Kvantitativni odnos strukturnih odlika i aktivnosti molekula

Šelko, Ivanka

Sveučilište u Zagrebu
PRIRODOSLOVNO-MATEMATIČKI FAKULTET
Kemijski odsjek

Ivanka Šelko

Studentica 3. godine Preddiplomskog sveučilišnog studija KEMIJA

KVANTITATIVNI ODNOS STRUKTURNIH ODLIKA I AKTIVNOSTI MOLEKULA

Završni rad

Rad je izrađen u Zavodu za fizikalnu kemiju

Mentor rada: prof. dr. sc. Branimir Bertoša

Zagreb, 2021.

Datum predaje prve verzije Završnog rada: 15. srpnja 2021.

Datum ocjenjivanja Završnog rada i polaganja Završnog ispita: 24. rujna 2021.

Mentor rada: prof. dr. sc. Branimir Bertoša Potpis:

Sadržaj

§ Sažetak

Kvantitativni odnos strukture i reaktivnosti (engl. *Quantitative structure–activity relationships*, QSAR) je metoda koja je pronašla svoje mjesto u znanosti, industriji, a koriste je i različita regulatorna tijela kako bi predvidjela utjecaj na okoliš pojedinih kemijskih tvari. Metoda se počela koristiti šezdesetih godina prošlog stoljeća i na njezin razvoj znatno utječe razvoj računala. Povezivanje strukturnih svojstava i reaktivnosti (najčešće biološka aktivnost) može omogućiti predviđanje aktivnosti novih, još nesintetiziranih spojeva, kod kojih će se strukturnim izmjenama poboljšati aktivnosti u odnosu na postojeće spojeve.

Izgraditi pouzdan QSAR model je izazovan zadatak koji započinje prikupljanjem podataka o strukturi i biološkoj aktivnosti izabranih spojeva. Različitim statističkim i matematičkim postupcima nastoji se pronaći jednadžba koja povezuje strukturu i biološku aktivnost. Važan korak je validacija modela koja se provodi, kako na istom skupu podataka s kojim je model izgrađen, tako i na skupu za vrednovanje koji nije korišten u razvoju modela. Ukoliko model zadovolji kriterije vrednovanja, može se smatrati pouzdanim unutar domene primjenljivosti koju je potrebno odrediti.

§ 1. UVOD

Kvantitativni odnos strukture i reaktivnosti (engl. *Quantitative structure–activity relationships*, QSAR) je metoda koja primjenjujući matematičke i statističke tehnike nastoji pronaći povezanost između strukture određenog spoja i njegove (često biološke) aktivnosti.[1] Cilj ovog rada je prikazati pojedine etape u izgradnji i validaciji pouzdanog QSAR modela.

Prvi radovi o povezanosti strukture i funkcije su toksikološke studije iz druge polovice 19. st. kada je Cros 1863. godine uočio povezanost toksičnosti primarnih alkohola i njihove topljivosti u vodi.[2] Prvi QSAR model ima svoje porijeklo u istraživanjima kojima su A. Crum -Brown i T. Fraser proučavali utjecaj prirodnih alkaloida na mišićnu paralizu. U svom radu 1868. godine, povezanost strukturnih promjena i aktivnosti prikazali su matematičkom jednadžbom koja se uz jasnije definiranje pojma konstitucija koristi i danas:[3]

$$\Delta \text{ (fiziološka aktivnost)} = f\,(\Delta \text{ konstitucija})$$

Uslijedio je rad u kojem je Richardson 1869. godine pokazao obrnuto proporcionalnu povezanost narkotičkog efekta primarnih alkohola i njihove molekulske mase,[4] te rad u kojem se Richet 1893. godine bavio istraživanjem povezanosti toksičnosti jednostavnih etera, alkohola i ketona i njihove topljivost u vodi.[5] H. Meyer, E. Overton i F. Baum su 1899. godine povezali djelotvornost narkotika i njegov koeficijent razdjeljenja između maslinovog ulja i vode.[6-8] Koeficijent razdjeljenja između oktanola i vode (maslinovo ulje je zamijenio oktanol) (P) poslužio je šezdesetak godina kasnije Corwinu Hanschu za izračunavanje konstanti hidrofobnosti (π). Corwin Hansch bavio se istraživanjem auksina, biljnih hormona rasta. Povezao je biološku aktivnost, hidrofobnost i Hammettove konstante te postao "otac modernog QSAR-a". Jedan od oblika te jednadžbe glasi:[9]

$$\log(1/c) = -k\pi^2 + k'\pi + \rho\sigma + k''$$

c – koncentracija tvari potrebna da izazove standardni odgovor (npr. koncentracija regulatora rasta koja uzrokuje 10 % veći rast biljnih stanica)

π – razlika u koeficijentu razdjeljenja oktanol – voda supstituiranog i osnovnog spoja

