	0.65625	0.64464	0.71591	1.652
18	0.50139	0.6125	0.60167	0.66818	2.098
19	0.47005	0.57422	0.56406	0.62642	2.673
20	0.47005	0.57422	0.56406	0.62642	2.641
21	0.490196	0.54044	0.53088	0.58957	2.827
22	0.490196	0.54044	0.53088	0.58957	3.135
23	0.490196	0.54044	0.53088	0.58957	3.091
24	0.490196	0.54044	0.53088	0.58957	3.06

25	0.462963	0.51042	0.50139	0.55682	3.404
26	0.83523	0.68371	0.82045	0.74587	2.447
27	0.62674	0.76563	0.75208	0.83523	1.05
28	0.62674	0.62674	0.75208	0.68371	1.591
29	0.75208	0.75208	0.9025	0.82045	1.611
30	0.47005	0.57422	0.56406	0.62642	1.488
31	0.55556	0.67222	0.66667	0.73333	2.999
32	0.57853	0.70673	0.69423	0.77098	0.649
33	0.5372	0.65625	0.64464	0.71591	0.984
34	0.50139	0.6125	0.60167	0.66818	0.819
35	1	0.84028	0.83333	0.75758	2.820
36	0.84028	1	0.83333	0.91667	2.730
37	0.83333	0.83333	1	0.90909	1.290
38	0.75758	0.91667	0.90909	1	1.940

Deoarece coeficientul de corelare este supus fluctuaţiilor de selecţie, o valoare r mare trebuie privită cu circumspecţie dacă numărul observaţiilor este mic şi, în plus, el nu poate fi utilizat ca termen de comparaţie pentru ecuaţii cu număr de date diferit.
Din tabelul 6.6 se observă similaritatea amprentelor moleculare pentru moleculele alese în setul de predicţie faţă de celelalte structuri din setul de bază. Putem observa că, cele patru structuri pot face parte dintr-un cluster deoarece similaritatea între ele este foarte ridicată. De asemenea am ales în setul şcoală acele structuri care au coeficientul de similaritate mai mare de 0.70 pentru cel puţin una dintre aceste structuri.

Setul şcoală şi setul de predicţie
Din setul şcoală fac parte structurile prezentate în tabelul 7, structuri cu coeficienţi de similaritate ridicaţi faţă de structurile lider (tabelul 8) din setul de predicţie, fiind prezentaţi aici şi cei mai semnificativi descriptori calculaţi cu programul Dragon alături de coeficientul de partiţie n-octanol/apă logP.

Tabelul 7. Descriptorii topologici şi coeficientul de partiţie observat log*P* pentru setul şcoală.

Nr. Structură	MW	X4v	G(N..O)	logP
1	197.16	0.799	9.619	1.879
2	211.19	0.956	9.613	2.439
3	276.05	1.15	9.626	2.739
8	139.12	0.52	4.386	1.583
17	197.16	0.748	14.97	1.652
26	218.01	0.853	0	2.447
27	164.13	0.649	22.162	1.05
28	169.15	0.665	6.598	1.591
29	139.12	0.55	0	1.611
31	323.05	1.265	9.641	2.999
32	182.15	0.694	29.088	0.649
33	196.18	0.771	29.063	0.984

Tabelul 8. Descriptorii topologici şi coeficientul de partiţie observat log*P* pentru setul de predicţie.

Nr. Structură	MW	X4v	G(N..O)	logP
35	296.9	1.121	4.405	2.49
36	232.04	0.984	4.516	2.37
37	139.12	0.52	4.386	1.56
38	153.15	0.632	4.517	1.92

Ecuţiile utilizate pentru predicţie sunt cele calibrate în clusterul şcoală (în forme lor normalizată de Matlab). Aceste ecuaţii prezintă un coeficient de corelaţie bun (R^2 = 0.9843 în cazul regresiei bivariate, ec. 39 şi R^2 = 0.98653 în cazul regresiei multiple, cu 3 descriptori, ec. 40). De precizat este faptul că în studiile anterioare la cel mai bun model obţinut în predicţia lui log P s-au folosit ecuaţii de regresie multiple cu 7 descriptori (ecuaţia 17 din publicaţia lui Yovani Marrero Ponce et all.) cu R^2 = 0.968 şi cu abiliatatea de prezicere R^2 = 0.938, valori mult mai mici decat rezultatele obţinute de noi.

Regresie bivariabilă

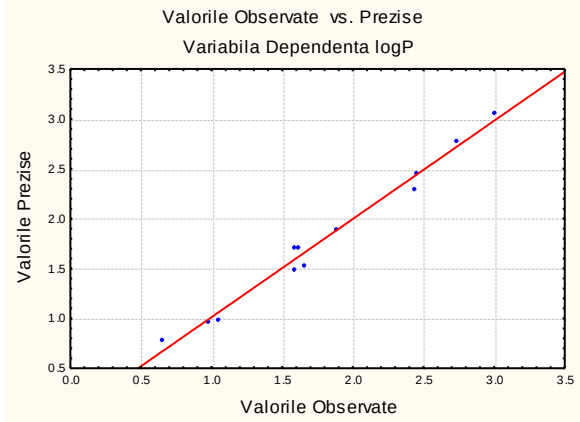
logP= 0.086403+3.442395· **X4v**+2.163165· **G(N..O)** (39)

n = 12 R^2 = 0.9843 s = 0.091837 F = 282.3465

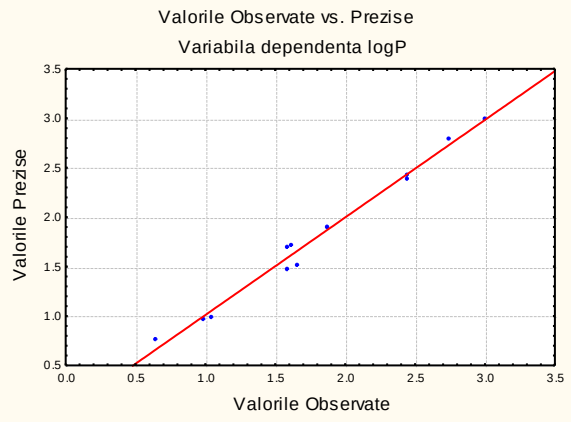
Regresie multi-variabilă

logP= 0.391261-0.003500· **MW** +3.287415· **X4v**-0.043587· **G(N..O)** (40)

