Univerza v Ljubljani

Fakulteta za elektrotehniko

UNIVERZITETNI ŠTUDIJ

DIPLOMSKO DELO

Napovedovanje biorazgradljivosti z regresijskimi drevesi

Bernard Ženko

Mentor: prof. dr. Nikola Pavešić

Somentor: doc. dr. Sašo Džeroski

Ljubljana, 2000

Zahvala

Zahvaljujem se mentorju prof. dr. Nikoli Pavešiću in somentorju doc. dr. Sašu Džeroskemu za strokovno pomoč ter vsestransko podporo pri delu. Zahvaljujem se dr. Jiřiju Damborskyju za izdelavo PLS modelov, komentar nekaterih zgrajenih modelov ter odgovore na množico vprašanj. Zahvaljujem se prof. dr. Borisu Komparetu za komentar modelov množic P1 in P2 ter pomoč pri slovenjenju nekaterih izrazov. Zahvaljujem se Andreju Koblerju za izdelavo modelov množic P1 in P2 s polno različico sistema Cubist. Zahvaljujem se Ernestu Ženku za jezikovni pregled besedila ter številne nasvete. Nenazadnje se zahvaljujem staršema, ki sta mi vedno stala ob strani in me podpirala pri študiju.

Kazalo

Povzetek

Biorazgradljivost spojine je ena pomembnejših lastnosti, ki jih moramo upoštevati pri ocenjevanju varnosti njene uporabe. Ker bi bilo eksperimentalno določanje biorazgradljivosti množice različnih kemikalij težko izvedljivo, se problema lotimo z modeliranjem količinskih odnosov med strukturo in določeno lastnostjo spojine (Quantitative Structure-Activity Relationships – QSAR). Za vzorčno množico spojin eksperimentalno določimo njihovo biorazgradljivost ter nato zgradimo model, ki zadovoljivo opisuje tako proučene kot neproučene spojine. Model lahko zgradimo s klasično metodo linearne regresije ali z metodami strojnega učenja; običajno so to metode za gradnjo regresijskih dreves. Ta dva tipa modelov sta bila primerjana v tem diplomskem delu. Za več različnih množic podatkov smo zgradili modele z orodjema za gradnjo regresijskih dreves Cubist in RETIS. Vsi zgrajeni modeli so bili prečno preverjeni. Najboljše med njimi sta pregledala strokovnjaka s področja biorazgradljivosti. Za majhne množice strukturno sorodnih spojin so modeli zgrajeni z linearno regresijo običajno bolj točni kot modeli z regresijskimi drevesi, čeprav imajo slednji včasih primerljivo točnost in so lahko razumljivi. Za večje množice strukturno različnih spojin so modeli z regresijskimi drevesi bolj točni kot linearni regresijski modeli.

Abstract

The biodegradability of a chemical compound must be considered when estimating the safety of its use for the environment. Because of the huge number of various chemicals, it is practically impossible to experimentally determine biodegradability for all or at least for a significant number of them. A possible solution to this problem is quantitative structure-activity relationships (QSAR) analysis. We experimentally test a representative group of chemicals and then build a model which satisfactorily describes the tested as well as unknown chemicals. The model can be built with classical linear regression methods or with machine learning methods, tipically regression tree building methods. A comparison between these two types of models is made in this work. For several data sets, models with Cubist and RETIS regression tree building systems are built. All models are cross-validated and the best ones are inspected by domain experts. For small sets of compounds with similar structure, models built with linear regression are usually more accurate than models built with regression trees, although the latter sometimes have comparable accuracy and are easily understandable. For large sets of structuraly diverse compounds, regression trees yield more accurate models than linear regression.

Ključne besede

strojno učenje	(machine learning)
odločitveno drevo	(decision tree)
regresijsko drevo	(regression tree)
biorazgradljivost	(biodegradability)
QSAR modeliranje	(QSAR modeling)

1. Uvod

Vpliv človeka na okolje v zadnjem desetletju priteguje vse več pozornosti tako strokovne kot laične javnosti. Vzroke za povečanje ekološke ozaveščenosti lahko iščemo tudi v množični in nekontrolirani uporabi raznovrstnih kemikalij v preteklosti. Mnogi kvarni učinki takega ravnanja (tanjšanje ozonskega plašča, učinki tople grede ipd.) so se pokazali ravno v iztekajočem se desetletju. Zaradi tega so raziskave lastnosti kemičnih spojin pridobile na pomembnosti.

Z okoljevarstvenega vidika je biorazgradljivost spojine ena njenih pomembnejših lastnosti; njeno poznavanje je predpogoj za oceno varnosti njene uporabe. Pomembno je tudi poznavanje mehanizmov biorazgradljivosti, saj bi s poznavanjem le teh lahko izdelali okolju manj škodljive kemikalije. Ker bi bilo eksperimentalno določanje biorazgradljivosti tako velike množice različnih spojin težko izvedljivo, se problema lotimo z modeliranjem količinskih odnosov med strukturo in določeno lastnostjo spojine (Quantitative Structure-Activity Relationships – QSAR). Postopek je naslednji. Za (majhno) množico spojin, katerih strukturo poznamo, eksperimentalno določimo njihovo biorazgradljivost. Na osnovi teh podatkov zgradimo model, ki opisuje biorazgradljivost spojine v odvisnosti od njenih strukturnih lastnosti. Dobljeni model lahko uporabimo za napovedovanje biorazgradljivosti spojin, katerih biorazgradljivost ni znana ali za pojasnjevanje povezav med strukturo spojine in njeno biorazgradljivostjo.

Običajno se za gradnjo QSAR modelov uporablja linearna regresija (glej npr. [1]–[7]), ki je opisana v poglavju 2. Alternativni pristop je uporaba metod strojnega učenja (glej [8]). Namesto modela v obliki ene linearne enačbe lahko tu model predstavimo npr. z regresijskim drevesom, ki v vsakem listu vsebuje linearni model. Regresijska drevesa in orodji, uporabljeni za njihovo gradnjo v tem diplomskem delu, so opisani v poglavju 3.

Cilj pričujočega diplomskega dela je bil primerjati modele, zgrajene s klasično metodo linearne regresije in z metodo regresijskih dreves. V ta namen je bilo uporabljenih več različnih množic podatkov ter orodji za gradnjo regresijskih dreves Cubist in RETIS. Modeli, zgrajeni z obema orodjema, so bili primerjani z modeli zgrajenimi s klasično metodo PLS. Opravljeni poskusi so opisani v poglavju 4.

Iz poskusov izhajajoči zaključki, kakor tudi nakaj smernic za nadaljnje delo so podani v poglavju 5.

2. Napovedovanje biorazgradljivosti

Proizvodnja raznovrstnih kemikalij v svetu nenehno narašča. Znanim vrstam se vsako leto pridruži množica novih, katerih lastnosti so še neraziskane. Zaradi pomanjkanja časa, denarja in človeških virov ni mogoče primerno proučiti njihovih lastnosti.

Prisotnost določene kemikalije je zaželjena le dokler ne izpolni svoje naloge, potem postane odpadek, onesnaževalec okolja [8]. Kemikalijo pojmujemo kot onesnaževalca tudi, če se znajde na napačnem mestu. Onesnaževalcev se lahko znebimo z uničevanjem, vendar to veliko stane. Druga, sicer zelo poceni, rešitev bi bilo razredčevanje kemikalij v okolju do neškodljivih koncentracij, kar pa z ekološkega stališča ni sprejemljivo. Nasprotno je razgradnja kemikalij v okolju na neškodljive snovi ekološko sprejemljiva ter tudi poceni. Poznamo različne načine razgradnje kemikalij: fizikalne (erozija, fotoliza, itd.), kemijske (hidroliza, oksidacija, itd.) in biološke (bioliza). Običajno nastopajo vse tri vrste razgradnje skupaj ter so med seboj prepletene, zaradi česar je razgradnja zelo zapleten proces. Pri oceni varnosti določene spojine za živa bitja in okolje moramo upoštevati možnost njene razgradnje v okolju. Za razvoj tehnologij čiščenja kontaminiranih področij je poleg tega zaželjeno tudi poznavanje mehanizmov razgradnje.

V našem primeru se bomo osredotočili na biorazgradljivost kemikalij. Zaradi zgoraj navedenih težav je nemogoče eksperimentalno proučiti biorazgradnjo vseh kemikalij. Pomagamo si lahko tako, da to naredimo za majhno skupino kemikalij ter na osnovi tako pridobljenega znanja zgradimo model, ki zadovoljivo opisuje oz. napoveduje tudi biorazgradljivost neproučenih kemikalij.

Na biorazgradljivost spojine gotovo vpliva njena struktura oz. z njo povezane lastnosti. V splošnem so spojine s strukturami, ki nastopajo tudi v naravnih spojinah, bolje razgradljive od spojin s strukturami, ki so v naravi neznane. Povezave med biorazgra-

dljivostjo spojine (njeno aktivnostjo) ter njeno strukturo, lahko raziščemo s t.i. QSAR (Quantitative Structure-Activity Relationships) modeliranjem.

2.1 QSAR modeliranje

QSAR modeliranje je postopek, s pomočjo katerega lahko raziščemo povezave med strukturo organskih spojin ter njihovo aktivnostjo. V našem primeru je proučevana aktivnost biorazgradljivost (Quantitative Structure-Biodegradability Relationships – QSBR), lahko pa je tudi kaj drugega: toksičnost (Quantitative Structure-Toxity Relationships – QSTR), stabilnost spojine (Quantitative Structure-Stability Relationships – QSSR), učinkovitost zdravila, itd. Strukturo proučevanih spojin opišemo z množico strukturnih, fizikalno-kemijskih ali kvantno-kemijskih lastnosti, t.i. *molekulskih deskriptorjev*, za katere želimo, da bi bili povezani z mehanizmom proučevane aktivnosti.

Pri QSAR modeliranju ločimo dve stopnji. Prva je razvoj modela na osnovi spojin, katerih aktivnost je bila eksperimentalno določena. Druga stopnja je uporaba modela, dobljenega v prvi stopnji za napovedovanje aktivnosti neznanih spojin. Neznane spojine so tiste, katerih aktivnost še ni bila eksperimentalno določena, znana pa je njihova struktura. Pri razvoju modela naj bi po [3] izvedli naslednje korake:

1. **Določitev razredov podobnih spojin,** za katere želimo zgraditi model. V začetno množico spojin poskušamo vključiti čimveč strukturno različnih pripadnikov izbranih razredov.

2. **Izbor molekulskih deskriptorjev obravnavanih spojin.** Poskušamo zbrati, izmeriti ali izračunati čim večje število molekulskih deskriptorjev za vsako spojino, izbrano v prvem koraku. Poseben poudarek damo deskriptorjem, ki so povezani z mehanizmom proučevane aktivnosti. Če teh ne poznamo, izberemo deskriptorje prostorske ureditve atomov v spojini, hidrofobičnosti in elektronske deskriptorje.

3. **Izbor učne in testne množice.** Množico vseh spojin z znanimi vrednostmi uporabljenih deskriptorjev razdelimo na učno in testno množico. Vsaka posebej naj pokriva celotno področje uporabe modela.

4. **Eksperimentalna določitev aktivnosti.** Za vse pripadnike učne in testne množice eksperimentalno določimo proučevano aktivnost. Vsi uporabljeni postopki morajo biti dosledni in ponovljivi, saj iz slabih podatkov ne moremo zgraditi dobrega modela.

5. **Gradnja QSAR modela.** S pomočjo matematične metode (običajno je to metoda PLS – glej razdelek 2.2) ali kako drugače (n.pr. s strojnim učenjem) poiščemo opis povezav med molekulskimi deskriptorji in aktivnostjo.

6. **Preverjanje veljavnosti zgrajenega modela.** Če bomo model uporabljali za napovedovanje, je ta korak še posebej pomemben. Obširneje je opisan v razdelku 2.3.

Vseh korakov včasih ni mogoče izvesti. Pri gradnji modela na osnovi objavljenih podatkov (modeli zgrajeni v okviru tega diplomskega dela) ne moremo izvesti korakov 1–4.