σ – Hammettova konstanta

ρ – reakcijska konstanta

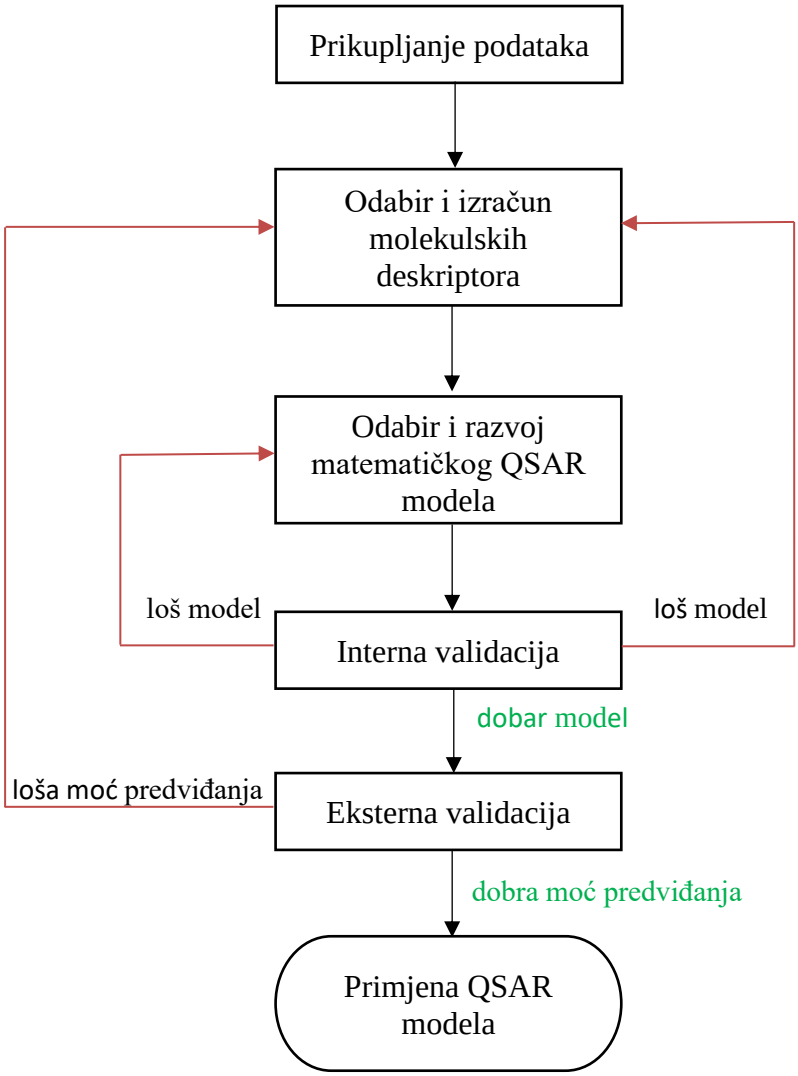
k, k', k'' – konstante dobivene regresijskom analizom

Od prve QSAR jednadžbe do danas objavljeno je približno 20 000 radova vezanih uz QSAR,[10] a razvoj QSAR tekao je paralelno s razvojem računala i dostupnošću kemijskih podataka. Uloga QSAR u farmaceutskoj industiji od velike je važnosti. Koristi se za optimizaciju vodećeg spoja (engl. *lead compounds*), predviđanje biološke aktivnosti, dizajn novih spojeva željene aktivnosti pri čemu uporaba QSAR metode smanjuje vrijeme i trošak potreban za sintezu novih lijekova i smanjuje broj eksperimenata na životinjama. Osim u farmaceutskoj industriji, metoda QSAR se koristi i u industriji proizvodnje i pronalaska pesticida i toksikologiji te je koriste regulatorna tijela pojedinih zemalja.

§ 2. PRIKAZ ODABRANE TEME

2.1. Metodologija QSAR-a

Razvoj pouzdanog QSAR modela sastoji se od nekoliko koraka koji su prikazani na slici 1. Organizacija za ekonomsku suradnju i razvoj (OECD) donijela je 2004. godine pet zahtjeva koje bi trebao ispuniti svaki model koji se koristi u regulatorne svrhe. Model treba imati definiran odgovor koji se prati, jasan algoritam, naznačenu domenu primjenljivosti, prikladne parametre za "dobrotu pristajanja" (engl. *goodness of fit*), robusnost i moć predviđanja te mehanističku interpretaciju ukoliko je moguća.[12]



Slika 1. Shematski prikaz razvoja i validacije QSAR modela (preuzeto i prilagođeno prema K. Z. Myint, X. Q. Xie, *Int. J. Mol. Sci.* **11** (2010) 3846 - 3866.).

2.2. Prikupljanje i obrada podataka

Ulazni podaci za razvoj QSAR modela su strukture izabranih spojeva i njihove (biološke) aktivnosti. Podatke možemo pronaći u znanstvenim radovima ili bazama podataka koje mogu biti javno dostupne (npr. PubChem[13], ChEMBL[14], ZINC[15]) ili komercijalne (npr. WDI[16], CAS REGISTRY[17]). Broj podataka koji ćemo koristiti ovisi o našem cilju, vremenu i računalu. Najmanji broj podataka iznosi 20 kako bi se izbjegla slučajna korelacija i prepodešavanje (engl. *overfitting)*. Optimalan broj iznosi oko 150 - 300.[18]

Poželjno je da su podaci o biološkim aktivnostima dobiveni pod istim eksperimentalnim uvjetima i iz istog izvora. Ukoliko to nije moguće, potrebno je uspostaviti korelaciju između mjerenja kako bi se ti podaci mogli koristiti u QSAR analizi. Mjerne jedinice kojima se iskazuje biološka aktivnost trebaju pripadati Međunarodnom sustavu mjernih jedinica (SI).

Prije početka razvoja QSAR modela potrebno je posebnu pažnju posvetiti anorganskim i organometalnim tvarima u odabranim podacima jer ih većina QSAR programskih paketa otežano procesira. Također, posebnu pažnju treba posvetiti mogućnosti da se u odabranim podacima nalaze isti spojevi pod različitim imenom i tretiranju izomera. Za takvu analizu podataka na raspolaganju su nam različiti programski alati poput MOE[19] ili ChemAxon[20].

2.3. Molekulski deskriptori

Molekulski deskriptor je konačni rezultat logičkog i matematičkog postupka koji transformira kemijsku informaciju kodiranu simboličkom reprezentacijom molekule u numeričku vrijednost.[21] Može se odrediti eksperimentalno ili teorijski (računalno). Eksperimentalno određeni molekulski deskriptori sadrže pogrešku mjerenja, dok oni određeni teorijski sadrže pogreške nastale zbog korištenja aproksimacija u računu. Prednost teorijski određenih deskriptora je u njihovoj lakoj dostupnosti. Deskriptore možemo klasificirati na nekoliko načina.

1) Prilikom izračuna deskriptora možemo:

a) istraživati cijelu molekulu, tako izračunane deskriptore nazivamo globalni (engl. *whole molecule descriptors*)

b) uzeti u obzir pojedini dio (supstituent), takvi se deskriptori nazivaju lokalni (engl. *fragment based*)

Najčešće korišteni globalni deskriptori su koeficijent razdjeljenja oktanol-voda, konstanta disocijacija kiselina, van der Waalsov obujam i sl.[22] Prednost lokalnih deskriptora je u njihovom lakšem izračunu i jednostavnoj interpretaciji utjecaja supstituenta na aktivnost čitave molekule. Hammettova elektronska konstanta je zasigurno najpoznatiji lokalni deskriptor.