n = 12 R^2 = 0.98653 s = 0.07885 F = 195.3105



a) R^2 = 0.9843, ecuaţia nr 6.16



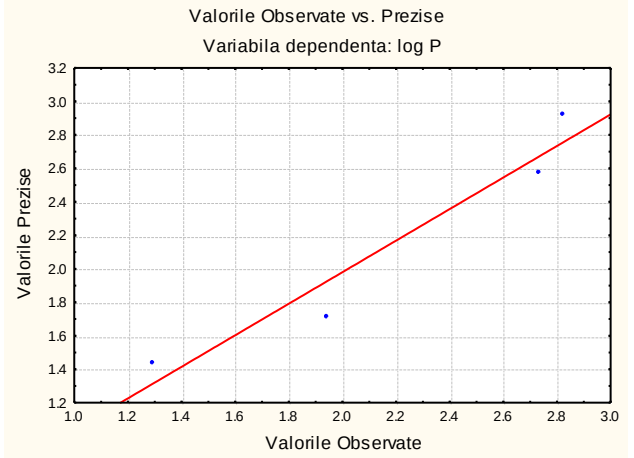
b) R^2 = 0.98653 ecuaţia nr 6.17

Figura 6. Reprezentarea grafică a proprietăţii observate vs. prezise pentru ec. 6.16 şi 6.17 pentru structurile din setul şcoală.

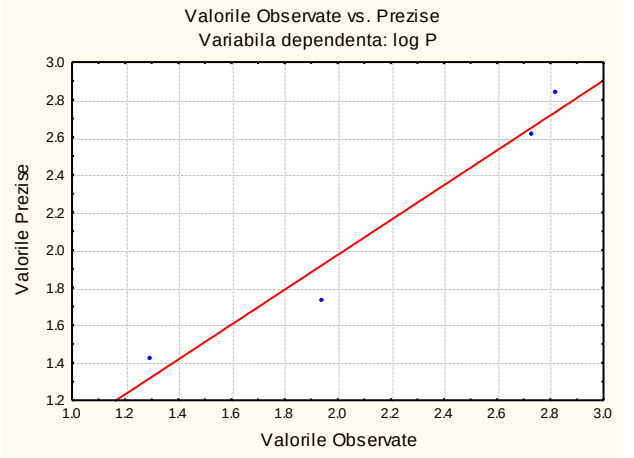
Orice model QSAR trebuie să fie validat cu un set de predicţie extern. În acest caz setul de predicţie este format din cei 4 derivaţi de 2-furiletenă la care prezicem valorile coeficientului de partiţie n-octanol/apă logP (Tabelul 9).

Tabelul 9. Proprietatea log P observată şi prezisă pentru structurile din setul de validare.

Nr. structură	log *P* **observat**	log *P* **prezis ec 45**	log *P* **prezis ec. 46**
35	2.820	2.924846	2.845186
36	2.730	2.580922	2.617008
37	1.290	1.438395	1.422570
38	1.940	1.709787	1.735940
R^2		**0.9333**	**0.9585**



a) **R^2** = 0.9333
log *P*(obs) = 0.9425*logP(calc) + 0.0948
log P (calc) – ecuaţia nr. 6.16



b) **R^2 =0.9585**
log *P*(obs) = 0.9278* logP(calc) + 0.1186
log P (calc) – ecuaţia nr. 6.17

Figura 7. Reprezentarea grafică a proprietăţilor oservate vs. prezise pe setul extern de validare.

Ecuaţia explică 95,85% din varianţa lui log P, această valoare arată abiliatatea crescută a modelului de predicţie (ecuaţia 39). În figura nr. 7.b este reprezentată relaţia dintre valorile observate şi cele prezise ale lui log P. În acest sens, ecuaţia obţinută cu descriptorii MW , X4v, G(N..O) este cea mai bună pentru predicţia proprietaţii log P.
Parcurgând aceleaşi etape am realizat prezicerea proprietăţii pe setul şoală şi validarea lui pe setul de predicţie a coeficientului de partiţie n-octanol/apă log P, dar de această dată cu ajutorul descriptorilor calculaţi cu programul TOPOCLUJ. Rezultatele obţinute sunt prezentate în continuare. Cei mai semnificativi descriptori calculaţi cu programul TOPOCLUJ sunt prezentaţi în tabelul 10 alături de proprietatea log P pentru structurile din setul şcoală determinat pe baza similarităţii moleculare cu liderii din setul de predicţie extern.

Tabelul 10. Descriptorii topologici şi coeficientul de partiţie observat log*P* pentru setul de predicţie.

Nr. Structură	CS[LM [Electronegativity]]	CS[Sh[W4 [Charge_Adjacency]]]	IE[CfMax [Density]]	VAA1	VAD1	Pobs
35	410	22	190	7.1	36	2.49
36	410	19	170	7.1	36	2.37
37	290	11	62	6.1	27	1.56
38	340	17	120	6.6	31	1.92

Ecuţiile utilizate pentru predicţie sunt cele calibrate în clusterul şcoală (în forma lor normalizată de Matlab). Aceste ecuaţii prezintă un coeficient de corelaţie bun (R^2 = 0.986337 - ec. 6.20 , R^2 = 0.960186 –ec. 6.19, R^2 = 0.911236 - ec.41) (Figura 8).