2.2 Metoda delnih najmanjših kvadratov (Partial Least Squares Method)

Klasični QSAR modeli imajo obliko linearne enačbe (linearni regresijski modeli), ki jo lahko zapišemo kot:

$$y = b_1 x_1 + b_2 x_2 + \ldots + b_m x_m. \tag{2.1}$$

Spremenljivka y predstavlja aktivnost, ki jo model opisuje (odvisna spremenljivka); x_j so deskriptorji, od katerih je aktivnost odvisna (neodvisne spremenljivke). Vrednosti b_j so parametri modela (regresijski koeficienti) in jih lahko za podano učno množico določimo na različne načine; največkrat se uporablja metoda delnih najmanjših kvadratov (Partial Least Squares Method – PLS). Opazimo lahko, da enačba 2.1 ne vsebuje konstantnega člena b_0. Razlog je v tem, da za odvisne in neodvisne spremenljivke običajno ne uporabljamo »surovih« vrednosti, temveč jih predhodno *centriramo* in *skaliramo*. Centriranje pomeni, da vsaki spremenljivki odštejemo njeno povprečje nad učno množico. Skaliranje največkrat izvedemo tako, da posamezno spremenljivko delimo z njeno standardno deviacijo v učni množici.

Preden si na kratko ogledamo metodo PLS, omenimo še reševanje sistema linearnih enačb z metodo najmanjših kvadratov, analizo glavnih komponent (Principal Component Analysis – PCA) in regresijo glavnih komponent (Principal Component Regression – PCR) na katerih temelji metoda PLS.

2.2.1 Reševanje sistema linearnih enačb

Parametre modela b_j lahko določimo z reševanjem sistema linearnih enačb. Za vsak primer iz učne množice napišemo po eno enačbo 2.1. Za n učnih primerov in m deskriptorjev dobimo tako sistem enačb:

$$\begin{aligned} y_1 &= b_{11}x_{11} + b_{12}x_{12} + \ldots + b_{1m}x_{1m} \\ y_2 &= b_{21}x_{21} + b_{22}x_{22} + \ldots + b_{2m}x_{2m} \\ &\vdots \\ y_n &= b_{n1}x_{n1} + b_{n2}x_{n2} + \ldots + b_{nm}x_{nm}, \end{aligned} \tag{2.2}$$

oziroma v matričnem zapisu

$$\mathbf{y} = \mathbf{Xb}. \tag{2.3}$$

Glede na število primerov (n) in število deskriptorjev (m) ločimo med tremi možnostmi:

1. $m > n$. Število neznank je večje od števila enačb. Obstaja neskončno rešitev za $\mathbf{b}$. Do ene same rešitve pridemo, če nekaj deskriptorjev zbrišemo, da dobimo enega od spodnjih dveh primerov.

2. $m = n$. Število neznank je enako številu enačb. V praksi redko naletimo na tak primer. Rešitev za $\mathbf{b}$ je enolično določena, če ima matrika $\mathbf{X}$ poln rang.

3. $m < n$. Število neznank je manjše od števila enačb. Sistem je predoločen in nima eksaktne rešitve. Poiščemo lahko le približek z metodo najmanjših kvadratov:

$$\hat{\mathbf{b}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}. \tag{2.4}$$

Matrika $\mathbf{X}'$ je transponirana matrika $\mathbf{X}$. Izpeljavo si lahko ogledamo v [9]. Šibka točka enačbe 2.4 je v tem, da zahteva izračun matriki $\mathbf{X}'\mathbf{X}$ inverzne matrike. Če ta ne obstaja,

oz. je slabo pogojena, si z enačbo 2.4 ne moremo pomagati. To je pri QSAR modeliranju pogost pojav, saj pogosto uporabljamo deskriptorje, ki so medsebojno močno korelirani.

Metodo najmanjših kvadratov lahko preprosto razširimo na več odvisnih spremenljivk (model opisuje več lastnosti hkrati). Edina sprememba v zgornjih enačbah je ta, da vektorja $\mathbf{y}$ in $\mathbf{b}$ zamenjamo z matrikama $\mathbf{Y}$ in $\mathbf{B}$. Metode, ki sledijo, so prav tako uporabne za eno ali več odvisnih spremenljivk.

2.2.2 Analiza glavnih komponent (Principal Component Analysis)

Z analizo glavnih komponent zapišemo matriko $\mathbf{X}$ ranga r kot vsoto r matrik ranga 1:

$$\mathbf{X} = \mathbf{M_1} + \mathbf{M_2} + \ldots + \mathbf{M_r}. \tag{2.5}$$

Vsako matriko $\mathbf{M_h}$ lahko zapišemo kot diadični produkt dveh vektorjev: $\mathbf{t_h}$ (score vector) in $\mathbf{p'_h}$ (loading vector), zato je enačbi 2.5 ekvivalentna enačba

$$\mathbf{X} = \mathbf{t_1 p'_1} + \mathbf{t_2 p'_2} + \ldots + \mathbf{t_r p'_r}, \tag{2.6}$$

oziroma v matričnem zapisu

$$\mathbf{X} = \mathbf{T P'}. \tag{2.7}$$

Vektorje $\mathbf{p}$ imenujemo *glavne komponente* (Principal Components) in so linearne kombinacije prvotnih neodvisnih spremenljivk (deskriptorjev). Glavne komponente so medseboj ortogonalne. Matrika $\mathbf{T}$ vsebuje informacije o učnih primerih, zapisane z glavnimi komponentami namesto z originalnimi deskriptorji. Razcep matrike $\mathbf{X}$ (enačba 2.6) lahko postopoma izračunamo z algoritmom NIPALS (Nonlinear Iterative PArtial Least Squares). Najprej s pomočjo matrike $\mathbf{X}$ izračunamo $\mathbf{t_1}$ in $\mathbf{p'_1}$. Diadični produkt tako dobljenih vektorjev odštejemo od matrike $\mathbf{X}$, da dobimo ostanek $\mathbf{E_1}$ ($\mathbf{E_1} = \mathbf{X} - \mathbf{t_1 p'_1}$). S pomočjo tega ostanka nato z algoritmom NIPALS izračunamo naslednja vektorja razcepa – $\mathbf{t_2}$ in $\mathbf{p'_2}$. Postopek ponavljamo, dokler je ostanek različen od nič. Zapišimo torej celoten postopek:

Za h, ki preteče vrednosti od 1 do r **ponavljaj:**

1. Iz $\mathbf{X}$ vzamemo poljuben vektor $\mathbf{x_j}$ in postavimo $\mathbf{t_h} = \mathbf{x_j}$.
2. Izračunamo $\mathbf{p'_h} = \mathbf{t'_h X}/\mathbf{t'_h t_h}$.
3. Normaliziramo $\mathbf{p'_h}$: $\mathbf{p'_{h}}_{novi} = \mathbf{p'_h}_{stari}/\left\|\mathbf{p'_h}_{stari}\right\|$.
4. Izračunamo $\mathbf{t_h} = \mathbf{X p_h}/\mathbf{p'_h p_h}$.
5. Če sta $\mathbf{t_h}$-ja iz drugega in četrtega koraka različna se vrnemo na drugi korak.
6. Izračunamo novo vrednost za $\mathbf{X}$ ($\mathbf{X} \leftarrow \mathbf{X} - \mathbf{t_h p'_h}$).

V ozadju algoritma NIPALS (koraki 1–5) je računanje lastnih vrednosti matrike $\mathbf{X'X}$ oz. $\mathbf{XX'}$. Če algoritem NIPALS konvergira, je dobljena rešitev enaka, kot bi jo dobili z računanjem lastnih vrednosti. S konvergenco v praktičnih primerih večinoma nimamo težav. Pomembna lastnost tako dobljenega razcepa matrike $\mathbf{X}$ je, da ima prva komponenta največjo težo, druga manjšo in tako naprej. Zato lahko sklepamo, da deskriptorji, ki nastopajo v prvih komponentah razcepa, najbolj vplivajo na odvisno spremenljivko.

2.2.3 Regresija glavnih komponent (Principal Component Regression)

Regresija glavnih komponent se izogne računanju matrike $(\mathbf{X'X})^{-1}$ tako, da namesto matrike $\mathbf{X}$ uporabi (v prejšnjem razdelku dobljeno) matriko $\mathbf{T}$. Namesto enačbe 2.3 imamo tako enačbo

$$\mathbf{y} = \mathbf{Tb} \tag{2.8}$$

ter rešitev v smislu najmanjših kvadratov

$$\hat{\mathbf{b}} = (\mathbf{T'T})^{-1}\mathbf{T'y}. \tag{2.9}$$

Zaradi medsebojne ortogonalnosti vektorjev $\mathbf{t_h}$, ki sestavljajo matriko $\mathbf{T}$, je ta dobro pogojena in ne pričakujemo težav pri računanju matrike $(\mathbf{T'T})^{-1}$. Dodatna prednost metode PCR je v tem, da lahko iz matrike $\mathbf{T}$ izpustimo vektorje $\mathbf{t_h}$, ki pripadajo majhnim lastnim vrednostim (zadnji členi razcepa 2.6). Na ta način še zmanjšamo kolinearnost vektorjev v matriki $\mathbf{T}$ ter odstranimo nekaj šuma. Seveda pa vedno obstaja nevarnost, da bomo s šumom zavrgli tudi informacije, ki so pomembne za naš model.

2.2.4 Metoda delnih najmanjših kvadratov (Partial Least Squares Method)

Metoda delnih najmanjših kvadratov temelji na algoritmu NIPALS. (Ker razširitev PLS na več odvisnih spremenljivk ni trivialna bo, za razliko od prejšnjih metod, podana za več neodvisnih spremenljivk – namesto vektorjev $\mathbf{y}$ in $\mathbf{b}$ imamo matriki $\mathbf{Y}$ in $\mathbf{B}$. Zožitev na eno odvisno spremenljivko je preprosta.) Kot pri PCA in PCR metodah je osnova razcep podatkovne matrike $\mathbf{X}$:

$$\mathbf{X} = \mathbf{TP}' + \mathbf{E} = \sum_{\mathbf{h=1}}^{\mathbf{a}} \mathbf{t_h p'_h} + \mathbf{E}. \tag{2.10}$$

Podobno lahko razcepimo tudi matriko $\mathbf{Y}$:

$$\mathbf{Y} = \mathbf{UQ}' + \mathbf{F}^* = \sum_{\mathbf{h=1}}^{\mathbf{a}} \mathbf{u_h q'_h} + \mathbf{F}^*. \tag{2.11}$$

V enačbah 2.10 in 2.11 nastopa vrednost a, ki predstavlja število členov razcepa, vključenih v matriki $\mathbf{T}$ in $\mathbf{P}$. Če upoštevamo vse člene, je $\mathbf{E} = \mathbf{F}^* = \mathbf{0}$, sicer pa ne. Oba razcepa bi lahko izračunali neodvisno z algoritmom NIPALS. Izkaže se (razlogov tu ne bomo obravnavali), da je bolje, če oba izračuna »prepletemo«. V algoritmu PLS so to koraki 1–8. Posledica takega izračuna je, da dobljeni vektorji $\mathbf{t}$ niso ortogonalni (vektorji $\mathbf{p}'$ so zamenjani z utežmi $\mathbf{w}'$). Ortogonalnost sicer ni nujna, je pa zaželjena, zato v agoritem dodamo korake 9–12. Podrobneje je metoda PLS opisana v [10].

Algoritem PLS:

1. Iz $\mathbf{Y}$ vzamemo vektor $\mathbf{y_j}$ in postavimo $\mathbf{u}_{start} = \mathbf{y_j}$.
2. Izračunamo $\mathbf{w}' = \mathbf{u}'\mathbf{X}/\mathbf{u}'\mathbf{u}$.
3. Normaliziramo $\mathbf{w}'_{novi} = \mathbf{w}'_{stari}/\left\|\mathbf{w}'_{stari}\right\|$.
4. Izračunamo $\mathbf{t} = \mathbf{Xw}/\mathbf{w}'\mathbf{w}$.
5. Izračunamo $\mathbf{q}' = \mathbf{t}'\mathbf{Y}/\mathbf{t}'\mathbf{t}$.
6. Normaliziramo $\mathbf{q}'_{novi} = \mathbf{q}'_{stari}/\left\|\mathbf{q}'_{stari}\right\|$.
7. Izračunamo $\mathbf{u} = \mathbf{Yq}/\mathbf{q}'\mathbf{q}$.