2) Drugi način podjele deskriptora je prema osobinama koje opisuju.

Fizičko-kemijski deskriptori (relativna molekulska masa, talište, vrelište, koeficijent razdjeljenja oktanol-voda, molekulska refraktivnost…) opisuju fizička i kemijska svojstva nekog spoja. S koeficijentom razdjeljenja oktanol-voda je započeo razvoj QSAR metode, a njegova važnost je i u tome što log*P* predstavlja lipofilnost molekule. Lipofilnost je važno svojstvo spojeva koji su (potencijalni) lijekovi jer ono utječe na njihovu topljivost, apsorpciju, distribuciju, metabolizam i eliminaciju te na interakciju ligand-receptor.

Topološki deskriptori imaju svoje porijeklo u teoriji grafova. Struktura molekule se prikazuje „kemijskim grafom" koji ima dva elementa: vrhove koji predstavljaju atome i bridove koji predstavljaju kemijsku vezu. Na temelju povezanosti unutar kemijske strukture konstruiraju se matrice iz kojih se izračunavaju deskriptori. Jednostavan izračun i povezivanje s mnogim kemijskim svojstvima prednost je topoloških deskriptora. Jedan od prvih topoloških deskriptora je Wienerov indeks koji je Wiener u svom radu nazvao "*path number w*" i definirao ga kao sumu udaljenosti između bilo koja dva ugljikova atoma u molekuli izraženu kao broj ugljik-ugljik veza.[23] U razvoju topoloških indeksa sudjelovali su znanstvenici iz Hrvatske (Randićev indeks, Zagreb indeksi).

Geometrijski deskriptori se izvode, ili iz optimizirane trodimenzionalne strukture, ili iz kristalografskih koordinata. Obuhvaćaju kvantno kemijske deskriptore, deskriptore koji opisuju volumen molekule (npr. van der Waalsov volumen, geometrijski volumen), deskriptori koji opisuju površinu dostupnu otapalu, WHIM deskriptore koji opisuju veličinu, oblik, simetriju molekule i mnoge druge. Kvantno kemijski deskriptori su dobiveni kvantno mehaničkim računom. Obuhvaćaju molekulske energije, elektronsku gustoću, molekulski elektrostatski potencijal (MEP), energiju najviše zauzete orbitale (E_{HOMO}), energiju najniže nezauzete orbitale (E_{LUMO}).

3) Deskriptore možemo podijeliti i s obzirom na njihovu dimenzionalnost. Dimenzionalnost deskriptora temelji se na prikazu molekule. Tako razlikujemo:

0D – deskriptori koji se temelje na prikazu spoja kemijskom formulom, npr. broj i vrsta pojedinih atoma, molekulska masa

1D – jednodimenzionalni deskriptori koji opisuju fragmente, supstituente, ne zahtijevaju potpuno poznavanje strukture

2D – dvodimenzionalni deskriptori koji se temelje na topološkom prikazu molekule

3D – trodimenzionalni deskriptori izvedeni iz trodimenzionalnog prikaza molekule npr. van der Waalsov volumen, GETAWAY deskriptori

4D – četverodimenzionalni deskriptori koji se temelje na stereoelektronskom prikazu molekule, povezani su sa svojstvima koja potječu od interakcije molekule s molekulama koje je okružuju.[21]

Razni programski paketi (npr. Dragon[24]) mogu generirati i više tisuća deskriptora, između kojih je potrebno odabrati najinformativnije za određenu QSAR analizu, jer velik broj deskriptora ne znači i veću moć predviđanja razvijenog modela, već često mogu stvarati i šum.

Proces odabira deskriptora započinje izuzimanjem deskriptora koji za cijeli niz spojeva imaju konstantne (ili približno konstantne) vrijednosti ili vrijednost nula.

Zatim je potrebno provesti skaliranje jer različiti deskriptori imaju različit raspon vrijednosti te se na taj način sprečava dominacija deskriptora s velikim vrijednostima u modelu. U tu svrhu koristimo normiranje u rasponu (engl. *range-scaling*) i standardno normiranje (engl. *autoscaling*). Kod normiranja u rasponu imamo

$$X_{ik}^{n} = \frac{X_{ik} - minX_k}{maxX_k - minX_k} \tag{2.3.1}$$

X_{ik} i X_{ik}^{n} – nenormirana i normirana vrijednost deskriptora k (k=1,…,N) za spoj i (i = 1, … ,M)

$minX_k = min_i X_{ik}$ – minimalna vrijednost k-tog deskriptora

$maxX_k = max_i X_{ik}$ – maksimalna vrijednost k-tog deskriptora

Deskriptori normirani na ovaj način imaju minimalnu vrijednost 0 i maksimalnu vrijednost 1.

Standardno normiranje vršimo prema formuli:

$$X_{ik}^{n} = \frac{X_{ik} - \mu_k}{\sigma_k}, \tag{2.3.2}$$

pri čemu je

$$\mu_k = \frac{1}{M} \sum_{i=1}^{M} X_{ik} \tag{2.3.3}$$

i $$\sigma_k = \sqrt{\frac{\sum_{i=1}^{M}(X_{ik}-\mu_k)^2}{M-1}} \quad (2.3.4)$$

Nakon normiranja deskriptori s niskom varijancom (manjom od 0,001) se izuzimaju.