Regresie bivariabilă

logP= 3.562116+0.015455· **IE[CfMax[Density]]** -0.109763· **VAD1** (41)
n = 12 R^2 = 0.911236 s = 0.519627 F = 46.19599

Regresie multi-variabilă

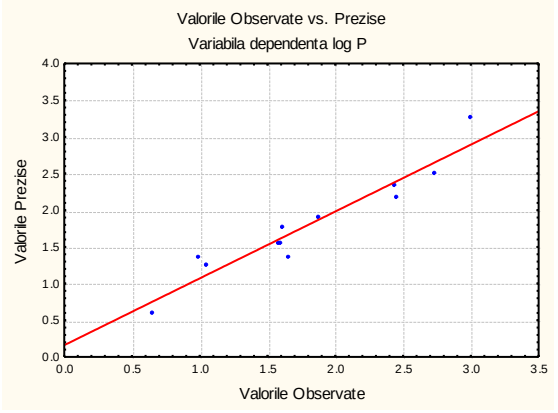
logP= 7.99004+0.05992· **CS[Sh[W4[Charge_Adjacency]]]** +0.01853· **IE[CfMax[Density]]** -1.45456· **VAA1** (42)
n = 12 R^2 = 0.960186 s = 0.233074 F = 64.31076

logP= 12.43950+0.01138· **CS[LM[Electronegativity]]**+ 0.06207· **CS[Sh[W4[Charge_Adjacency]]]**+0.01658· **IE[CfMax[Density]]** -2.70391· **VAA1** (43)
n = 12 R^2 = 0.986337 s = 0.079982 F = 126.3344

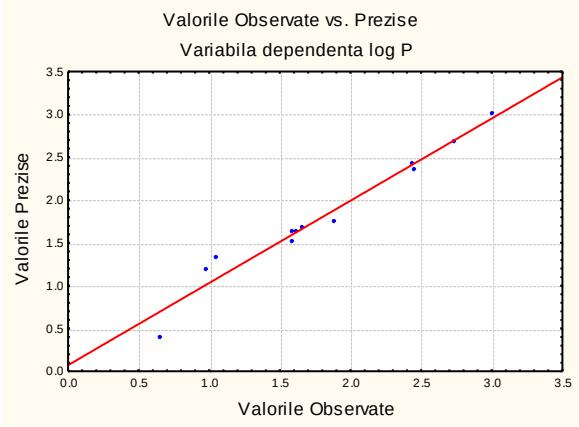
Modelele QSAR trebuie să fie validate pe un set de predicţie extern. Setul de predicţie este format din cei 4 derivaţi de 2-furiletenă la care prezicem valorile coeficientului de partiţie n-octanol/apă logP (Tabelul 12).

Tabelul 12. Proprietatea log P observată şi prezisă pentru structurile din setul de validare

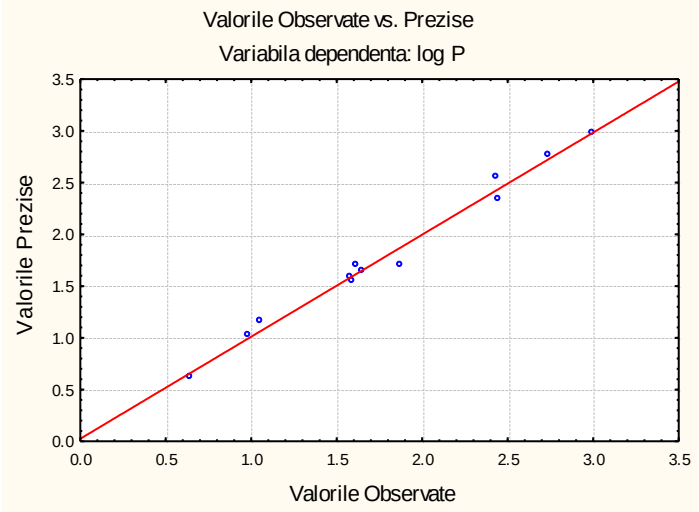
Nr. structură	log *P* observat	log *P* prezis ec 47	log *P* prezis ec. 48	log *P* prezis ec. 49
35	2.820	2.547183	2.501652	2.424993
36	2.730	2.238073	1.951290	1.907205
37	1.290	1.556749	0.925237	0.957875
38	1.940	2.014115	1.632225	1.509150
R^2		**0.924**	**0.9037**	**0.9143**



a) R^2 = 0.911236, ecuaţia nr 41



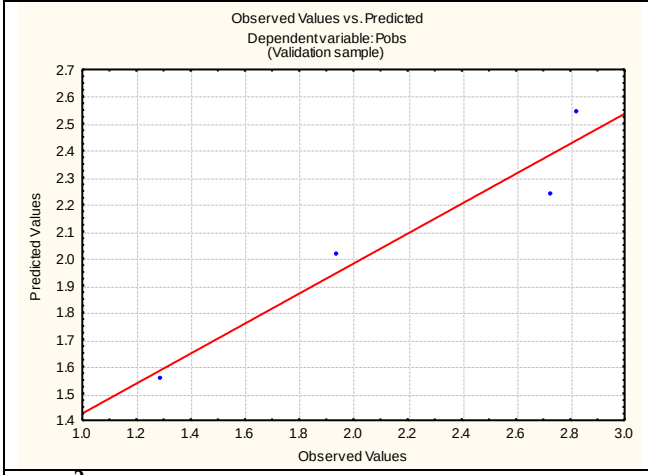
b) R^2 = 0.960186, ecuaţia nr 42



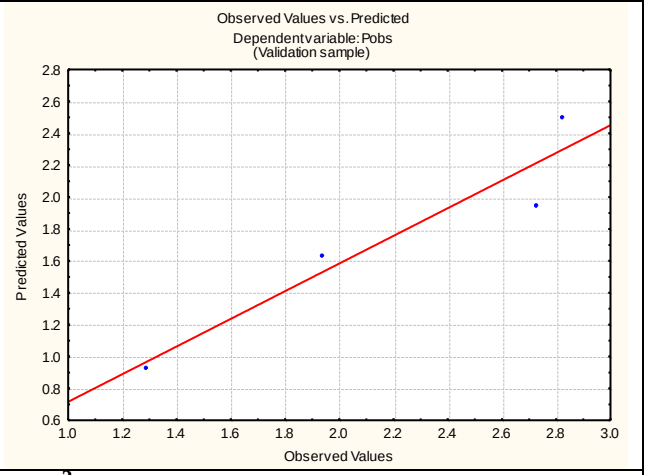
c) R^2 = 0.986337, ecuaţia nr 43

Figura 8. Reprezentarea grafică a proprietăţii observate vs. prezise pentru ec. 6.18, 6.19 şi 6.20 pentru structurile din setul şcoală.