8. Primerjamo $\mathbf{t}$ iz četrtega koraka s tistim iz prejšnje iteracije. Če sta enaka (v okviru določene napake), gremo na korak 9, sicer na korak 2. (Če matrika $\mathbf{Y}$ vsebuje le eno spremenljivko, lahko korake 5–8 izpustimo, postavimo $q=1$ ter končamo z iteracijo.)
9. Izračunamo $\mathbf{p}' = \mathbf{t}'\mathbf{X}/\mathbf{t}'\mathbf{t}$.
10. Normaliziramo $\mathbf{p}'_{novi} = \mathbf{p}'_{stari}/\left\|\mathbf{p}'_{stari}\right\|$.
11. Normaliziramo $\mathbf{t}'_{novi} = \mathbf{t}'_{stari}/\left\|\mathbf{p}'_{stari}\right\|$.
12. Normaliziramo $\mathbf{w}'_{novi} = \mathbf{w}'_{stari}/\left\|\mathbf{p}'_{stari}\right\|$.
13. Izračunamo regresijski koeficient $b = \mathbf{u}'\mathbf{t}/\mathbf{t}'\mathbf{t}$.
14. Izračunamo ostanke (za h-to komponento):
    $\mathbf{E_h} = \mathbf{E_{h-1}} - \mathbf{t_h}\mathbf{p}'_\mathbf{h}$; $\quad \mathbf{X} = \mathbf{E_0}$;
    $\mathbf{F_h} = \mathbf{F_{h-1}} - b_h\mathbf{t_h}\mathbf{q}'_\mathbf{h}$; $\quad \mathbf{Y} = \mathbf{F_0}$.
15. Za izračun naslednje komponente se vrnemo na korak 1.

Po izračunu prve komponente zamenjamo $\mathbf{X}$ v korakih 2, 4 in 9 z ostankom $\mathbf{E_h}$ ter $\mathbf{Y}$ v korakih 5 in 7 z ostankom $\mathbf{F_h}$.

2.2.5 Določitev najpomembnejših deskriptorjev

Ena bistvenih prednosti metode PLS (v primerjavi z navadno metodo najmanjših kvadratov) je v tem, da nam da model z deskriptorji, ki nosijo največ informacije o odvisni spremenljivki. To je posledica uporabe algoritma NIPALS, zato lahko ta algoritem oz. analizo glavnih komponent v ta namen uporabimo neodvisno od metode PLS.

Za določitev najpomembnejših deskriptorjev lahko uporabimo tudi uteži $\mathbf{w}$, ki smo jih dobili pri PLS modeliranju (glej algoritem PLS). Utež w_{hk} namreč kaže pomembnost k-tega deskriptorja v h-ti komponenti modela PLS. Na tej osnovi lahko sestavimo vrednost VIP (Variable influence on projection):

$$VIP_k = \sqrt{\sum_h \frac{w_{hk}^2 \; SSY_h \; m}{SSY_{skupni} \; a}}, \tag{2.12}$$

ki podaja pomembnost k-te neodvisne spremenljivke. V enačbi pomeni m število vseh deskriptorjev, a število komponent PLS modela, SSY_h je vsota kvadratov odvisne spremenljivke, ki jo pojasni h-ta komponenta modela, SSY_{skupni} pa je vsota kvadratov odvisne spremenljivke, ki jo pojasni model z vsemi a-timi komponentami. Tipične vrednosti VIP so okrog 1, zato imajo pomembni deskriptorji VIP vrednost večjo od 1, nepomembni pa manjšo od 0.8. Več o koeficientu VIP lahko preberemo v [11].

2.3 Ocenjevanje veljavnosti modelov

V prejšnjem razdelku smo si ogledali klasično metodo QSAR modeliranja. Preden lahko model, ki ga dobimo s to ali katero drugo metodo (npr. regresijsko drevo) uporabimo za razlago mehanizmov aktivnosti ali za napovedovanje neznanih primerov, moramo oceniti njegovo veljavnost. Če je model veljaven pomeni, da je primerna predstavitev, približek dejanskega sistema znotraj proučevanega območja vrednosti neodvisnih in odvisnih spremenljivk.

Prvi pogoj za veljavnost modela je, da zadovoljivo opisuje primere iz učne množice. To pomeni, da se dejanske vrednosti odvisnih spremenljivk ne razlikujejo »preveč« od vrednosti, ki jih napove model. Mera za linearno odvisnost dveh nizov števil je korelacijski koeficient

$$r = \frac{\overline{(Y_d - \overline{Y}_d)(Y_n - \overline{Y}_n)}}{\sqrt{\overline{(Y_d - \overline{Y}_d)^2}\,\overline{(Y_n - \overline{Y}_n)^2}}}\,, \tag{2.13}$$

kjer Y_d pomeni dejansko vrednost ter Y_n napovedano vrednost odvisne spremenljivke. Korelacijski koeficient lahko zavzame vrednosti med -1 in 1. Za dober model si želimo vrednost r čim bližje 1. V QSAR modeliranju je razširena še ena mera za »podobnost« dejanskih in napovedanih vrednosti. To je koeficient R^2 (Multiple Corelation Coefficient, explained variance), podan z enačbo:

$$R^2 = 1 - \frac{\sum(Y_d - Y_n)^2}{\sum(Y_d - \overline{Y}_d)^2}\,. \tag{2.14}$$

Zaloga vrednosti koeficienta R^2 je navzgor omejena z 1, kar pomeni popolno ujemanje dejanskih in napovedanih vrednosti.

Veliki vrednosti koeficientov r in R^2 še ne pomenita, da je model veljaven, saj upoštevata le učne primere. Navadno si niti ne želimo prevelikega ujemanja modela z učnimi

primeri (overfitting). Vrednosti spremenljivk učnih primerov so izmerjene in zato vsebujejo šum. Če bi z modelom dosegli popolno ujemanje, bi to pomenilo, da model opisuje tudi šum. Tega pa vsekakor ne želimo.

Drugi pogoj za veljavnost modela je, da nam da zadovoljive napovedi za neznane primere (eden od ciljev QSAR modeliranja). Z modelom napovemo vrednosti testnih primerov ter jih primerjamo z dejanskimi. Ujemanje lahko ovrednotimo z enačbama 2.13 in 2.14. Za tak način ocenjevanja veljavnosti moramo celotno množico znanih primerov pred začetkom modeliranja razdeliti na dva dela: na učno in na testno množico (standardna delitev je 60% primerov za učno in 40% za testno množico). Iz tega sledi, da bo zgrajeni model slabši kot bi lahko bil, saj bo vseboval podatke iz samo 60% primerov. Poleg tega imamo pri QSAR modeliranju pogosto le nekaj deset primerov in moramo vse uporabiti za učenje modela. Če ne moremo ali ne želimo uporabiti testne množice primerov, jo lahko simuliramo. Ena od metod, ki nam to omogoča je t.i. prečno preverjanje (Cross-Validation).

Pri prečnem preverjanju sestavimo več spremenjenih učnih množic, tako da iz prvotne množice odstranimo manjšo skupino primerov. Vsak primer moramo izvzeti natanko enkrat (za vsak primer mora obstajati natanko en model, ki tega primera ne vsebuje). Pri t.i. »izloči enega« prečnem preverjanju (Leave One Out Cross-Validation) tako dobimo enako število spremenjenih učnih množic kot je število učnih primerov (število primerov v teh množicah je za 1 manjše od števila primerov v prvotni učni množici). Z vsako, tako dobljeno množico, zgradimo model in z njim napovemo vrednost odvisne spremenljivke za primer(e), ki v tej množici ne nastopajo. Tako dobljene napovedane vrednosti primerjamo z dejanskimi in njihovo ujemanje ovrednotimo s korelacijskim koeficientom prečnega preverjanja

$$q = \frac{\overline{(Y_d - \overline{Y}_d)(Y_n - \overline{Y}_n)}}{\sqrt{\overline{(Y_d - \overline{Y}_d)^2}\ \overline{(Y_n - \overline{Y}_n)^2}}} \tag{2.15}$$

ter koeficientom Q^2 (Cross-Validated Multiple Corelation Coefficient, predicted variance)

$$Q^2 = 1 - \frac{\sum (Y_d - Y_n)^2}{\sum (Y_d - \overline{Y}_d)^2}\ . \tag{2.16}$$

Oba koeficienta sta analogna koeficientoma 2.13 in 2.14 in opisujeta napovedno moč modela. Za veljaven model zahtevamo njuni vrednosti čim bližje 1. Okvirno pomeni vrednost

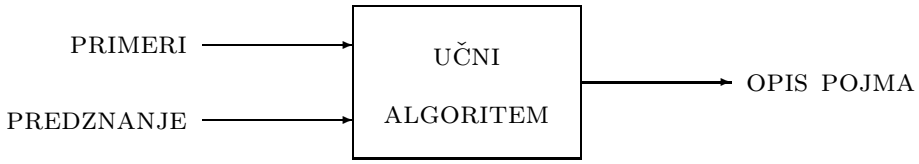
Q^2 večja od 0.6 smiseln model, vrednost večja od 0.9 pa odličen model. Dodaten pogoj za veljaven model je, da ni preveč prilagojen učnim primerom. Tak model ima približno enake vrednosti r in q (ali R^2 in Q^2). Več o preverjanju veljavnosti QSAR modelov najdemo v [12].

Opisane metode ocenjevanja veljavnosti modelov niso omejene zgolj na linearne regresijske modele. So splošno uporabne, med drugim tudi za ocenjevanje modelov dobljenih z metodami strojnega učenja opisanimi v naslednjem poglavju.

3. Strojno učenje

Strojno učenje (Machine Learning – ML) je veja raziskav umetne inteligence. Osnovni princip strojnega učenja je avtomatsko opisovanje (modeliranje) pojavov iz podatkov (glej [13]). Rezultat učenja iz podatkov so lahko so lahko pravila, funkcije, sistemi enačb ipd., ki so lahko predstavljeni z različnimi formalizmi: odločitvenimi pravili, odločitvenimi drevesi, regresijskimi drevesi, nevronskimi mrežami itn.

Algoritem za strojno učenje je prikazan na sliki 3.1. Cilj učečega se sistema je, da na osnovi podatkov (običajno v obliki množice primerov) in predznanja (background knowledge) zgradi opis danega pojma (model). Učne primere pripravi strokovnjak na področju uporabe; za vsak primer mora biti podan njegov opis in razvrstitev – klasifikacija. Predznanje vsebuje informacije o jeziku, s katerim so opisani primeri in pojmi (npr. možne vrednosti spremenljivk, s katerimi so opisani primeri, njihovo hierarhijo in podobno). Pomembna lastnost učnega algoritma je, da lahko uporablja nepopolne podatke



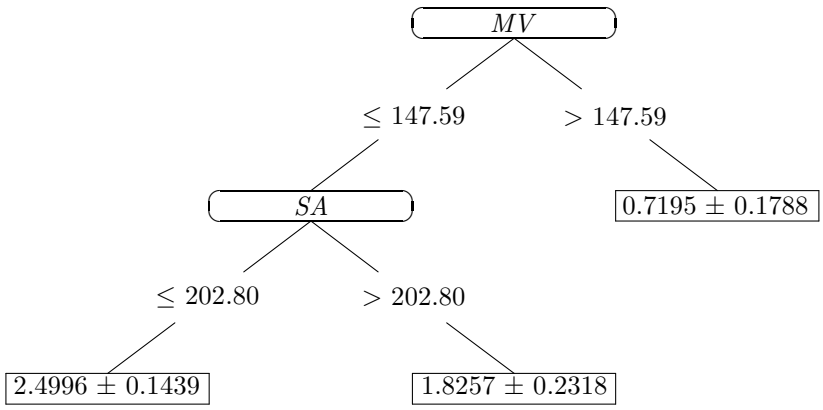
Slika 3.1: Algoritem za strojno učenje

(primerjaj [14]). Primeri pogosto vsebujejo določeno količino šuma (napak) tako v opisih kot v klasifikacijah. Pomanjkljivosti so lahko tudi v znanju s področja uporabe. Učne algoritme lahko v grobem razdelimo v dve skupini: metode na principu črne škatle (npr. nevronske mreže in statistične metode) in metode, ki omogočajo interpretacijo naučenega znanja (Knowledge-Oriented Methods). S pomočjo prvih dobimo opis pojma, uporaben za razpoznavanje pojmov, ki pa ga uporabnik le težko interpretira. Cilj druge skupine

metod je naučeno znanje zapisati v strukturirani, uporabniku razumljivi obliki. V tem diplomskem delu nas zanima zapis v obliki *odločitvenih dreves*.