Sljedeći korak uključuje izuzimanje međusobno kolinearnih vektora. Najčešće se koristi korelacijska analiza parova (engl. *pairwise correlation analysis*). Metoda se temelji na računanju korelacijskog koeficijenta deskriptora s najvećom varijancom i svih ostalih deskriptora. Korelacijski koeficijent između dva deskriptora X_a i X_b se izračunava prema formuli:

$$R(X_a, X_b) = \frac{\sum_{i=1}^{M}(X_{ia}-\mu_a)(X_{ib}-\mu_b)}{\sqrt{\sum_{i=1}^{M}(X_{ia}-\mu_a)^2 \sum_{i=1}^{M}(X_{ib}-\mu_b)^2}} \quad (2.3.5)$$

μ_a – srednja vrijednost deskriptora X_a

μ_b – srednja vrijednost deskriptora X_b

Granična vrijednost koeficijenta korelacije ovisi i o skupu podataka i metodi. Metoda k-najbliži susjed (k-NN) dopušta relativno veliki broj deskriptora pa kao graničnu vrijednost koeficijenta korelacije možemo uzeti 0,90 – 0,95. Višestruka linearna regresija zahtijeva manji broj deskriptora pa je potrebno koristiti i manju graničnu vrijednost.[18]

Redukcija skupa deskriptora se vrši filter metodama (engl. *Filter methods*) i metodama "omotača" (engl. *Wrapper methods*). Filter metode su jednostavnije, brže i primjenjive na višedimenzionalne skupove podataka.

2.4. Podjela QSAR pristupa

Poput deskriptora, i QSAR metode možemo podijeliti na temelju prikaza strukture[25]:

1D–QSAR – povezuje aktivnost s globalnim molekulskim svojstvom npr. log*P*

2D–QSAR – temelji se na topološkom prikazu molekula

3D–QSAR – temelji se na trodimenzionalnom prikazu, povezuje aktivnost i nekovalentne molekulske interakcije

4D–QSAR – nadogradnja 3D–QSAR, dok u 3D–QSAR-u promatramo samo jedan konformer, u 4D-QSAR-u skup konformera

5D–QSAR – poput 4D–QSAR u kojem osim skupa liganada imamo i informaciju o receptoru i njegovoj fleksibilnosti (induciranom pristajanju)

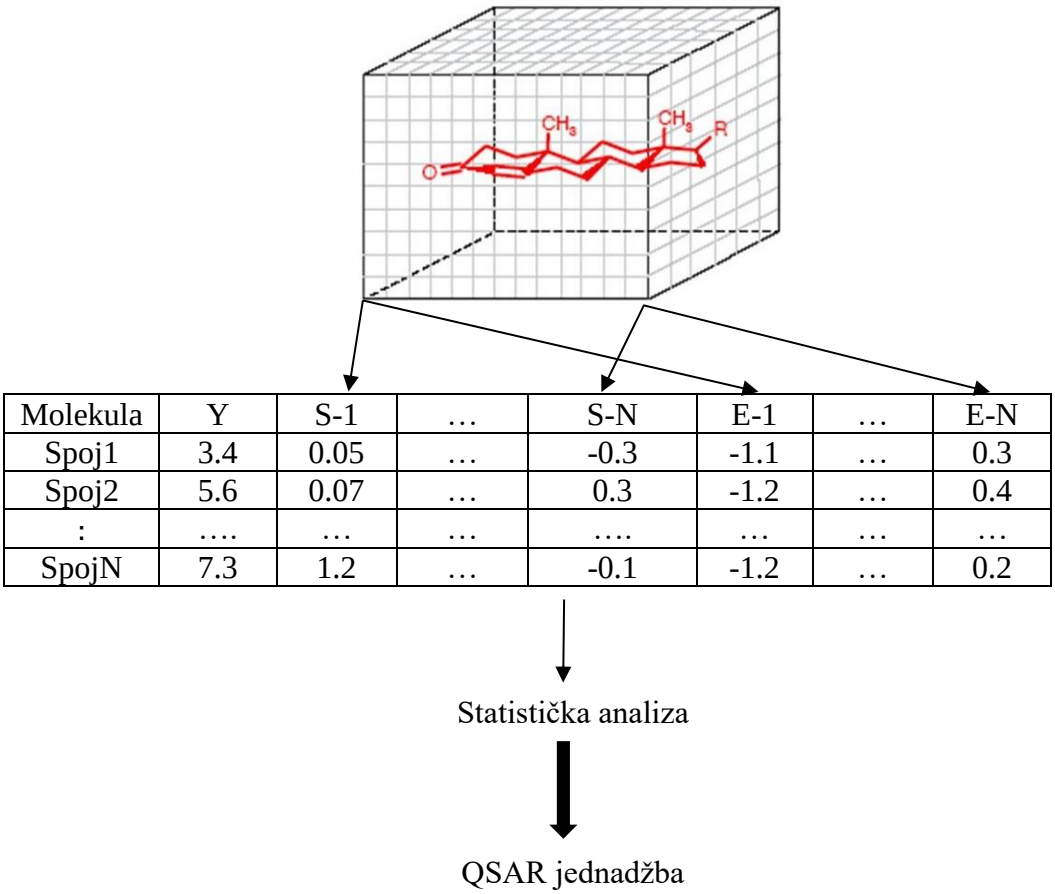
6D–QSAR – poput 5D-QSAR koji uključuje različite solvatacijske modele

Za razliku od 1D i 2D–QSAR pristupa, kod višedimenzionalnih pristupa obavezno koristimo trodimenzionalnu strukturu molekula. Prednost 1D i 2D metoda je u njihovoj jednostavnosti i brzini, a ograničenja proizlaze iz nedostatka parametara za opisivanje pojedinih interakcija (npr. ligand-receptor).