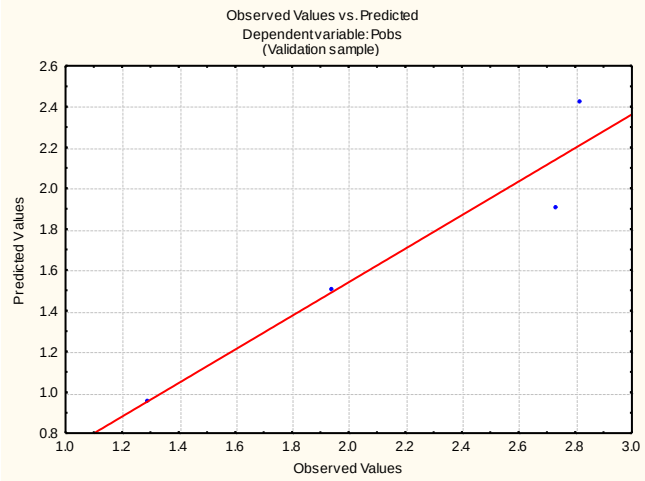
Descriptorii topologici implicaţi în această ecuaţie sunt IE[CfMax[Density]] şi VAD1. Reprezentarea grafica a acestora este prezentat in figura 9.



a) $\mathbf{R^2 = 0.924}$
log *P*(obs) = 0.5554*logP(calc)+0.87
log P (calc) – ecuaţia nr. 47



b) $\mathbf{R^2 = 0.9037}$
log *P*(obs) = 0.8674* logP(calc) - 0.1513
log P (calc) – ecuaţia nr. 48



c) $\mathbf{R^2 = 0.9143}$
log *P*(obs) = 0.8228*logP(calc) - 0.1062
log P (calc) – ecuaţia nr. 49

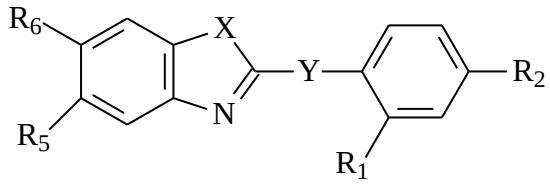
Figura 9. Reprezentarea grafică a proprietăţilor oservate vs. prezise pe setul extern de validare.

Noi modele QSAR pentru prezicerea activitaţii biologice a derivaţilor de benzoxazol/ benzimidazol

Bolile micotice sunt produse de microorganisme încadrate sistematic în regnul Fungi. Ele se găsesc în număr mare în mediul înconjurător. Majoritatea au adoptat mediul de viaţă saprobiotă, dar unele dintre ele s-au adaptat la viaţa parazitară. Se citează că peste 300 de specii au fost găsite a fi patogene pentru animale. Ciupercile obligatoriu parazite aparţin, în cea mai mare parte, categoriei dermatomicetelor. Ele nu se dezvoltă în mediu ci doar pot supravieţui şi se pot transmite la organismele vii prin contagiune directă sau indirectă.

Obţinerea descriptorilor moleculari

Recent s-au raportat sinteza şi activităţiile a câtorva derivaţi benzoxazolici/ benzimidazolici împotriva speciei C. Albicans. Analiza relaţiilor cantitative structură activitate (QSAR) a fost mai mult răspândită şi eficient folosită în studiile teoretice ale designului medicamentelor. Modelele QSAR propun cercetătorilor în domeniu sinteze orientate, permiţând realizarea unor economii de timp, bani şi energie, ameliorând cercetarea farmaceutică (Figura 10, Tabelul 13).



R_1= H, Cl, F, NO_2, CH_3, OCH_3	Y= —, CH_2, CH_2O, CH_2S
R_2= H, Cl, Br, NH_2, CH_3, OCH_3	X= O, NH
R_5=H, CH_3	
R_6= H, CH_3	

Figura 10. Activitatea antimicotică a derivaţilor de benzoxazol şi benzimidazol împotriva speciei C.Albicans

Tabelul 13. Valorile log 1/C observat pentru setul de derivati benzoxazolici/benzimidazolici.

	R_1	**R_2**	**R_5**	**R_6**	**X**	**Y**	**log 1/C Obs.**
1	Cl	H	CH_3	H	O	-	3.989
2	OCH_3	H	CH_3	H	O	-	3.980
3	NO_2	H	CH_3	H	O	-	4.007
4	Cl	Cl	CH_3	H	O	-	4.046
5	CH_3	CH_3	CH_3	H	O	-	3.977
6	OCH_3	OCH_3	CH_3	H	O	-	4.032
7	Cl	H	H	CH_3	O	-	3.989
8	OCH_3	H	H	CH_3	O	-	3.980
9	F	H	H	CH_3	O	-	3.958
10	NO_2	H	H	CH_3	O	-	4.007
11	Cl	Cl	H	CH_3	O	-	4.046
12	CH_3	CH_3	H	CH_3	O	-	3.977
13	OCH_3	OCH_3	H	CH_3	O	-	4.032
14	H	H	CH_3	H	O	CH_2	4.251
15	H	Br	CH_3	H	O	CH_2	4.383
16	H	NH_2	CH_3	H	O	CH_2	4.280
17	H	H	H	CH_3	O	CH_2	4.251
18	H	H	CH_3	H	NH	CH_2	4.249
19	H	Cl	CH_3	H	NH	CH_2	4.312
20	H	Br	CH_3	H	NH	CH_2	4.382
21	H	NH_2	CH_3	H	NH	CH_2	4.278
22	H	H	CH_3	H	O	CH_2O	3.980
23	H	H	CH_3	H	O	CH_2S	4.009
24	H	H	CH_3	H	NH	CH_2S	4.007
25	H	Cl	CH_3	H	NH	CH_2O	4.037

Descriptorii topologici utilizaţi în studiile QSAR sunt accesibili şi pot fi uşor calculaţi cu ajutorul programelor software. Setul descriptorilor moleculari utilizaţi în acest studiu a fost calculat cu pachetul software DRAGON.[45] Structurile au fost optimizate folosind metoda Hamiltoniană semi empirică PM3 disponibilă în softul HyperChem.[137]

Prelucrarea şi analiza datelor

Scopul acestui studiu este de a dezvolta un model QSAR nou şi mult mai eficient decât celelalte obţinute anterior pentru prezicerea activităţii antifungice a bezoxazolilor /benzimidazolilor substituiţi în poziţia 2 şi cu grupare metil în poziţia 5 sau 6.