3.1 Odločitvena drevesa

Odločitveno drevo (decision tree) je drevo, ki ima v vsakem notranjem vozlišču test, ki testira vrednost nekega *atributa*, v listih pa ima predpis, ki priredi neko vrednost *razredu* (glej [17]). Atributi so neodvisne spremenljivke (deskriptorji), s katerimi so opisani primeri. Razred je odvisna spremenljivka in je lahko diskretna ali zvezna. V prvem primeru pravimo takemu drevesu *klasifikacijsko drevo* (classification tree), v drugem pa *regresijsko drevo* (regression tree). Interpretacija drevesa se začne pri korenu. V vsakem vozlišču izvršimo test zapisan v tem vozlišču ter se glede na rezultat testa usmerimo v eno izmed poddreves. Ko prispemo v list, primeru pripišemo vrednost razreda zapisano v tem listu. Na sliki 3.2 vidimo primer regresijskega drevesa, ki ga bomo srečali v poglavju 4.4. Na



Slika 3.2: Primer regresijskega drevesa.

osnovi vrednosti dveh deskriptorjev spojine (MV in SA) nam da drevo oceno za njeno biorazgradljivost. Ta lahko zavzame eno od treh možnih vrednosti, ki so zapisane v listih. Zaloga vrednosti drevesa je torej diskretna množica. Če imamo namesto konstant v listih drevesa linearne enačbe, postane zaloga vrednosti zvezna množica. Takim regresijskim drevesom pravimo tudi *modelska drevesa* (model trees). Razlika med modelskim drevesom in regresijo, opisano v razdelku 2.2 je v tem, da slednja predpostavi enotno obliko funkcijske odvisnosti med atributi in razredom po celem problemskem prostoru, medtem

ko modelsko drevo dopušča v različnih območjih problemskega prostora različne oblike funkcijske odvisnosti.

3.1.1 Gradnja odločitvenih dreves

Za gradnjo odločitvenih dreves so v obstoječih sistemih za strojno učenje najbolj razširjeni algoritmi iz družine TDIDT (Top-Down Induction of Decision Trees). Njihova prednost je relativno majhna časovna zahtevnost, vendar nam ti algoritmi ne jamčijo izgradnje optimalnega drevesa, ker uporabljajo metodo lokalne optimizacije s samo enim korakom gledanja naprej (glej [17]). Algoritem TDIDT zapišemo takole:

ZgradiPoddrevo(*MnožicaPrimerov*)
Če *MnožicaPrimerov* ustreza ustavitvenemu pogoju **potem**
Naredi list, ki vsebuje celotno *MnožicoPrimerov*.
sicer
Izberi »najboljši« atribut A.
Atribut A razdeli *MnožicoPrimerov* v množici *Primeri1* in *Primeri2*.
ZgradiPoddrevo(*Primeri1*).
ZgradiPoddrevo(*Primeri2*).

Ustavitveni pogoj je lahko (glej [13]):

1. dovolj »čista« učna množica (npr. vsi ali večina primerov je iz istega razreda),
2. premalo učnih primerov za zanesljivo nadaljevanje gradnje drevesa,
3. ali pa je zmanjkalo (dobrih) atributov.

Za izbiro najboljšega atributa uporabimo izbrano mero (ne)čistoče (impurity measure) množice. Običajno je to entropija za odločitvena drevesa in varianca za regresijska drevesa. Za slednjo sta dva primera podana v nadaljevanju pri opisih orodij Cubist in RETIS.

Drevesu, zgrajenem z zgornjim algoritmom, pravimo *popolno drevo* in je običajno preveč prilagojeno učnim primerom. Temu se izognemo z *rezanjem* drevesa. Poznamo sprotno rezanje (pre-pruning) in naknadno rezanje (post-pruning) dreves. Prvo realiziramo z ustavitvenim pogojem, ki prepreči nadaljnjo gradnjo drevesa. Pogoj običajno

upošteva minimalno število primerov v vozlu, nehomogenost lista ter minimalen napredek, ki ga moramo doseči z razvojem vozla v poddrevo. Naknadno rezanje realiziramo tako, da najprej zgradimo popolno drevo, nato pa ocenimo, katere veje so *nezanesljive* in jih odstranimo z drevesa. Dobra ocena nezanesljivosti oz. pričakovane klasifikacijske napake na neznanih primerih je predpogoj za uspešnost naknadnega rezanja.

Algoritem TDIDT je implementiran v večih orodjih za gradnjo odločitvenih dreves. V tem diplomskem delu sta bila uporabljena sistema Cubist in RETIS, zato si ju na kratko oglejmo.

3.2 Cubist

Sistem Cubist je nadgradnja sistema za gradnjo modelskih dreves M5, opisanega v [15]. Ker gre za komercialno orodje, podrobnosti algoritma niso objavljene. Demonstracijska različica (za operacijski sistem Windows ter več vrst Unix-a) je na voljo na spletni strani podjetja *RuleQuest* (`www.rulequest.com`) in je omejena na največ 200 učnih primerov.

3.2.1 Algoritem

V nadaljevanju bo opisan algoritem sistema M5 ter njegove znane izboljšave v sistemu Cubist. Učna množica T vsebuje primere, opisane z določeno množico numeričnih ali diskretnih atributov ter pripadajočo ciljno vrednostjo. Najprej izračunamo standardno deviacijo ciljnih vrednosti v množici T. Če T ne vsebuje zelo malo primerov ali si njihove vrednosti niso zelo podobne, T razdelimo glede na rezultat testa. Naj bo T_i podmnožica primerov, ki ustreza i-temu izidu določenega testa. Če obravnavamo standardno deviacijo $sd(T_i)$ ciljnih vrednosti primerov v T_i kot mero napake, lahko pričakovano zmanjšanje napake zaradi obravnavanega testa zapišemo kot

$$\Delta error = sd(T) - \sum_i \frac{|T_i|}{|T|} \times sd(T_i). \tag{3.1}$$

Med vsemi možnimi testi M5 izbere tistega, ki maksimizira pričakovano zmanjšanje napake. Po tem postopku dobimo začetno drevo. Nadaljnji postopek je precej obširen, zato omenimo le nekaj bistvenih značilnosti.

Ocene napak. M5 večkrat potrebuje oceno točnosti modela za neznane primere. *Ostanek* modela za določen primer je absolutna vrednost razlike med dejansko ciljno vrednostjo in vrednostjo, ki jo napove model. Za oceno napake modela, zgrajenega na osnovi učnih primerov, M5 najprej izračuna povprečje ostankov modela za učne primere. Tako dobljena ocena napake v splošnem podceni napako modela za neznane primere, zato jo M5 pomnoži z $(n + \nu)/(n - \nu)$, kjer n pomeni število učnih primerov, ν pa je število parametrov v modelu. Posledica je povečanje ocene napake za modele z veliko parametri, zgrajene na podlagi majhnega števila učnih primerov.

Linearni modeli. S pomočjo klasičnih regresijskih metod je v vsakem vozlišču zgrajen linearni model, ki pa ne vsebuje vseh atributov, temvešamo tiste, ki se pojavljajo v pogojih ali linearnih modelih v pripadajočem poddrevesu.

Poenostavitev linearnih modelov. Ko M5 po zgoraj opisanem postopku dobi linearni model, ga poenostavi z izločanjem parametrov, tako da minimizira ocenjeno napako.

Rezanje drevesa. Za vsako notranje vozlišče M5 oceni napako poenostavljenega linearnega modela in pripadajočega poddrevesa ter izbere tistega z manjšo oceno napake. Če izbere linearni model, vozlišče postane list.

Glavna izboljšava sistema Cubist v primerjavi z M5 je ta, da omogoča gradnjo *sestavljenih* modelov. Sestavljeni modeli napovedujejo ciljne vrednosti neznanih primerov tako, da hkrati upoštevajo modelsko drevo in pravilo »k-najbližjih sosedov«. Slednje napove ciljno vrednost novega primera kot povprečje ciljnih vrednosti k najbolj podobnih primerov v učni množici. Cubist upošteva obe metodi na naslednji način. Najprej poišče pet (k=5) neznanemu primeru najbolj podobnih primerov v učni množici. Namesto da bi takoj izračunal povprečje njihovih ciljnih vrednosti, jih pred tem *popravi*. Označimo z $\mathbf{x}$ neznan primer in z $\mathbf{n}$ enega od $\mathbf{x}$-ovih petih najbližjih sosedov. Ciljno vrednost primera $\mathbf{n}$ poznamo, označimo jo z $\mathbf{T}(\mathbf{n})$. Z modelskim drevesom lahko napovemo ciljni vrednosti za $\mathbf{x}$ in $\mathbf{n}$; označimo ju z $\mathbf{M}(\mathbf{x})$ in $\mathbf{M}(\mathbf{n})$. Model torej napove razliko ciljnih vrednosti $\mathbf{x}$ in $\mathbf{n}$, ki znaša $\mathbf{M}(\mathbf{x})-\mathbf{M}(\mathbf{n})$. S to razliko Cubist popravi napovedano ciljno vrednost primera $\mathbf{x}$, ki jo napove sosed $\mathbf{n}$. Namesto vrednosti $\mathbf{T}(\mathbf{n})$ torej uporabi $\mathbf{T}(\mathbf{n})+\mathbf{M}(\mathbf{x})-\mathbf{M}(\mathbf{n})$. Tak način kombiniranja modelskih dreves in pravila »k-najbližjih sosedov« je podrobneje opisan v [16].

3.2.2 Uporabniški vmesnik

Zaradi sodobnega grafičnega uporabniškega vmesnika je uporaba programa zelo preprosta. Pred gradnjo modela moramo zapisati podatke o učnih primerih v obliki razumljivi sistemu Cubist (glej prilogo B). To pomeni, da v eni datoteki podamo atribute s katerimi so opisani učni primeri, v drugi pa za vsak primer podamo vrednosti atributov ter ciljno vrednost. Zgrajeno modelsko drevo nam Cubist poda kot množico pravil. Vsakemu listu drevesa ustreza eno pravilo. Na gradnjo modela lahko vplivamo z nastavitvijo naslednjih parametrov.

Oblika modela. Izbiramo lahko med modelom v obliki pravil ter modelom, ki uporablja kombinacijo pravil in metode »k-najbližjih sosedov«. Izbiro lahko prepustimo tudi Cubistu. Privzeta vrednost je model v obliki pravil.

Najmanjše število primerov, ki ga vsebuje posamezno pravilo: Podamo ga v odstotkih vseh primerov; privzeta vrednost je 1%.

Dovoljena ekstrapolacija. Pri vsakem pravilu si Cubist zapomni največjo in najmanjšo ciljno vrednost učnih primerov, ki ustrezajo pogojem pravila. Napovedana ciljna vrednost novega primera lahko pade iz tega območja. S tem parametrom določimo za koliko se lahko ciljna vrednost novega primera razlikuje od vrednosti učnih primerov. Vrednost podamo v odstotkih območja ciljnih vrednosti učnih primerov, privzeta vrednost je 10%.