U širem smislu, 3D–QSAR obuhvaća sve metode koje se temelje na trodimenzionalnom prikazu molekule. Koriste se "zamrznuti" modeli, što znači da je sustav u ravnoteži i vremenski neovisan. Temeljne postavke 3D–QSAR metode su sljedeće:[26]

- na opaženu biološku aktivnost utječe ligand, a ne njegov metabolit ili derivat
- zanemaruje se dinamička priroda vezanja liganda koja podrazumijeva promjenu konformacije uslijed vezanja
- "kruta" geometrija veznog mjesta receptora
- gubitak translacijske i rotacijske entropije zbog vezanja jednako se razmatra kod svih spojeva
- broj veza koje mogu rotirati predstavljaju entropijski gubitak zbog zamrzavanja neterminalnih rotora
- svi ligandi vežu se na isto vezno mjesto na proteinu
- utjecaj otapala, temperature, difuzija, transport, pH, koncentracija soli i ostali faktori koji pridonose slobodnoj energiji vezanja ligand-receptor se ne razmatraju

Metoda komparativne analize molekulskog polja (engl. *Comparative Molecular Field Analysis*, CoMFA), kao prototip 3D-QSAR metoda, izračunava steričke i elektrostatske interakcije liganda s izabranim probama na način da se ligand smjesti u trodimenzionalnu rešetku, a probe u točke te rešetke. Razmak unutar rešetke je važno definirati jer određuje detaljnost opisa polja interakcija oko molekule. U svaku točku rešetke stavlja se proba, to može biti sp^3 hibridizirani ugljikov atom za istraživanje steričkih interakcija ili proton (H^+) za istraživanje elektrostatske interakcije ili neka druga dobro definirana kemijska vrsta. Za modeliranje van der Waalsovih interakcija koristi se Lennard-Jonesov potencijal, a elektrostatske interakcije se opisuju Coulombovim zakonom. Ove dvije funkcije pokazuju velike vrijednosti na atomskim pozicijama pa se uvodi granična vrijednost za steričke i elektrostatske vrijednosti ($\pm$ 30 kcal / mol). Dobivene vrijednosti zapisuju se u obliku 3D matrice, kao što je prikazano na slici 2.



Slika 2. Osnovna shema metode CoMFA (preuzeto i prilagođeno prema T. I. Oprea, u P. Bultinck, H. De Winter, W. Langenaeker, J. P. Tollenaere (ur.), *Computational Medicinal Chemistry for Drug Discovery*, Marcel Dekker, Inc., New York, 2004).

Primjenom statističkih metoda, najčešće metode parcijalnih projekcija najmanjih kvadrata (engl. *Partial Least Square*, PLS) dobiva se QSAR jednadžba koja povezuje aktivnost i interacijski potencijal koji vlada oko molekule. Rezultati se mogu vizualizirati konturnim dijagramima gdje se obično različitim bojama prikazuju različita područja (npr. crvena boja, preferiraju se elektronegativni supstituenti).

Nedostaci ove metode, poput problema orijentacije molekula, nedovoljno kvantizirane hidrofobnosti, upotrebe graničnih vrijednosti i sl., nastojali su se prevladati u metodi koja koristi Gaussove funkcije kako bi riješila problem graničnih vrijednosti (engl. *Comparative Molecular Similarity Indices Analysis*, CoMSIA).

2.5. Razvoj modela

U izgradnji QSAR modela koristimo različite statističke i kemometrijske metode. Važno je pronaći model koji s minimalnim brojem deskriptora daje najbolju korelaciju s aktivnošću, a uz to je robusan i prediktivan.

2.5.1. Linearni modeli

Linearni modeli predviđaju linearni odnos aktivnosti i odabranih molekulskih deskriptora. Često se koriste u QSAR analizi zbog jednostavnosti, jasne interpretacije i zadovoljavajuće točnosti ukoliko se radi o manjem skupu podataka za koji su molekulski deskriptori pažljivo odabrani.

- Višestruka linearna regresija

Višestruka linearna regresija (engl. *Multiple linear regression*, MLR) je proširenje jednostavne linearne regresije (engl. *simple linear regression*, SLR), gdje umjesto jedne imamo više nezavisnih varijabli.
Opća jednadžba za MLR model glasi:[27]

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_n X_n \qquad (2.5.1.1.)$$

β_0 – konstanta modela (odsječak)
Y – zavisna varijabla
$X_1, X_2, \ldots, X_n$ – nezavisne varijable koje predstavljaju molekulske deskriptore s odgovarajućim koeficijentima $\beta_1, \beta_2, \ldots, \beta_n$

Iznos koeficijenta opisuje utjecaj molekulskog deskriptora kao nezavisne varijable u opisivanju aktivnosti ukoliko su korištene normirane vrijednosti nezavisnih varijabli.[28] Potrebno je izbjeći kolinearnost deskriptora, a preporučeni omjer broja spojeva i broja deskriptora trebao bi iznositi 5 : 1 kako bi se ovom metodom razvili zadovoljavajući QSAR modeli.

- Metoda parcijalnih projekcija najmanjih kvadrata

Metoda parcijalnih projekcija najmanjih kvadrata (engl. *partial least square*, PLS) je iterativni regresijski postupak koji koristimo kada imamo velik broj kolinearnih deskriptora ili u slučajevima kada je broj deskriptora veći od broja podataka.

Deskriptori (X_1, …, X_m) transformiraju se u međusobno ortogonalne „latentne varijable, LVs", vrijednosti t_1, …, t_n koje su linearna kombinacija nezavisnih varijabli. Jednadžba za PLS glasi:[27]

$$Y = a_1t_1 + a_2t_2 + \ldots + a_nt_n \quad (2.5.1.2)$$

pri čemu su LVs:

$$t_1 = b_{11}X_1 + b_{12}X_2 + \ldots + b_{1m}X_m \quad (2.5.1.3)$$

$$t_2 = b_{21}X_1 + b_{22}X_2 + \ldots + b_{2m}X_m \quad (2.5.1.4)$$

.

.

$$t_n = b_{n1}X_1 + b_{n2}X_2 + \ldots + b_{nm}X_m \quad (2.5.1.5)$$

Ova metoda je jedna od najčešće korištenih u izgradnji 3D-QSAR modela.

- Analiza glavnih komponenata

Analiza glavnih komponenata (eng. *principal component analysis*, PCA) je statistička tehnika (slična PLS-u) koja linearno transformira nezavisne varijable u manji skup nekoreliranih varijabli koje nazivamo glavnim komponentama (engl. *Prinicipal Components*, PC). Glavne komponente se grade na način da maksimalno opisuju varijanciju matrice nezavisnih varijabli, pri čemu prva glavna komponenta opisuje najveći udio varijancije, a svaka sljedeća manji, te sadrže glavninu informacija sadržanih u nezavisnim varijablama.[29] Glavne komponente su međusobno ortogonalne. Takva metoda redukcije varijabli poznata je i kao "parsimonious summarization".