Descriptorii topologici obţinuţi cu programul Dragon şi activitatea biologică (antifungică) log 1/C pentru acest set de structuri sunt prezentaţi în tabelul 14.

Tabelul 14. Descriptorii topologici obţinuţi cu programul Dragon, valorile experimentale şi valorile prezise log 1/C ale derivaţilor de benzoxazoli/ benzimidazoli.

	MW	nCs	nHDon	Obs. log 1/C	Calc. log 1/C
1	243.7	0	0	3.989	3.988
2	239.29	0	0	3.980	3.981
3	254.26	0	0	4.007	4.006
4	278.14	0	0	4.046	4.046
5	237.32	0	0	3.977	3.977
6	269.32	0	0	4.032	4.031
7	243.7	0	0	3.989	3.988
8	239.29	0	0	3.980	3.981
9	227.25	0	0	3.958	3.960
10	254.26	0	0	4.007	4.006
11	278.14	0	0	4.046	4.046
12	237.32	0	0	3.977	3.977
13	269.32	0	0	4.032	4.031
14	223.29	1	0	4.251	4.251
15	302.18	1	0	4.383	4.384
16	238.31	1	2	4.280	4.279
17	223.29	1	0	4.251	4.251
18	222.31	1	1	4.249	4.251
19	256.75	1	1	4.312	4.309
20	301.2	1	1	4.382	4.384
21	237.33	1	3	4.278	4.278
22	239.29	0	0	3.980	3.981
23	255.36	0	0	4.009	4.008
24	254.38	0	1	4.007	4.007
25	272.75	0	1	4.037	4.038

Descriptorii moleculari semnificativii calculaţi cu programul Dragon sunt: MW- masa moleculară, nCs- numărul atomilor de C secundar (sp^3) şi nHDon- numărul atomilor donori ai legăturii de H (N si O).

S-a găsit ecuaţia QSAR utilizând analiza de regresie multiliniară. Calitatea modelului obţinut este dată de pătratul coeficientului de regresie(R^2) , raportul Fischer, eroarea standard estimată (s) şi de procedeul leave-one-aut (LOO) ca şi procedură de cross-validare.
Ecuaţia prezintă o valoare bună a coeficientului de corelaţie R^2=0.999 în regresia multivariabilă cu 3 descriptori. Aceasta corelaţie este semnificativ mai bună decât cele raportate anterior (R^2=0.94). Valorile experimentale şi cele calculate ale lui log 1/C pentru acest set sunt reprezentate grafic în Figura 11.

log 1/C= 3.577225 + 0.001686* **MW** + 0.297347 * **nCs** + 0.00122 * **nHDon** (44)
n=25 R2=0.999932 s=0.0029 F= 103092.3

Pobs Std.Dev.= 0.145464
Pobs Std.Err = 0.029093

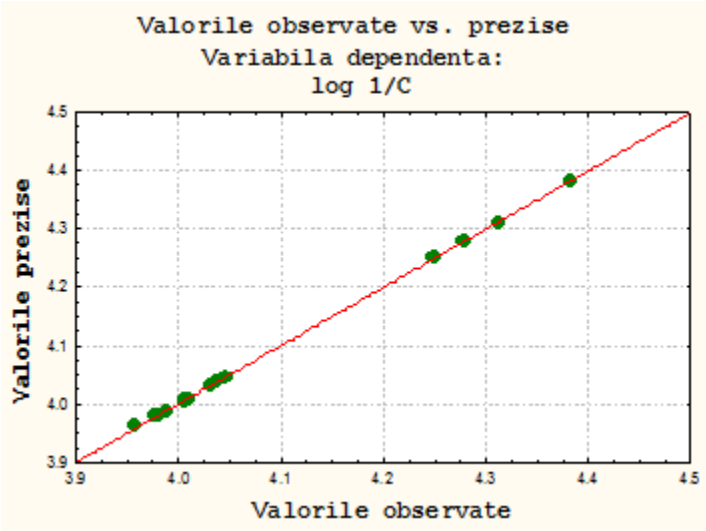


Figura 11. Reprezentarea grafica a valorilor observate vs. prezise pentru ecuaţia 6.21.

Pe baza datelor obţinute s-a propus un nou model dezvoltat pentru predicţia directă a activităţii biologice. S-au construit clusteri de similaritate pentru fiecare moleculă din setul scoală şi s-au realizat predicţii în subseturile congenerice obţinute. Descriptorii moleculari semnificativi sunt MW şi nCs. S-au găsit ecuaţiile QSAR utilizând analiza regresiei multivariate (Tabelul 15).
Scopul acestei metode este acela de obţine cel mai bun model pentru prezicerea viitoare a activităţii antifungice pentru alte structuri noi. Pentru fiecare structură, luată ca şi lider în testul de similaritate, s-a calculat ecuaţia de predicţie bivariată pentru fiecare subset de similaritate. Fiecare din cele 25 de ecuaţii de învăţare au R^2=0.9999 iar acestă cifră variază doar începând de la a 5-a zecimală. Ecuaţia de predicţie obţinută prezintă un coeficient de corelaţie foarte bun, R^2=0.999, fiind o regresie bivariabilă (Figura 12).
Corelaţia este de asemenea mult mai bună decât cea raportată anterior (R^2=0.94) astfel încât ea este excelentă.

Tabelul 15. Valorile parametrilor regresiei precum şi valorile observate şi calculate pentru log 1/C, ec. Y_{calc}=a + b*MW +c*nCs.