Faktor jedrnatosti (brevity factor). V prejšnjem razdelku smo videli, da Cubist najprej zgradi začetni model ter ga nato poenostavi. Na stopnjo poenostavitve vplivamo s faktorjem jedrnatosti. Izbiramo lahko vrednosti med 0 in 100%; 100% pomeni najbolj jedrnat (poenostavljen) model.

Prečno preverjanje. Z izbiro te možnosti izvedemo prečno preverjanje veljavnosti modela (postopek je opisan v razdelku 2.3). Določiti moramo število podmnožic na katere razdelimo celotno učno množico. Če je izbrano število podmnožic enako številu učnih primerov, izvedemo t.i. »izloči enega« prečno preverjanje.

Uporabniški vmesnik vsebuje še nekatera druga orodja. Ugotavljamo lahko kateri primeri pripadajo kakšnemu pravilu, iz grafa napovedanih vrednosti, v odvisnosti od dejanskih, pa lahko identificiramo izstopajoče primere. Cubist lahko kličemo tudi iz paketne datoteke ter ga tako npr. uporabimo za gradnjo optimiziranih modelov.

3.3 RETIS

Sistem za gradnjo regresijskih dreves RETIS (REgression Tree Induction System) je bil razvit na Institutu Jožef Stefan v Ljubljani. Deluje pod operacijskim sistemom DOS in omogoča opis primerov z največ 30 atributi.

3.3.1 Algoritem

Algoritem pripada že omenjeni skupini algoritmov za gradnjo odločitvenih dreves TDIDT. Zaradi njegove obsežnosti si bomo na tem mestu ogledali le nekaj njegovih značilnosti (podrobneje je opisan v [17]).

Ocena verjetnosti. V postopku gradnje klasifikacijskega drevesa večkrat potrebujemo oceno verjetnosti posameznega (diskretnega) razreda, oz. v primeru regresijskega drevesa, oceno porazdelitve (zveznega) razreda. Ocenjevanje verjetnosti z relativnimi frekvencami je zanesljivo le za množice z velikim številom primerov. Pri gradnji drevesa pa se število primerov hitro zmanjšuje, ko potujemo iz korena proti listom. V tem primeru nam bolj realistično oceno verjetnosti nastopa posameznega razreda da t.i. m-ocena:

$$p_{\oplus} = \frac{n_{\oplus} + mp_a}{N + m}, \tag{3.2}$$

kjer je $n_{\oplus}$ število pripadnikov razreda $\oplus$, N je skupno število vseh primerov v podmnožici, p_a je *apriorna* verjetnost razreda $\oplus$ in m parameter ocene. Izbira vrednosti p_a in m je odvisna od področja uporabe. V splošnem velja, da naj ima m, za podatke z veliko šuma, večjo vrednost kot za podatke z malo šuma. Analogno m-oceni verjetnosti lahko razvijemo m-oceno porazdelitve, ki jo pri gradnji regresijskih dreves uporablja RETIS.

Izbor »najboljšega« atributa. Bistven del TDIDT algoritma je izbor atributa, ki omogoča razbitje na podmnožice z minimalno pričakovano nehomogenostjo primerov. Za mero nehomogenosti RETIS uporablja m-oceno porazdelitve razreda in m-oceno standardne deviacije razreda.

Naknadno rezanje drevesa. Naknadno rezanje poteka po naslednjem postopku. V vsakem vozlišču ocenimo *statično* in *vzvratno* (backed-up) napako. Statična napaka nam pove, kakšna bi bila pričakovana napaka na neznanih primerih, če bi bilo to vozlišče list (torej, če bi pripadajoče poddrevo odrezali). Vzvratna napaka je pričakovana ocena

napake, če poddrevesa v tem vozlišču ne odrežemo. Če je statična napaka manjša ali enaka vzvratni, poddrevo odrežemo in vozlišče spremenimo v list. Pri ocenah napak ima pomembno vlogo m-ocena verjetnosti.

Linearni modeli. Za izračun linearnih modelov v listih drevesa RETIS uporablja metodo uteženih najmanjših kvadratov. Pri tem uporabi vse atribute.

3.3.2 Uporabniški vmesnik

Uporabniški vmesnik programa je enostaven in pregleden. Podobno kot pri sistemu Cubist, tudi tukaj v eni datoteki opišemo atribute učnih primerov, v drugi pa učne primere same (glej prilogo C). Pred gradnjo drevesa moramo nastaviti naslednje parametre.

Najmanjše število primerov v posameznem listu. Pomeni isto kot pri Cubistu, le da tukaj podamo število primerov in ne njihov odstotek.

Lokalno regresijo. Izberemo lahko ali bo imelo zgrajeno drevo v listih linearne modele (modelsko drevo) ali le kostantne vrednosti (regresijsko drevo).

Parameter m za učenje. Izberemo vrednost m, ki jo bo RETIS uporabljal za ocenjevanje verjetnosti med učenjem drevesa. Če izberemo m=0, bomo dobili popolno, neobrezano drevo.

Parameter m za naknadno rezanje. Izberemo vrednost m, ki jo bo RETIS uporabljal za ocenjevanje verjetnosti med naknadnim rezanjem drevesa. Večja vrednost pomeni močneje porezano drevo.

Običajno najprej zgradimo popolno, neobrezano drevo ter ga nato naknadno režemo (lahko tudi večkrat). Zgrajeno drevo si lahko ogledamo na zaslonu ali ga npr. zapišemo v formatu LaTeX. Podobno kot Cubist lahko tudi RETIS kličemo iz paketne datoteke.

4. Poskusi

Glavni namen pričujočega diplomskega dela je bil v tem, da ugotovimo v kolikšni meri lahko na področju QSAR modeliranja nadomestimo klasične regresijske metode z metodami strojnega učenja. Zato so bili izvedeni poskusi na več različnih množicah podatkov. Dobljeni modeli in njihova primerjava s klasičnimi modeli je podana v tem poglavju, ki predstavlja jedro diplomskega dela.

4.1 Metodologija poskusov

Uporabljenih je bilo devet različnih množic podatkov ter orodji za gradnjo regresijskih dreves Cubist in RETIS. Prvih sedem množic podatkov je javno dostopnih na spletnem naslovu `www.chemi.muni.cz/~jiri/`. Vsebujejo majhno število primerov, večinoma s področja biorazgradljivosti spojin. Za te množice so bili, poleg modelov na osnovi celotnih množic, zgrajeni tudi nekateri modeli na osnovi njihovih izpeljank. Te vsebujejo zožen nabor deskriptorjev in/ali primerov in so bile dobljene na osnovi klasičnega QSAR modeliranja. Zadnji dve množici (P1 in P2) so pripravili avtorji članka [8]. Vsebujejo bistveno več primerov, zato se tudi parametri modeliranja nekoliko razlikujejo.

Za množice z malo učnimi primeri sta bila z orodjem Cubist zgrajena po dva modela. Eden s privzetimi ter eden z optimiziranimi parametri. Optimizirani parametri so tisti, pri katerih je imel z njimi dobljen model, največji korelacijski koeficient prečnega preverjanja q; za vsako množico podatkov so različni. Za množici z večjim številom primerov sta bila zgrajena le modela s privzetimi parametri.

S sistemom RETIS je bilo za množice z malo primeri zgrajenih po šest modelov: z vključeno in izključeno linearno regresijo v listih dreves, ter s tremi različnimi vrednostmi parametra m za naknadno rezanje dreves (0, 0.5 in 1). Za množici z več primeri so

bili zgrajeni po štirje modeli: z vključeno in izključeno linearno regresijo v listih dreves ter z dvema različnima vrednostima parametra m za naknadno rezanje dreves. Prva vrednost parametra je bila vedno 1, druga pa je bila interaktivno določena tako, da je imelo porezano drevo največ osem listov. S tem smo omogočili strokovno interpretacijo modelov. V vseh primerih je bila vrednost parametra m za učenje enaka 0, najmanjše dovoljeno število primerov v listih drevesa pa 1.

Morebitno odstopanje od zgoraj navedene sheme izdelanih modelov je opisano pri posameznih množicah podatkov.

Za vse modele dobljene po opisanem postopku smo izvedli »izloči enega« prečno preverjanje natančnosti napovedi neznanih primerov. Izbrane modele malih množic z dovolj veliko napovedno močjo je pregledal J. Damborsky in jih strokovno komentiral. Modele množic P1 in P2 je strokovno komentiral B. Kompare. Vse klasične PLS modele je izdelal J. Damborsky.

4.2 Toksičnost ter biorazgradljivost anilinov in fenolov

4.2.1 Uvod

Na področju varovanja okolja lahko QSAR metode uporabimo za napovedovanje toksičnosti (QSTR) in biorazgradljivosti (QSBR). V primerjavi s številom primerov uporabe QSTR je znanih relativno malo primerov uporabe QSBR. Pri slednjem je še vedno veliko neznank glede izbire primernih molekulskih deskriptorjev, glede najbolj zanesljivih eksperimentalnih podatkov, kot tudi glede samega procesa biorazgradnje. Zato bi bilo koristno, če bi lahko obširno znanje s področja modeliranja toksičnosti uporabili pri modeliranju biorazgradljivosti. V nadaljevanju so prikazani modeli dveh skupin spojin (anilinov in fenolov) za toksičnost in biorazgradljivost.

4.2.2 Množica podatkov

Toksičnost anilinov in fenolov je bila izmerjena s populacijsko rastjo *Tetrahymena pyriformis*, soj GL-C. Toksičnost je podana kot koncentracija, pri kateri je rast zavrta 50% (IGC_{50}). Pri biorazgradljivosti je upoštevana le prva stopnja oksidacije; fenoli so bili oksigenirani z oksigenazo, anilini pa z dioksigenazo. V tem primeru gre bolj za biotransformacijo, kot za biorazgradnjo. Hitrost oksidacije je podana s kinetično konstanto drugega reda (k_b).

Podatki (kot primer so podani v prilogi A) in PLS modeli so povzeti po [1]. Za opis strukturnih razlik med obravnavanimi spojinami je bilo *a priori* izbranih devet molekulskih deskriptorjev (tabela 4.1). Izbor je bil omejen na hidrofobične in stereoelektronske

$\log K_{ow}$	logaritem porazdelitvenega koeficienta oktanol/voda (hidrofobičnost)	r_w	Van der Waalsov radij
		V_w	Van der Waalsov volumen
		μ	dipolni moment
HOMO	energija najvišje zasedene molekulske orbitale	M_w	molekulska teža
		σ	Hammettova sigma konstanta
LUMO	energija najnižje nezasedene molekulske orbitale	pK_a	ionizacijska konstanta

Tabela 4.1: Izbrani molekulski deskriptorji za aniline in fenole.

parametre, za katere je znano, da so pomembni pri modeliranju toksičnosti oziroma biorazgradljivosti.

4.2.3 Rezultati

Cilj modeliranja je bil ugotoviti razlike in povezave med toksičnostjo in biorazgradljivostjo anilinov in fenolov, zato so bili zgrajeni modeli za toksičnost in biorazgradljivost za aniline in fenole posebej ter za oboje skupaj.

Modeliranje s PLS metodo zahteva smiseln izbor deskriptorjev. V tem primeru so bili deskriptorji izbrani na podlagi (čim večjega) pomembnostnega faktorja VIP in koeficienta Q^2.

PLS modeli so podani v nadaljevanju.

• Toksičnost anilinov:

ni uporabnega modela.

• Toksičnost fenolov:

$$\log IGC_{50}^{-1} = 0.5742\ \log K_{ow} - 0.5292\ HOMO - 0.5597\ LUMO - \\ - 0.0122\ V_w - 5.043. \tag{4.1}$$

• Toksičnost anilinov in fenolov skupaj:

ni uporabnega modela.