Regresija glavnih sastavnica (engl. *principal component regression*, PCR) je metoda linearne regresije u kojoj se kao nezavisne varijable koriste glavne sastavnice (PC).

- Free–Wilson analiza

Metoda koju su 1964. Godine razvili Free i Wilson temelji se na pretpostavci da supstituenti na određenim položajima u molekuli jednako pridonose opaženoj aktivnosti u svim spojevima kongenerične serije. Matematički metodu možemo opisati sljedećim izrazom:

$$Y = \mu + \sum_{ij} \alpha_{ij} R_{ij} \quad (2.5.1.6)$$

μ – konstantna koja pokazuje aktivnost nesupstituiranog spoja

α_{ij} – doprinos supstituenta R_i na mjestu j

R_{ij} - može imati vrijednost 1 ili 0, ovisno o tome da li je supstituent R_i prisutan ili odsutan na mjestu j

Pretpostavke na kojima se temelji ova metoda su:

1) stalan doprinos pojedinog supstituenta aktivnosti

2) doprinosi supstituenata su aditivni

3) između supstituenata ne postoje interakcije

Glavni nedostatak ove metode je nemogućnost predviđanja aktivnosti za spojeve koji sadrže supstituente koji nisu bili uključeni u analizu. Fujita i Ban su 1971. godine modificirali ovu metodu uvodeći referentni spoj (može biti bilo koji spoj) za koji su sve binarne vrijednosti jednake nuli. Doprinos pojedinog supstituenta dobiva se kao razlika doprinosa tog supstituenta i referentnog spoja.[27] Prednosti ovih metoda su u tome da ne zahtijevaju izračune deskriptora.

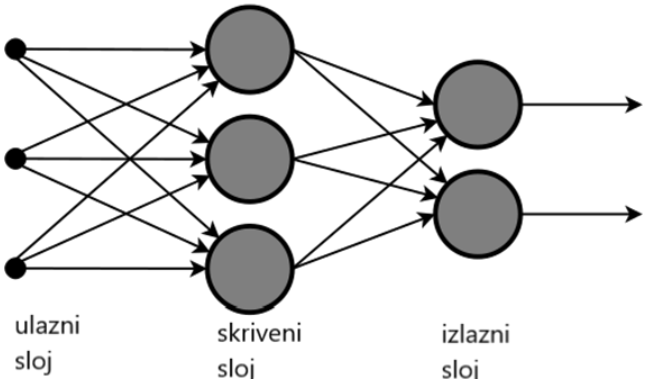
2.5.2. Nelinearni modeli

U QSAR analizi ponekad postoji nelinearan odnos između molekulskih parametara i aktivnosti. U tom slučaju koristimo nelinearne metode koje slove kao točne, ali upitne interpretabilnosti s obzirom da se temelje na iteracijskom podešavanju (engl. *fitting*) te postoji mogućnost prepodešavanja (engl. *overfitting*).

- Umjetne neuronske mreže

Metoda umjetnih neuronskih mreža (engl. *Artificial neural networks,* ANN) temelji se na simulaciji rada neurona u ljudskom mozgu. Svaka umjetna neuronska mreža sastoji se od nekoliko slojeva: ulaznog, jednog ili više slojeva koji se nazivaju skrivenim, te izlaznog sloja.[25] Jedinice od kojih se sastoje slojevi nazivaju se umjetnim neuronima i svaki neuron iz jednog sloja povezan je sa svim neuronima sljedećeg sloja. Neuroni ulaznog sloja primaju signale (vrijednost deskriptora pomnožena s težinskim faktorom), zatim ih prosljeđuju u skrivene

slojeve gdje ih procesira nelinearna funkcija, a rezultat se predaje neuronima u izlaznom sloju gdje se interpretiraju.



Slika 3. Shematski prikaz umjetne neuronske mreže. Krugovi predstavljaju umjetne neurone u pojedinim slojevima (prema M. K. Macan, M. Petrović, *Analitika okoliša*, Hinus, Zagreb, 2013 str. 371.).

- k-najbliži susjed

Metoda k-najbliži susjed (engl. *k-Nearest Neighbor*, k-NN) klasificira spojeve u nekoliko k grupa. K je unaprijed zadana veličina, a zasniva se na mjerenju udaljenosti (Euklidove ili manhattan udaljenosti) između molekulskih deskriptora. Pouzdanost metode opada ukoliko k nije dobro zadan.

2.5.3. Vrijednosti koje značajno odstupaju (engl. Outliers)

Razlikujemo prave i tzv. prividne *outliere*, odnosno vrijednosti koje značajno odstupaju.[31] Pravi nastaju zbog eksperimentalnih pogrešaka mjerenja aktivnosti ili pogrešno izračunatih deskriptora. Prividni nastaju zbog pogrešno modelirane linearnosti te mogu pomoći u interpretaciji modela.

U QSAR istraživanjima *outlierima* se pristupa na dva načina. Različitim tehnikama (statističke, algoritam potpornih vektora) nastoje se *outlieri* identificirati i eliminirati, odnosno ukloniti te spojeve iz modela. Nedostatak ovog pristupa je u tome da *outlieri* imaju utjecaj na model korišten u njihovoj identifikaciji. Drugi pristup je korištenje robusnih metoda u izgradnji

modela na koji *outlieri* nemaju toliki utjecaj. Pri tom je vrlo važno adekvatno i s čim većom sigurnošću identificirati prave *outliere*.

2.6. Validacija modela

Da bi QSAR model bio prihvaćen kao pouzdan treba zadovoljiti validacijski postupak. Razlikujemo validacijske postupke u kojima koristimo iste spojeve koje smo koristili za učenje modela, tzv. interna validacija (engl. *internal validation*), i validacijske postupke u kojima koristimo spojeve iz skupa za vrednovanje modela, tzv. vanjska validacija (engl. *external validation*).