Nr.	MW	nCs	log 1/C Obs.	Parametrii ecuaţilor de predicţie			log 1/C Calc.
				a	b	c	
1	243.70	0	3.989	3.529574	0.001885	0.300421	3.989006
2	239.29	0	3.980	3.578017	0.001684	0.298368	3.980871
3	254.26	0	4.007	3.577699	0.001684	0.298482	4.005963
4	278.14	0	4.046	3.577586	0.001685	0.298405	4.046282
5	237.32	0	3.977	3.577923	0.001684	0.298388	3.97753
6	269.32	0	4.032	3.586488	0.001650	0.299659	4.030916
7	243.70	0	3.989	3.577964	0.001683	0.299659	3.988105
8	239.29	0	3.980	3.578017	0.001684	0.298368	3.980871
9	227.25	0	3.958	3.579723	0.001677	0.298309	3.960813
10	254.26	0	4.007	3.577699	0.001684	0.298482	4.005963
11	278.14	0	4.046	3.586385	0.001651	0.299609	4.045527
12	237.32	0	3.977	3.586968	0.001649	0.299502	3.978302
13	269.32	0	4.032	3.586488	0.001650	0.299659	4.030916
14	223.29	1	4.251	3.529649	0.001885	0.300392	4.250924
15	302.18	1	4.383	3.582466	0.001666	0.299986	4.385951
16	238.31	1	4.280	3.585095	0.001656	0.299324	4.279045
17	223.29	1	4.251	3.529649	0.001885	0.30032	4.250924
18	222.31	1	4.249	3.529748	0.001884	0.300463	4.249153
19	256.75	1	4.312	3.585044	0.001656	0.298955	4.309212
20	301.20	1	4.382	3.583753	0.001661	0.299843	4.383946
21	237.33	1	4.278	3.585631	0.001654	0.299458	4.277595
22	239.29	0	3.980	3.587052	0.001649	0.299486	3.981574
23	255.36	0	4.009	3.585960	0.001652	0.299638	4.007876
24	254.38	0	4.007	3.585968	0.001652	0.299608	4.006286
25	272.75	0	4.037	3.586233	0.001651	0.299593	4.036649

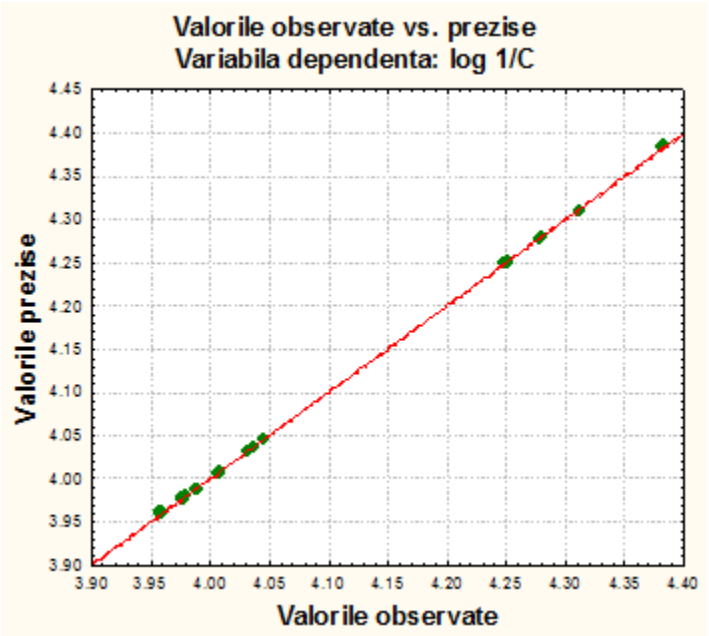


Figura 12. Reprezentarea grafica a log 1/C observat fata de log 1/C prezis.

Setul de derivaţi de benzoxazoli/benzimidazoli substituiţi în poziţia 2 şi având gruparea metil în poziţia 5 sau 6, testat anterior pentru activitatea antibacteriană împotriva lui C. Albicans, a fost analizat prin relaţia cantitativă structură-activitate biologică şi contribuţiile la activitate pentru efectele structurale şi funcţionale au fost determinate utilizând procedeul de regresie multiplă. Rezultatele QSAR obţinute relevă faptul că substituentul din poziţia Y (descriptorul nCs) joacă un rol important şi îşi aduce o contribuţie importantă la activitatea antibacteriană.

CONCLUZII

Sinteza şi implementarea pe piaţă a unor medicamente performante care să raspundă într-un grad cât mai ridicat la soluţionarea unor probleme de sănătate acute societăţii reprezintă un deziderat important al industriei farmaceutice în special, şi al cercetatorului în particular.

Găsirea unor metode teoretice prin care se reduc costurile şi timpul necesar găsirii şi sintezei compuşilor biologic activi cu aplicabilitate practică, reprezintă ţelul oricărui cercetator. Una dintre aceste metode îl oferă chimia matematică prin folosirea diferitelor tehnici şi modele, prin care structurile sunt evaluate şi cuantificate într-un numar.

Teza de faţă urmareşte tratarea şi obţinerea diferitelor modele teoretice cu aplicabilitate în design-ul de molecule bioactive, în special prin prisma similarităţii structurale. Obţinerea unor metode de evaluare cantitativă a similarităţii este doar una dintre problemele care trebuie rezolvate de către *"proiectantul"* de substanţe bioactive. Noţiunile centrale ale oricărui sistem de căutare a similarităţii intermoleculare sunt criteriul de similaritate şi măsura utilizată pentru cuantificarea similarităţii.

Astfel, este tratată problema similarităţii moleculare, urmărind şi dezvoltând diferite aspecte ale acestei teorii. În acest sens au fost prezentate pe larg:

- descrierea cantitativă a similarităţii
- indici de similaritate moleculară
- descriptori graf-moleculari
- programul software TOPOCLUJ-SIMIL

Au fost analizate diferite tehnici şi metode de suprapunere moleculară, similaritate, aplicate pe clase de structuri cu activitate biologică.

Aceste metode includ:

- analiza de clusteri
- analiza componentelor principale
- analiza de factori

S-au propus şi realizat diferite modele de clusterare pe baza similarităţii, utile pentru modelarea şi prezicerea proprietăţilor moleculare (biologice sau fizico-chimice), având o contribuţie majoră la studiile cantitative ale relaţiei structură - activitate biologică QSAR pentru diferite tipuri de molecule.