• Biorazgradljivost anilinov:

$$\log k_b = -11.237\ r_w + 0.0092\ M_w + 0.3737\ pK_a - 14.194. \tag{4.2}$$

• Biorazgradljivost fenolov:

$$\log k_b = -13.7430 + 0.0351\ V_w + 0.1946\ pK_a - 13.427. \tag{4.3}$$

• Biorazgradljivost anilinov in fenolov skupaj:

$$\log k_b = -11.233\ r_w + 0.3147\ pK_a - 12.738. \tag{4.4}$$

S sistemom Cubist je bil uporaben model dobljen le za biorazgradljivost anilinov in fenolov skupaj (enačba 4.5). V vseh ostalih primerih so bile vrednosti Q^2, ki nam kažejo napovedno moč modela, bistveno nižje od 0.5 (tabela 4.2). Dobljeni model je skoraj enak PLS modelu (enačba 4.4). Vsebuje samo eno pravilo z istima deskriptorjema, majhne razlike so le v utežeh.

$$\log k_b = -10.5\ r_w + 0.328\ pK_a - 12.983 \tag{4.5}$$

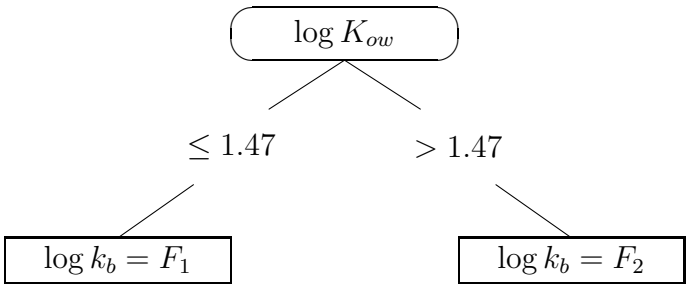
Vrednost koeficienta R^2=0.98 sicer kaže na boljše ujemanje učnih primerov kot pri PLS modelu (R^2=0.955), vendar smemo zaradi nižje vrednosti Q^2 (Cubist: 0.82; PLS: 0.949) pričakovati, da se bo bo model pri ocenjevanju novih primerov slabše obnesel kot PLS model.

Model	PLS		Cubist							
			Privzeti parametri				Optimizirani parametri			
	R^2	Q^2	r	q	R^2	Q^2	r	q	R^2	Q^2
Toksičnost anilinov	–	–	0.00	-0.39	0.00	-0.96	0.00	-0.39	0.00	-0.96
Toksičnost fenolov	0.96	0.83	0.83	-0.15	0.69	-0.44	0.83	-0.15	0.69	-0.44
Toksičnost anilinov in fenolov skupaj	-	-	0.51	0.05	0.26	-0.24	0.51	0.05	0.26	-0.24
Biorazgradljivost anilinov	0.95	0.89	0.97	-0.40	0.93	-0.77	0.97	-0.40	0.93	-0.77
Biorazgradljivost fenolov	0.99	0.93	0.98	0.48	0.96	0.16	0.98	0.48	0.96	0.16
Biorazgradljivost anilinov in fenolov skupaj	0.96	0.95	0.98	0.91	0.96	0.82	0.98	0.91	0.96	0.82

Model	PLS		RETIS							
			Regresija (m=1)				Brez regresije (m=1)			
	R^2	Q^2	r	q	R^2	Q^2	r	q	R^2	Q^2
Toksičnost anilinov	–	–	1.00	0.24	1.00	-9.02	0.99	-0.11	0.75	-0.47
Toksičnost fenolov	0.96	0.83	1.00	0.96	1.00	0.18	0.98	-0.08	0.81	-0.25
Toksičnost anilinov in fenolov skupaj	-	-	0.97	-0.39	0.94	-4.05	0.94	0.18	0.79	-0.10
Biorazgradljivost anilinov	0.95	0.89	1.00	-0.59	1.00	-135	0.85	0.55	0.68	0.28
Biorazgradljivost fenolov	0.99	0.93	1.00	-0.43	1.00	-670	0.95	-0.24	0.80	-0.91
Biorazgradljivost anilinov in fenolov skupaj	0.96	0.95	1.00	0.92	1.00	0.84	0.94	0.72	0.84	0.49

Tabela 4.2: Korelacijski koeficienti modelov za biorazgradljivost in toksičnost anilinov in fenolov.

S sistemom RETIS smo smiseln model dobili le za skupno biorazgradljivost anilinov in fenolov (slika 4.1). Kljub malenkost večjemu koeficientu Q^2 od Cubistovega modela (enačba 4.5), je model precej bolj zapleten. Upošteva namreč vseh devet deskriptorjev. Iz dobljenih modelov lahko sklepamo, da je kvaliteta modelov z regresijskimi drevesi zelo odvisna od števila učnih primerov. Pri modeliranju samo anilinov in samo fenolov nismo dobili uporabnega modela, ker smo imeli le 7 oziroma 8 primerov. Pri modeliranju obeh skupin spojin skupaj (15 primerov), pa smo dobili model za biorazgradljivost primerljiv s klasičnim. Podobno kot PLS, nam za toksičnost tudi Cubist in RETIS nista dala uporabnega modela. Sklepamo lahko, da je toksičnost odvisna še od drugih dejavnikov, ki niso bili zajeti v uporabljenih deskriptorjih.



$$
\begin{aligned}
F_1 &= 3.77 - 0.8474\ \log K_{ow} + 2.3240\ HOMO - 11\ r_w - 0.0939\ V_w \\
&\quad +0.7066\ \mu + 0.0604\ M_w - 3\ \sigma + 0.8604\ pK_a \\
F_2 &= -21.29 - 0.9011\ \log K_{ow} + 1.01\ HOMO + 11\ LUMO - 35\ r_w \\
&\quad +0.1435\ V_w + 2.2642\ \mu + 0.0173\ M_w + 9\ \sigma + 1.1061\ pK_a
\end{aligned}
$$

Slika 4.1: Model za biorazgradljivost anilinov in fenolov zgrajen z RETIS-om.

4.3 Akutna toksičnost nasičenih in nenasičenih alifatskih ogljikovodikov

4.3.1 Uvod

Halogenirani alifatski ogljikovodiki so bili dolgo časa množično uporabljani kot organska topila in hladilni medij, kot posredniki pri kemijski sintezi in na mnogih drugih področjih. Ocene svetovne proizvodnje se merijo v milijonih ton na leto. Zaradi njihove množične uporabe je bil velik del te proizvodnje izpuščen neposredno v okolje. Medtem so odkrili, da je veliko klorovih ogljikovodikov (vinilklorid, kloroform, itn.) kancerogenih ali drugače škodljivih človeku in živalim. QSAR modeliranje nam lahko pomaga pri ocenjevanju toksičnosti spojin, katerih toksičnost še ni bila eksperimentalno določena.

4.3.2 Množica podatkov

Obstaja veliko različnih načinov ocenjevanja vpliva kemikalij na žive organizme. Microtox test je eden najpogosteje uporabljanih sistemov za ocenjevanje akutne toksičnosti čistih

spojin. Merimo zmanjšanje bioluminiscence morske bakterije *Photobacterium phosphoreum*, na podlagi katere določimo efektivno koncentracijo (*EC50*) preverjane spojine.

Učna množica 19. spojin, 11 haloalkanov in 8 haloalkenov, je bila izbrana iz skupine 58. halogeniranih alifatskih ogljikovodikov. Za tri izmed teh spojin se je izkazalo, da niso stabilne v 2% vodni raztopini NaCl, zaradi česar njihove toksičnosti ni bilo moč oceniti. Na podlagi izkušenj s QSAR modeliranjem ogljikovodikov je bilo določenih 23 molekulskih deskriptorjev (tabela 4.3). Njihove vrednosti so bile zbrane iz literature, oziroma izračunane z metodami računske kemije. Podatki in PLS modeli so iz [2].

MW	molekulska teža	*IP*	ionizacijski potencial
bp	vrelišče	*CR*	odboj jedro-jedro
n	refrakcijski indeks	*Dip*	dipolni moment
D	gostota	*BCHO*	prispevek HOMO k vezi
Solu	topnost v vodi	*BCLU*	prispevek LUMO k vezi
$\log P$	logaritem porazdelitvenega koeficienta oktanol/voda	*BO*	razpored vezi
		Ha	celotna trdota
Hf	toplota tvorbe	*EV*	elipsoidna prostornina
TE	skupna energija	*SA*	površina
EE	energija elektronov (velikost molekule)	*MV*	molarni volumen
		MR	molarna refrakcija
HOMO	energija HOMO (najvišja zasedena molekulska orbitala)	*Kow*	porazdelitveni koeficient oktanol/voda (iz TSAR programa)
LUMO	energija LUMO (najnižja nezasedena molekulska orbitala)	*TLi*	celotna lipofilnost

Tabela 4.3: Izbrani molekulski deskriptorji za halogenirane alifatske ogljikovodike.

4.3.3 Rezultati

Modeli so bili zgrajeni za haloalkane (11 spojin), ter za haloalkane in haloalkene skupaj (16 spojin). Modeliranje haloalkenov ni bilo mogoče zaradi premajhnega števila spojin (5). S klasično (PLS) metodo sta bila dobljena dva modela za haloalkane (enačbi 4.6 in 4.7), ter trije za haloalkane in haloalkene skupaj (enačbe 4.8, 4.9 ter 4.10). Pri gradnji modela 4.9 je bilo uporabljenih le 14 spojin, izvzeta sta bila dva izstopajoča haloalkena.

PLS modeli so podani v nadaljevanju.

• Haloalkani (vsi deskriptorji):

$$\begin{aligned}\log EC_{50}^{-1} &= 0.0008\ Mw + 0.0016\ bp + 0.2683\ n + 0.0038\ D + \\ &+ 0.0747\ \log P - 0.0013\ Hf - 0.0003\ TE - 0.0001\ EE + \\ &+ 0.0001\ CR - 0.0683\ IP + 0.0683\ HOMO + \\ &+ 0.0024\ LUMO + 0.1289. \end{aligned} \tag{4.6}$$

• Haloalkani (samo deskriptorja *MR* in *EE*):

$$\log EC_{50}^{-1} = -0.0003\ EE + 0.0671\ MR - 2.6298. \tag{4.7}$$

• Haloalkani in haloalkeni skupaj (samo deskriptorja *MR* in *EE*):

$$\log EC_{50}^{-1} = -0.0003\ EE + 0.0847\ MR - 2.8159. \tag{4.8}$$

• Haloalkani in haloalkeni brez dveh spojin (samo deskriptorja *MR* in *EE*):

$$\log EC_{50}^{-1} = -0.0003\ EE + 0.0726\ MR - 2.8722. \tag{4.9}$$

• Haloakani in haloalkeni (samo deskriptorji *MR, EE, BO, Hf* in *CR*):

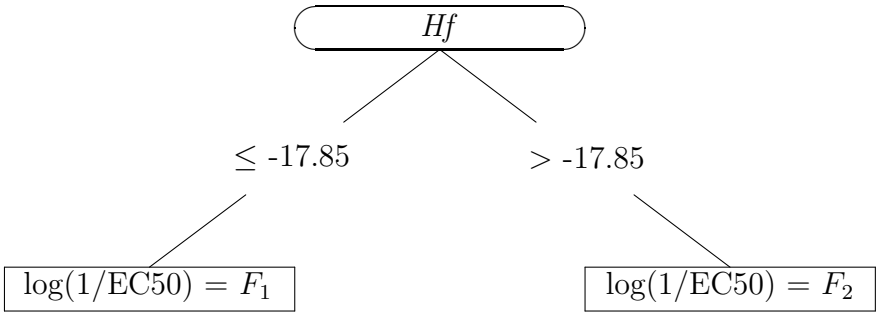
$$\begin{aligned}\log EC_{50}^{-1} &= -0.0002\ EE + 0.1136\ MR + 0.0347\ Hf + \\ &+ 0.0002\ CR - 20.6560\ BO + 17.291. \end{aligned} \tag{4.10}$$

Iz Cubistovega modela za haloalkane z vsemi deskriptorji (enačba 4.11) in RETISovega z deskriptorji *MR, EE, BO, Hf* in *CR* (slika 4.2) vidimo, da niti Cubist niti RETIS nista uspela določiti obeh deskriptorjev, za katera je znano da sta pomembna za toksičnost (*MR* in *EE*).