2.6.1. Interna validacija

Najčešće korištena metoda interne validacije je unakrsna validacija.

Unakrsna validacija je metoda koja se temelji na izuzeću jednog spoja, (engl. *Leave-One-Out Cross-Validation,* LOO) ili više njih (engl. *Leave-Group-Out Cross-Validation,* LGO) iz skupa za učenje modela. Spojevi koji nisu izuzeti iz skupa služe za razvoj modela kojim se zatim predviđa aktivnost izuzetog spoja. Postupak se ponavlja tako da se svaki spoj izuzme i za njega predvidi aktivnost. Kriterij predikcijske sposobnosti je korelacijski koeficijent q^2 koji se izračunava na sljedeći način:

$$q^2 = 1 - \frac{\sum\left(Y_{obs(train)} - Y_{pred(train)}\right)^2}{\sum\left(Y_{obs(train)} - \bar{Y}_{(train)}\right)^2} \qquad (2.6.1.1)$$

gdje su:

Y_{obs} – eksperimentalno opažena vrijednost aktivnosti

Y_{pred} – predviđena vrijednost aktivnosti

$\bar{Y}$ - srednja vrijednost aktivnosti

Vrijednosti q^2 su manje od 1, a mogu biti i negativne. Negativne vrijednosti q^2 ukazuju da je model opisao potpuno pogrešnu korelaciju. Da bi model bio prihvatljiv, poželjno je da je $q^2 >$ 0,4.[18]

Standardna devijacija predviđanja iznosi:

$$\text{SDER} = \sqrt{\frac{\sum\left(Y_{obs(train)} - Y_{pred(train)}\right)^2}{n}} \qquad (2.6.1.2)$$

pri čemu je *n* broj spojeva.[27]

Velika vrijednosti q^2 nije dovoljan kriterij da bi se smatralo da model ima dobru moć predviđanja. K. Roy je razvio novi parameter r_m^2:[32]

$$\bar{r}_m^2 = \frac{(r_m^2 + r_m'^2)}{2} \quad (2.6.1.3)$$

$$\Delta r_m^2 = |r_m^2 - r_m'^2| \quad (2.6.1.4)$$

$$r_m^2 = r^2 \times \left(1 - \sqrt{r^2 - r_0^2}\right) \quad (2.6.1.5)$$

$$r_m'^2 = r^2 \times \left(1 - \sqrt{r^2 - r_0'^2}\right) \quad (2.6.1.6)$$

r^2 – kvadrat korelacijskog koeficijenta između opažene i predviđene vrijednosti

r_0^2 – kvadrat korelacijskog koeficijenta između opažene i predviđene vrijednosti kada regresijski pravac prolazi kroz ishodište

$r_0'^2$ – kvadrat korelacijskog koeficijenta između predviđene i opažene vrijednosti kada regresijski pravac prolazi kroz ishodište

Preporuča se da je $\bar{r}_m^2 > 0{,}5$ i $\Delta r_m^2 < 0{,}2$.

2.6.2. Eksterna validacija

U eksternoj validaciji (engl. *external validation*) koristimo spojeve koji nisu korišteni pri izgradnji modela i koji pripadaju skupu za vrednovanje (testni skup). Predviđaju se njihove aktivnosti kako bi se testirala predikcijska moć modela. Jedan od najčešćih parametara koji se izračunavaju je R^2 čija vrijednost ne bi trebala biti manja 0,6.

$$R^2 = 1 - \frac{\sum(Y_{obs(test)} - Y_{pred(test)})^2}{\sum(Y_{obs(test)} - \bar{Y}_{(train)})^2} \quad (2.6.2.1)$$

$Y_{obs(test)}$ – opažena aktivnost u skupu za vrednovanje

$Y_{pred(test)}$ – predviđena aktivnost u skupu za vrednovanje

$\overline{Y}_{train}$ – srednja vrijednost aktivnosti u skupu za učenje

R^2 nije dovoljan kriterij za validaciju modela. Potrebno je izračunati korelacijske koeficijente R_0^2 (predviđena vs. opažena aktivnost) i R'^2_0 (opažena vs. predviđena aktivnost) za regresiju kroz ishodište. Nagibi iznose k i k'. Model je prihvatljiv ukoliko vrijedi:[32]

$$\frac{(R^2-R_0^2)}{R^2} < 0{,}1 \qquad \text{i} \qquad 0{,}9 \leq k \leq 1{,}1 \tag{2.6.2.2}$$

ili $$\frac{(R^2-R\prime_0^2)}{R^2} < 0{,}1 \qquad \text{i} \qquad 0{,}9 \leq k' \leq 1{,}1 \tag{2.6.2.3.}$$

$$|R_0^2 - R'^2_0| < 0{,}3 \tag{2.6.2.4}$$

Također, postoje i kriteriji za eksternu validaciju predstavljeni Q^2 funkcijom koja se temelji na korelacijskom koeficijentu q^2 interne unakrsne LOO validacije. Shi je 2001. predložio prvu takvu funkciju:[27]

$$Q_{F1}^2 = 1 - \frac{\sum_{i=1}^{n_{EXT}}\left(Y_{obs(test)} - Y_{pred(test)}\right)^2}{\sum_{i=1}^{n_{EXT}}\left(Y_{obs(test)} - \bar{Y}_{train}\right)^2} \tag{2.6.2.5}$$

Schüürmann je 2008. predložio poboljšanu verziju koja ne zahtijeva poznavanje skupa za učenje:[27]

$$Q_{F2}^2 = 1 - \frac{\sum_{i=1}^{n_{EXT}}\left(Y_{obs(test)} - Y_{pred(test)}\right)^2}{\sum_{i=1}^{n_{EXT}}\left(Y_{obs(test)} - \bar{Y}_{test}\right)^2} \tag{2.6.2.6}$$

Nedostatak ovih funkcija je u tome što ovise o raspodijeli podataka. Consonni je taj nedostatak pokušao ukloniti u novoj funkciji, Q_{F3}^2:[27]

$$Q_{F3}^2 = 1 - \frac{\sum_{i=1}^{n_{EXT}}\left(Y_{obs(test)} - Y_{pred(test)}\right)^2 \Big/ n_{EXT}}{\sum_{i=1}^{n_{train}}\left(Y_{obs(train)} - \bar{Y}_{train}\right)^2 \Big/ n_{train}} \tag{2.6.2.7}$$

Parametar r_m^2 koji se koristio u internoj validaciji za određivanje predikcijske sposobnosti može se koristiti i u vanjskoj validaciji.