Rezultatele modelelor QSAR propuse urmăresc într-un mod excelent proprietăţile biologice ale claselor de structuri propuse. Astfel, modelele propuse cu 3 variabile independente, MW, X4v şi G(N...O), având R^2=0.98653, prezintă o capacitate avansată de predicţie a proprietăţii log P în setul de validare având R^2= 0.9585.

Rezultatele obţinute devansează rezultatele similare din literatură, având o abilitate mai bună de predicţie şi în acelaşi timp un numar redus de descripori utilizaţi în ecuaţia de regresie, fiind utile în estimarea proprietăţilor fizico-chimice şi biologice a unor noi derivaţi de 2-furiletenă.

Modelele QSAR de succes obţinute prezintă o corelaţie statistică semnificativă între caracteristicile chimice ale compuşilor (descriptorilor) şi activitatea biologică. În urma analizelor ecuaţiilor obţinute în urma modelarii seturilor de compuşi, s-a ajuns la concluzia, că, este necesar în toate cazurile să se evite predicţia activităţii biologice a compuşilor care au structură foarte diferită de a compuşilor din setul şcoală. Oricum, oricare dintre aceste proceduri poate fi folosită separat ca parte a studiilor computaţionale cu multiple scopuri de atins. Sunt necesare aceste tehnici analitice din ce în ce mai mult pentru generarea,

interpretarea şi redarea mai multor informaţii despre molecule cu potenţial biologic şi nu numai.

De asemenea, s-a observat faptul că, numai un subset de descriptori ai structurilor moleculare, care sunt cei mai importanţi şi semnificativi din punct de vedere statistic, sunt selectaţi pentru a descrie o activitate biologică aleasă.

Rezultatele obţinute au fost publicate in urmatoarele reviste:

1. Costescu, A., Moldovan, C.D., Diudea, M.V., QSAR modeling of steroid hormones, *Match* 55 (2), pp. 315-329, **2006**.
2. Moldovan, C.D., Costescu, A., Katona, G., Diudea, M.V., A novel QSAR approach in modeling antifungal activity of some 5-or 6-methyl-2-substituded benzoxazoles/benzimidazoles against C. albicans using molecular descriptors, *Match* 60 (3), pp. 977-984, **2008**.
3. Costescu, A., Moldovan, C.D., Katona, G., Diudea, M.V., QSAR modeling of human catechol O-methyltransferase enzyme kinetics, *Journal of Mathematical Chemistry* 45 (2), pp. 287-294, **2009**.
4. Moldovan, C.D., Costescu, A., Katona, G., Diudea, M.V., Application to QSAR studies of 2-furylethylene derivatives, *Journal of Mathematical Chemistry* 45 (2), pp. 442-451, **2009**.

INDEX

Bibliografie selectiva

1 Rosen, R. in: Johnson, M. A.; Maggiora, G. M. Eds. Concepts and Applications of Molecular Similarity, Wiley, New York, **1990**, Chap. 12, 369-382.

2 Mezey, P. G. Three-Dimensional Topological Aspects of Molecular Similarity. In: Johnson, M.A.; Maggiora, G. M. Eds. Concepts and Applications of Molecular Similarity, Wiley, New York, **1990**, Chap. 11, 321-368.

3 Randić, M. Design of Molecules with Desired Properties. A Molecular Similarity Approach to Property Optimization. In: Johnson, M. A.; Maggiora, G. M. Eds. Concepts and Applications of Molecular Similarity, Wiley, New York, **1990**, Chap. 5, 77-145.

4 Maggiora, G. M.; Johnson, M. A. Introduction to Similarity în Chemistry. In: Johnson, M. A.; Maggiora, G. M. Eds. Concepts and Applications of Molecular Similarity, Wiley, New York, **1990**, Chap. 1, 1-13.

5 Tsai, C. -c.; Johnson, M. A.; Nicholson, V.; Naim, M. Eds., Graph Theory and Topology in Chemistry, Elsevier, Amsterdam, **1987**, 231.

6 Balaban, A. T.; Chiriac, A.; Motoc, I.; Simon, Z. Steric Fit în QSAR (Lecture Notes în Chemistry, Vol. 15), Springer, Berlin, **1980**, Chap. 6.

7 Kvasnička, V.; Pospichal, J. Fast Evaluation of Chemical Distance by Tabu Search Algorithm. *J. Chem. Inf. Comput. Sci.* **1994**, 34, 1109-1112.

8 Diudea, M.V. Layer Matrices în Molecular Graphs, *J. Chem. Inf. Comput. Sci.* **1994**, 34, 1064-1071.

9 Ugi, I.; Wochner, M.A.; Fontain, E.; Bauer, J.; Gruber, B.; Karl, R. Chemical Similarity, Chemical Distance, and Computer Assisted Formalized Reasoning by Analogy, in: Maggiora, G. M. Eds. Concepts and Applications of Molecular Similarity, Wiley, New York, **1990**, Chap. 9, 239-288.

10 Basak, S.C.; Magnusson, V.R.; Niemi, G.J.; Regal, R.R., Determining Structural Similarity of Chemicals Using Graph-Theoretic Indices, *Discr. Appl. Math.* **1988**, 19, 17-44.

11 M. Randić, W.L. Woodworth, A. Graovac, Unusual random Walks, *Int. J. Quantum Chem.*, **1983**, 24, 435-452.

12 M. Randić, Generalized molecular descriptors, *J. Math. Chem.*, **1991**, 7, 155-168.

13 N. Trinajstić, Chemical Graph Theory, CRC Press. Inc., Boca Raton Florida, **1983**.

14 A.T. Balaban, I. Moţoc, D. Bonchev, O. Mekenyan, Topological Indices for Structure - Acvtivity Correlations, *Top. Curr. Chem.*, **1993**, 114, 21-55.

15 D.H. Rouvray, The challange of characterizing branching in molecular species, *Discr. Appl. Math.*, **1988**, 19, 317-338.