$$\log EC_{50}^{-1} = -3.399 + 0.0316\ SA - 0.57\ BCHO \tag{4.11}$$

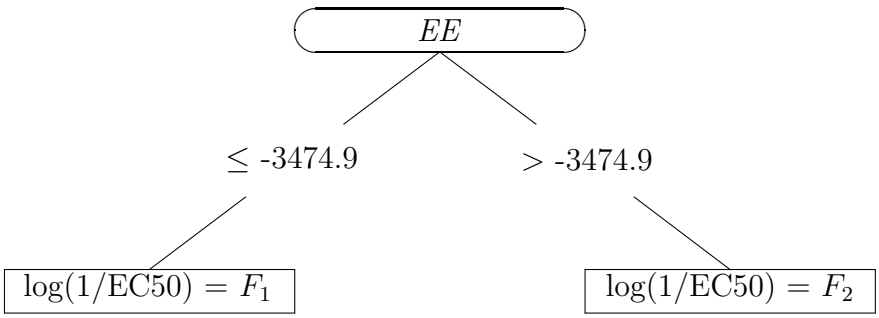
Modeli, dobljeni z omejitvijo deskriptorjev na *MR* in *EE*, so bolj primerljivi s klasičnimi. Za haloalkane dobimo s Cubistom model 4.12, v katerem je le deskriptor *MR*:

$$\log EC_{50}^{-1} = -3.195 + 0.1225\ MR. \tag{4.12}$$



$$F_1 = 25.7 + 0.016\ Hf - 0.0014\ EE - 0.0015\ CR - 30.2\ BO + 0.11\ MR$$
$$F_2 = 14.9 + 0.09\ Hf + 0.00014\ EE + 0.0011\ CR - 17.4\ BO + 0.058\ MR$$

Slika 4.2: Model za haloalkane in haloalkene(deskriptorji *MR, EE, BO, Hf* in *CR*) zgrajen s sistemom RETIS, z regresijo.



$$F_1 = 4.66 - 0.0039\ EE - 0.72\ MR$$
$$F_2 = -2.72 + 0.000069\ EE + 0.1\ MR$$

Slika 4.3: Model za haloalkane (samo deskriptorja *MR* in *EE*) zgrajen s sistemom RETIS, z regresijo, m=1.

Odsotnost deskriptorja *EE* je vzrok za slabšo napovedno moč tega modela v primerjavi s klasičnim. RETIS nam da model na sliki 4.3.

Molekule razdeli glede na velikost (*EE*); model za manjše molekule je zelo podoben

modelu PLS (enačba 4.7). Korelacijski koeficienti vseh modelov so v tabeli 4.4.

Model	PLS		Cubist							
			Privzeti parametri				Optimizirani parametri			
	R^2	Q^2	r	q	R^2	Q^2	r	q	R^2	Q^2
Haloalkani (vsi deskriptorji)	0.90	0.77	0.92	0.79	0.84	0.61	0.92	0.84	0.84	0.71
Haloalkani (samo deskriptorja *MR* in *EE*)	0.90	0.88	0.93	0.83	0.86	0.65	0.93	0.88	0.86	0.75
Haloalkani in haloalkeni skupaj (samo deskriptorja *MR* in *EE*)	0.42	0.30	0.71	0.62	0.51	0.38	0.71	0.62	0.51	0.38
Haloalkani in haloalkeni skupaj brez dveh spojin (samo deskriptorja *MR* in *EE*)	0.89	0.88	0.94	0.92	0.88	0.84	0.98	0.91	0.96	0.82
Haloakani in haloalkeni (samo deskriptorji *MR, EE, BO, Hf* in *CR*)	0.85	0.68	0.71	0.62	0.51	0.38	0.71	0.62	0.51	0.38

Model	PLS		RETIS							
			Regresija (m=1)				Brez regresije (m=1)			
	R^2	Q^2	r	q	R^2	Q^2	r	q	R^2	Q^2
Haloalkani (vsi deskriptorji)	0.90	0.77	0.97	0.36	0.83	0.06	0.97	0.36	0.83	0.06
Haloalkani (samo deskriptorja *MR* in *EE*)	0.90	0.88	0.98	0.93	0.95	0.85	0.95	0.59	0.78	0.31
Haloalkani in haloalkeni skupaj (samo deskriptorja *MR* in *EE*)	0.42	0.30	0.87	0.35	0.75	-0.23	0.91	0.22	0.73	-0.10
Haloalkani in haloalkeni skupaj brez dveh spojin (samo deskriptorja *MR* in *EE*)	0.89	0.88	0.98	0.94	0.97	0.88	0.96	0.81	0.88	0.65
Haloakani in haloalkeni (samo deskriptorji *MR, EE, BO, Hf* in *CR*)	0.85	0.68	0.99	0.78	0.97	0.59	0.95	0.36	0.79	0.07

Tabela 4.4: Korelacijski koeficienti modelov zgrajenih za nasičene in nenasičene alifatske ogljikovodike.

4.4 Biorazgradljivost dioksinov in furanov

4.4.1 Uvod

Dioksini in furani so človeku in okolju škodljive spojine. Nastajajo npr. pri sežiganju komunalnih odpadkov. Zaradi velikih količin le-teh je zelo pomembno, da poznamo hitrost in mehanizme njihove biorazgradnje.

4.4.2 Množica podatkov

Aerobična degradacija proučevanih spojin je bila izvedena s *Sphingomonas sp.*, soj RW1. Iz skupine 210. kloriranih dibenzo-p-dioksinov in dibenzofuranov je bila s pomočjo PCA metode (razdelek 2.2.2) izbrana učna množica 16. spojin. Strukturne lastnosti spojin so bile opisane s 50. molekulskimi in atomskimi deskriptorji (tabela 4.5).

$\log P$	logaritem porazdelitvenega koeficienta oktanol/voda	*HOMO*	energija najvišje zasedene molekulske orbitale (MOPAC)
MM	molekulska masa	*homo*	energija najvišje zasedene molekulske orbitale (Turbomole)
SA	površina	*LUMO*	energija najnižje nezasedene molekulske orbitale (MOPAC)
MV	molekulski volumen	*lumo*	energija najnižje nezasedene molekulske orbitale (Turbomole)
IM1s – IM3s	vztrajnostni momenti (velikost) – skupaj trije deskriptorji	*IP*	ionizacijski potencial
IM1l – IM3l	vztrajnostni momenti (dolžina) – skupaj trije deskriptorji	*Ded*	elektrofilne krajevne nedoločenosti – skupaj 9 deskriptorjev
MR	molarna refrakcija	*Dnd*	nukleofilne krajevne nedoločenosti – skupaj 9 deskriptorjev
TE	skupna energija (MOPAC)	*qd*	delni atomski naboji – skupaj 9 deskriptorjev
te	skupna energija (Turbomole)		
BCLU	prispevek LUMO k vezi		
DIP	dipolni moment (MOPAC)		
dip	dipolni moment (Turbomole)		
HARD	trdota		
HF	toplota tvorbe		

Tabela 4.5: Molekulski deskriptorji za furane in dioksine.

4.4.3 Rezultati

S klasično analizo so bili razviti trije modeli. Začetni model (ni prikazan) vsebuje vse deskriptorje, razmerje R^2:Q^2 je 0.94:0.78. Njegova uporabnost je predvsem zaradi zapletenosti omejena; pri razumevanju mehanizmov biorazgradljivosti si z njim ne moremo pomagati. S PCA analizo sta bili izbrani še dve podmnožici deskriptorjev. Prva s 15. elementi ($\log P$, *MM, SA, MV, IM1s, IM2s, IM3s, MR, te, BCLU, dip, HARD, De5, Dn5, Dn4*) in druga z 9. elementi ($\log P$, *MM, SA, MV, IM2s, IM3s, MR, te, dip*). Modela s tem naborom deskriptorjev (enačbi 4.13 in 4.14) imata boljše razmerje R^2:Q^2 (0.95:0.88 in 0.94:0.92) ter sta bistveno preprostejša. Predvsem zadnji ima veliko napovedno moč.

PLS modeli so podani v nadaljevanju.

• Model s petnajstimi deskriptorji:

$$\begin{aligned}\log k \;=\;& -0.1161\ \log P - 0.0019\ MM - 0.0048\ SA - 0.0041\ MV - \\ & - 0.0015\ IM1s - 0.0009\ IM2s - 0.0008\ M3s - 0.0128\ MR + \\ & + 0.0001\ te - 0.3571\ BCLU + 0.2042\ dip + 0.2807\ HARD - \\ & - 56.133\ De5 + 52.746\ Dn5 + 8.9055\ Dn4 - 1.24. \end{aligned} \tag{4.13}$$

• Model z devetimi deskriptorji:

$$\begin{aligned}\log k \;=\;& -0.1581\ \log P - 0.0030\ MM - 0.0053\ SA - 0.0063\ MV - \\ & - 0.0011\ IM2s - 0.0011\ IM3s - 0.0184\ MR + 0.0002\ te + \\ & + 0.2638\ dip + 6.6951. \end{aligned} \tag{4.14}$$

Korelacijski koeficienti modelov z regresijskimi drevesi (tabela 4.6) so slabši od koeficientov klasičnih modelov. Zato so bili zgrajeni tudi modeli, kjer je bil izbor deskriptorjev omejen na 15 oziroma 9 deskriptorjev, dobljenih s PCA analizo pri klasičnem modeliranju. Ker je maksimalno število deskriptorjev, ki jih lahko uporabi RETIS enako 30, sta bila z njim zgrajena le modela, ki sta upoštevala 15 oz. 9 deskriptorjev. Cubistova modela z vsemi in s 15. deskriptorji (nista prikazana) vsebujeta deskriptor *SA* (površina), za katerega je poleg *MV* (molekulski volumen) znano, da najbolje opisuje biotransformacijo dioksinov in furanov s *Sphingomonas sp.*, soj RW1. Kot zelo dober je bil ocenjen Cubistov

Model	PLS		Cubist							
			Privzeti parametri				Optimizirani parametri			
	R^2	Q^2	r	q	R^2	Q^2	r	q	R^2	Q^2
Model z vsemi deskriptorji	0.94	0.78	0.98	0.78	0.97	0.6	0.98	0.88	0.96	0.73
Model s 15. deskriptorji	0.95	0.88	0.93	0.82	0.85	0.64	0.92	0.88	0.83	0.75
Model z 9. deskriptorji	0.94	0.92	0.89	0.87	0.79	0.76	0.88	0.89	0.77	0.79

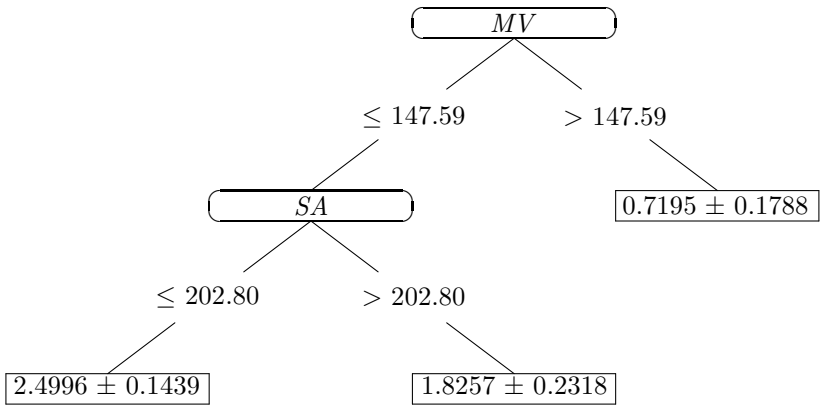
Model	PLS		RETIS							
			Regresija (m=1)				Brez regresije (m=1)			
	R^2	Q^2	r	q	R^2	Q^2	r	q	R^2	Q^2
Model z vsemi deskriptorji	0.94	0.78	–	–	–	–	–	–	–	–
Model s 15. deskriptorji	0.95	0.88	1.00	0.89	1.00	0.71	0.96	0.75	0.88	0.55
Model z 9. deskriptorji	0.94	0.92	1.00	0.55	1.00	-0.01	0.96	0.75	0.88	0.55

Tabela 4.6: Korelacijski koeficienti modelov za dioksine in furane.

model dobljen z naborom 9. deskriptorjev (enačba 4.15):

$$\log k = 7.651 - 0.0424\ MV. \tag{4.15}$$

Vsebuje le deskriptor MV, njegova odlika pa je preprostost. Skladen z dosedanjim znanjem je tudi RETISov model brez regresije s 15. ali 9. deskriptorji (slika 4.4), ki uporabi le oba pomembna deskriptorja.