Preciznost i točnost postavljenog modela mjeri se CCC (engl. *concordance correlation coefficient*) parametrom.

$$\text{CCC} = \frac{2\sum_{i=1}^{n}\left(Y_{obs(test)}-\bar{Y}_{obs(test)}\right)\left(Y_{pred(test)}-\bar{Y}_{pred(test)}\right)}{\sum_{i=1}^{n}\left(Y_{obs(test)}-\bar{Y}_{obs(test)}\right)^2+\sum_{i=1}^{n}\left(Y_{pred(test)}-\bar{Y}_{pred(test)}\right)^2+n\left(Y_{pred(test)}-\bar{Y}_{pred(test)}\right)} \quad (2.6.2.8)$$

Idealna vrijednost CCC je blizu 1, a prihvatljive su one veće od 0,85.[27]

Kako bi razvijeni model mogao poslužiti za predviđanje aktivnosti za nove spojeve potrebno je odrediti domenu primjenljivosti. Domena primjenljivosti se definira kao fizičko-kemijski, strukturni ili biološki prostor, znanje ili informacija na kojima je razvijen skup za učenje modela i na kojima se vrši predviđanje za nove spojeve. Opisuje se deskriptorima modela.[18] Za utvrđivanje domene primjenljivosti koriste se različite metode temeljene na deskriptorima, na odgovoru, geometrijske metode, metode temeljene na distribuciji gustoće vjerojatnosti.

§ 3. LITERATURNI IZVORI

1. Tropsha, u K. M. Merz, D. Ringe, C. H. Reynolds (ur.), *Drug design: structure- and ligand- based approaches*, Cambridge University Press, New York, 2010, str. 152.
2. F. A. Cros, *Action de l'alcool amylique sur l'organisme,* Doktorski rad, University of Strasbourg**,** 1863.
3. A. Crum Brown, T. R. Fraser, *J Anat Physiol.* **2 (2)** (1868), 224 - 242.
4. B. W. Richardson, *Medical Times and Gazette* **2** (1869) 703 - 706.
5. C. Richet, *Comptes Rendus Société Biologie* **54** (1893) 775 - 776
6. H. Meyer, N. Schmiedebergs, *Arch. Exp. Pathol. Pharmakol*. **42** (1899) 109 – 118.
7. F. Baum, N. Schmiedebergs, *Arch. Exp. Pathol. Pharmakol*. **42** (1899) 119 – 137.
8. E. Overton, *Vierteljahrsschr. Naturforsch. Ces*. Zurich **44** (1899) 88 – 135.
9. C. Hansch, T. Fujita, *J. Am. Chem. Soc.* **86 (8)** (1964) 1616 – 1626.
10. Tropsha et al., *Chem. Soc. Rev*. **49** (2020) 3525 - 3564.
11. K. Z. Myint, X. Q. Xie, *Int. J. Mol. Sci.* **11** (2010) 3846 - 3866.
12. https://www.oecd.org (datum pristupa 5. 9. 2021.)
13. http://pubchem.ncbi.nlm.nih.gov/ (datum pristupa 5. 9. 2021.)
14. https://www.ebi.ac.uk/chembl/ (datum pristupa 5. 9. 2021.)
15. http://zinc.docking.org (datum pristupa 5. 9. 2021.)
16. http://www.daylight.com/products/wdi.html (datum pristupa 5. 9. 2021.)
17. http://cas.org/expertise/cascontent/registry/index.html (datum pristupa 5. 9. 2021.)
18. A. Tropsha, A. Golbraikh, u J .L. Faulon, A. Bender (ur.), *Handbook of chemoinformatics algorithms*, Chapman & Hall/CRC, Boca Raton, 2010
19. http://www.chemcomp.com/ (datum pristupa 5. 9. 2021.)
20. http://www.chemaxon.com/ (datum pristupa 5. 9. 2021.)
21. R. Todeschini, V. Consonni, *Handbook of Molecular Descriptors*,Wiley-VCH, Weinheim, 2000
22. K. Roy, S. Kar, R. N. Das, *Understanding the Basics of QSAR for Applications in Pharmaceutical Sciences and Risk Assessment*, Academic Press, London, 2015, str. 54.
23. H. Wiener, *J. Am. Chem. Soc.*, **69** (1947) 2636
24. http://www.vcclab.org (datum pristupa 5. 9. 2021.)

25. J. Verma, V. M. Khedkar, E. C. Coutinho, *Current Topics in Medicinal Chemistry*, **10** (2010) 95 - 115.
26. T. I. Oprea, u P. Bultinck, H. De Winter, W. Langenaeker, J. P. Tollenaere (ur.), *Computational Medicinal Chemistry for Drug Discovery*, Marcel Dekker, Inc., New York, 2004
27. S. Dastmalchi, M. Hamzeh-Mivehroud, B. Sokouti, *Quantitative structure-activity relationship: a practical approach*, CRC Press Taylor & Francis Group, Boca Raton, 2018
28. S. Yousefinejad, B. Hemmateenejad, *Chemometrics and Intelligent Laboratory Systems* **149** (2015) 177 - 204.
29. G. H. Dunteman, *Principal Components Analysis*, Sage Publications, Newbury Park, 1989
30. M. K. Macan, M. Petrović, *Analitika okoliša*, Hinus, Zagreb, 2013 str. 371.
31. T. Scior et al., *Current Medicinal Chemistry* **16** (2009), 4297 – 4313
32. K. Roy et al., *Journal of Computational Chemistry*, **34** (2013) 1071 - 1082