16 M. Randić, Design of molecules with desired properties. A molecular similarity approach to property optimization, in "Concepts and Applications of Molecular Similarity, M.A. Johnson and G.M. Maggiora, Eds., John Wiley & Sons, Inc., **1990**.

17 R. Carbó, L. Leyda, and M. Arnau, *Int. J. Quantum Chem.*, **1980**, 17, 1185–1189.

18 D. Ciubotariu, S. Mureşan, V. Gogonea, M. Medeleanu, D. Dragoş, Relaţii Cantitative Structură Chimică-Activitate Biologică (QSAR), Ed.Mirton, Timişoara, **1996**

19 E. Overton, Studien über die Narkose, Fischer, Jena, **1901**; A. Meyer, *Arch. Exptl. Pathol. Pharmakol.*, **1899**, 42, 110.

20 J.N. Langley, *J. Physiol.(*London), **1908**, 1, 339.

21 P. Ehrlich, *Ber. Dtsch. Chem. Ges.*, **1909**, 42, 17.

22 A. R. Cushny, Biological Relations of Optically Isomeric Substances, Balliere, Tindall and Cox, London, **1926**.

23 I. Moţoc, Structura moleculelor şi activitatea biologică, Ed. Facla, Timişoara, 1980.

24 E.J. Ariëns, Stereochemistry: A Source of Problems in Medicinal Chemistry, *Med.Res.Rev.*, **1986**, 6, 451-466.

25 E.J. Ariëns, Stereochemistry in the Analysis of Drug-Action, Part II. *Med. Res. Rev.*, **1987**, 7, 367-387.
26 E.J. Ariëns, Stereochemical Implications of Hybrid and Pseudohybrid Drugs, Part III. *Med. Res. Rev.*, **1988**, 8, 309-320.
27 I.D. Resa, S. Petrescu, M. Precupas, A. Câra, Probleme de statistica rezolvate pe calculator, Ed. Facla, Timisoara, **1984**.
28 D. McCormick, A. Roach, Measurement, Statistics and Computation, John Wiley & Sons, London, **1987**.
29 Katritzky, A. R.; Maran, U.; Lobanov, V. S.; Karelson, M. Perspective: Structurally diverse quantitative structure-property relationship correlations of technologically relevant physical properties. *J. Chem. Inf. Comput. Sci.* **2000**, 40, 1-18.
30 Robinson, D. D.; Winn, P. J.; Lyne, P. D.; Richards, W. G. Self-Organizing Molecular Field Analysis: A Tool for Structure-Activity Studies. *J. Med. Chem.* **1999**, 42, 573-583.
31 Cramer, R. D., I.; Patterson, D. E.; Bunce, J. D. Comparative Molecular Field Analysis (CoMFA). 1. Effect of Shape on Binding of Steroids to Carrier Proteins. *J. Am. Chem. Soc.* **1988**, 110, 5959-5967.
32 Coats, E. The CoMFA Steroids as a Benchmark Dataset for Development of 3D QSAR Methods. Perspect. *Drug DiscoV. Design.* **1998**, 12/13/14, 199-213.
33 Dunn, J. F.; Nisula, B. C.; Rodbard, D. Transport of Steroid Hormones: Binding of 21 Endogeneous Steroids to Both Testosterone- Binding Globulin and Corticosteroid-Binding Globulin in Human Plasma. *J. Clin. Endocrin. Metab.* **1981**, 53, 58-68.
34 Tuppurainen K, Viisas M, Laatikainen R, Perakyla M.: Evaluation of a Novel Electronic Eigenvalue (EEVA) Molecular Descriptor for QSAR/QSPR Studies: Validation Using a Benchmark Steroid Data Set. *J. Chem. Inf. Comput. Sci.* **2002**, 42, 607-613
35 Rios-Santamaria I, Garcia-Domenech R, Cortijo J, Santamaria P, Morcillo E J and Galvez J, *Internet Electronic Journal of Molecular Design*, **2002**, 1, 70.
36 Galvez J, Garcia-Domenech R, Salabert M T and Soler R, *J Chem Inf Comput Sci*, **1994**, 34, 520.
37 Todeschini, R.; Consonni, V. Handbook of molecular descriptors. Wiley-VCH: Weinheim, Germany, **2000**.
38 Polanski, J.; Walczak, B. The Comparative Molecular Surface Analysis (COMSA): a novel Tool for Molecular Design. *Comput. Chem.* **2000**, 24, 615-625.
39 M. Pastor, G. Cruciani, I. McLay,S. Pickett and Sergio Clementi: Grid Independent Descriptors (GRIND): A Novel Class of Alignment- Independent Three-Dimensional Molecular Descriptors. *J. Med. Chem.* **2000**, 43, 3233-3243.
40 M.A. Cabrea Pérez et al.: Experimental and theoretical determination of physicochemical properties in a novel family of microcidal compounds. *European Bulletin of Drug Research*, **2001**, 9, 1.
41 Yovani Marrero Ponce et al.: Atom, Atom-Type, and Total Linear Indices of the Molecular Pseudograph's Atom Adjacency Matrix: Application to QSPR/QSAR Studies of Organic Compounds, *Molecules* **2004**, 9, 1100-1123.
42 Dore, J. Ch.; Viel, C. Antitumoral Chemoterapy. X. Cytotoxic and Antitumoral Activity of β-Nitrostyrenes and Nitrovinyl Derivatives. *Farmaco.* **1975**, 30, 81-109.
43 Castañedo, N.; Goizueta, R.; Perez, J.; Gonzalez, J.; Silveira, E. Cuesta, M.; Martinez, A.; Lugo,E.; Estrada, E.; Carta, A.; Navia, O.; Delgado, M. Cuban Pat. 22446, 1994; Can. Pat. 2,147,594,**1999**.
44 Blondeau, J. M.; Castañedo, N.; Gonzalez, O.; Medina, R.; Silveira, E. In Vitro Evaluation of G-1: A Novel Antimicrobial Compound. Antimicrob. Agents *Chemother*. **1999**, 11, 1663-1669.

45 DragonProfessional version 5.4, software -**2006** (http://www.talete.mi.it/dragon_net.htm)