Slika 4.4: Model za dioksine in furane (izbor 15. oz. 9. deskriptorjev) zgrajen s sistemom RETIS, brez regresije.

4.5 Biorazgradljivost haloalifatskih spojin

4.5.1 Uvod

Poznavanje biorazgradljivosti haloalifatskih spojin je zaradi njihove množične uporabe zelo pomembno. Z njimi smo se srečali že v razdelku 4.3. Tokrat nas zanimajo nekateri drugi predstavniki tega razreda spojin, pa tudi način ocenjevanja biorazgradljivosti je drugačen.

4.5.2 Množica podatkov

Množica vsebuje naslednjih 27 haloalifatskih spojin:

1-kloropropan,	1-klorobutan,	1-kloropentan,	1-kloroheksan,	1-kloroheptan,
1-klorooktan,	1-klorononan,	1-klorodekan,	1-klorododekan,	1-klorotetradekan,
1-kloroheksadekan,	1-klorooktadekan,	1-bromoetan,	1-bromobutan,	1-bromoheksan,
1-bromotetradekan,	1-jodobutan,	1-jodopentan,	1-jodoheksan,	1,1-diklorometan,
1,2-dikloroetan,	1,3-dikloropropan,	1,4-diklorobutan,	1,6-dikloroheksan,	1,9-diklorononan,
1,10-diklorodekan,	1,2-dibromoetan.			

Ocena dehalogenizacije je bila izvedena z nepoškodovanimi celicami bakterije *Rhodococcus erythropolis Y2*. Ciljna vrednost predstavlja naklon premice v grafu časovnega poteka koncentracije klorovih ionov. Iz skupine petih fizikalno-kemičnih in 15. kvantno-kemičnih deskriptorjev je bilo s pomočjo PCA analize izbranih devet deskriptorjev (tabela 4.7). Podatki so povzeti po [4].

Mw	molekulska teža	TE	skupna energija
IX	vztrajnostni moment okrog x osi	EE	energija elektronov
$\log P$	logaritem porazdelitvenega koeficienta oktanol/voda	$HOMO$	energija najvišje zasedene molekulske orbitale
Hf	toplota tvorbe	Dip	dipolni moment
		$BCLU$	prispevek $LUMO$ k vezi

Tabela 4.7: Molekulski deskriptorji haloalifatskih spojin.

4.5.3 Rezultati

Začetni PLS model (enačba 4.16) je upošteval vseh 27 spojin. Iz grafa napovedane dehalogenizacije, v odvisnosti od izmerjene (ni prikazan), so bile določene tri spojine (1,1-diklorometan, 1-kloropropan in 1,2-dikloroetan), ki so močno odstopale od dobljenega modela. Zaradi tega so tudi statistični pokazatelji tega modela (tabela 4.8) slabi. Model brez teh treh spojin (enačba 4.17) je bistveno boljši.

PLS modeli so podani v nadaljevanju.

• Model z vsemi (27.) spojinami:

$$\begin{aligned} Y2 \;=\; & -0.069\ Mw - 0.9045\ IX - 1.9837\ \log P + 0.1165\ Hf + \\ & + 0.0063\ TE + 0.0009\ EE - 1.4024\ HOMO - 6.1407\ Dip - \\ & - 1.2534\ BCLU + 106.16. \end{aligned} \tag{4.16}$$

• Model brez treh spojin:

$$\begin{aligned} Y2 \;=\; & -0.1692\ Mw - 1.0745\ IX - 3.8575\ \log P + 0.0708\ Hf + \\ & + 0.0059\ TE + 0.001\ EE - 37.817\ HOMO - 15.072\ Dip - \\ & - 673.68\ BCLU - 1581.5. \end{aligned} \tag{4.17}$$

Model	PLS		Cubist							
			Privzeti parametri				Optimizirani parametri			
	R^2	Q^2	r	q	R^2	Q^2	r	q	R^2	Q^2
Model z vsemi spojinami	0.34	0.20	0.55	0.12	0.30	-0.38	0.65	0.60	0.42	0.33
Model brez treh spojin	0.92	0.87	0.89	0.80	0.80	0.63	0.97	0.85	0.94	0.71
Model	PLS		RETIS							
			Regresija (m=1)				Brez regresije (m=1)			
	R^2	Q^2	r	q	R^2	Q^2	r	q	R^2	Q^2
Model z vsemi spojinami	0.34	0.20	0.94	0.01	0.88	-2.60	0.97	0.19	0.89	-0.13
Model brez treh spojin	0.92	0.87	0.99	0.88	0.98	0.74	0.96	0.74	0.88	0.54

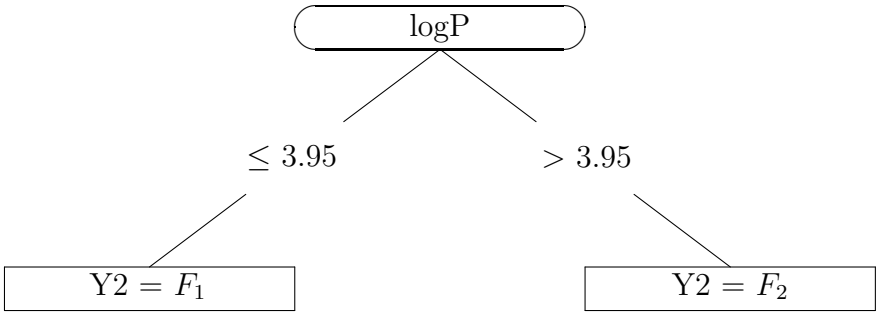
Tabela 4.8: Korelacijski koeficienti modelov za dehalogenizacijo haloalifatskih spojin.

Orodja za gradnjo regresijskih dreves nam za vseh 27 spojin niso dala veljavnih modelov (glej vrednosti Q^2 in q v tabeli 4.8). Regresijska drevesa so močnejši formalizem

od linearnih enačb in bi zato lahko pričakovali, da bodo orodja sposobna zgraditi drevo z modelom za 24 podobnih spojin v enem listu ter modeli za tri izstopajoče spojine v ostalih. Vzrok, da se to ni zgodilo, je verjetno v majhnem številu primerov (predvsem izstopajočih), ali pa je biorazgradnja odvisna od deskriptorjev, ki tu niso bili upoštevani. Modeli brez treh izstopajočih spojin so bistveno boljši. Cubistov model (enačba 4.18) ima sicer nižjo napovedno moč kot PLS model (enačba 4.17), vendar vsebuje le dva deskriptorja:

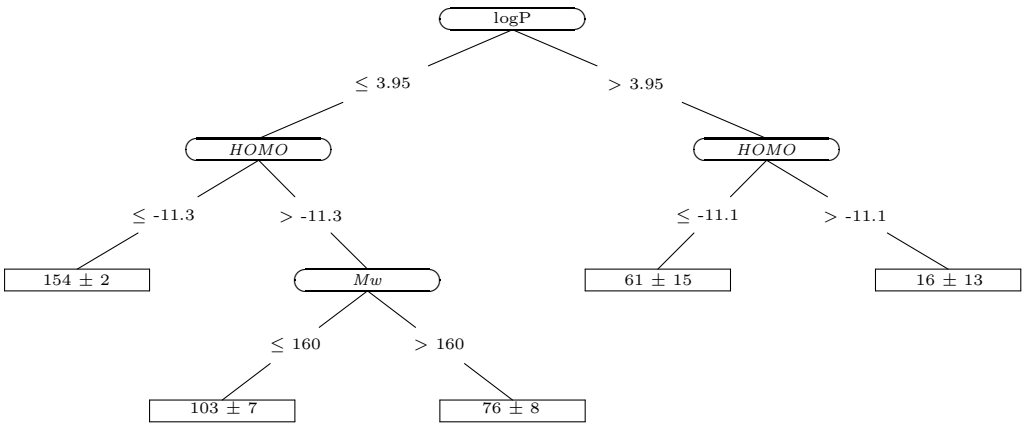
$$Y2 \;=\; -679.3 + 0.0562\ TE - 76.9\ HOMO. \tag{4.18}$$

Celotna energija TE je verjetno povezana z velikostjo molekule (pomembno za vezavo substrata na aktivno mesto encima), medtem ko energija najvišje zasedene molekulske orbitale $HOMO$ opisuje elektronske lastnosti, pomembne za reakcijo dehalogenizacije. Model je preprost, kar je dobro za interpretacijo, vendar so njegove napovedi manj natančne od napovedi kompleksnejšega modela PLS (enačba 4.17). Zaradi preprostosti je ekspertova ocena tega modela zelo dobra. Smiselna modela nam je dal tudi RETIS. Model z regresijo (slika 4.5) je podoben PLS modelu (enačba 4.17), model brez regresije (slika 4.6) pa je podoben Cubistovemu modelu (enačba 4.18). Statistični parametri (tabela 4.8) obeh RETISovih modelov so precej slabši od PLS modela (enačba 4.17).



$$
\begin{aligned}
F_1 &= -598 + 17.1704\, Mw + 12.7029\, IX - 257\, \log P - 66.8074\, Hf \\
&\quad +2.47860\, TE + 0.203530\, EE + 142.5142\, HOMO + 51\, Dip - 893\, BCLU \\
F_2 &= 4028 - 9.7202\, Mw + 1.9403\, IX + 1\, \log P + 12.6063\, Hf - 1.36163\, TE \\
&\quad +0.007442\, EE + 729.8901\, HOMO - 6\, Dip - 2161\, BCLU
\end{aligned}
$$

Slika 4.5: Model za dehalogenizacijo haloalifatskih spojin zgrajen z RETISom, z regresijo.



Slika 4.6: Model za dehalogenizacijo haloalifatskih spojin zgrajen z RETISom, brez regresije.

4.6 Biorazgradljivost mutantov haloalkanske dehalogenaze

4.6.1 Uvod

Za ugotavljanje povezave med strukturo proteinov in njihovo funkcijo se pogosto uporabljajo pozicijsko naravnani mutageni poskusi. Z naključnimi mutagenimi poskusi, ki jim sledi določitev funkcionalnosti mutantov, lahko dobimo informacije o pomembnosti določene aminokisline za aktivnost in stabilnost proteina. Funkcionalnost mutantov lahko določimo s QSFR (Quantitative Structure-Function Relationships) modeliranjem. Za razliko od primerov v prejšnjih razdelkih (QSBR, QSTR), ko smo gradili model za napoved neke lastnosti (preproste) spojine na osnovi deskriptorjev te spojine, tu gradimo model za napoved lastnosti (kompleksne) spojine (proteina) na osnovi deskriptorjev njenega sestavnega dela (aminokisline).

4.6.2 Množica podatkov

Zanima nas dehalogenizacija nespremenjene haloalkanske dehalogenaze (*Xanthobacter autotrophicus GJ10*) in njenih 15. enotočkovnih (pozicija 172) mutantov. Izmerjena je bila spektrofotometrično z detekcijo sproščenih halidnih ionov in je izražena kot konstanta prvega reda (k). Kot že rečeno, so za opis dehalogenizacije uporabljeni deskriptorji aminokislin. Posamezno lastnost aminokislin lahko opišemo na različne načine, npr. hidrofobičnost opisujejo deskriptorji *Ht, He, Hk, -Hp, Hf, Gh* in *Ha*. Eksaktna povezava med sorodnimi deskriptorji ni znana, zato na začetku upoštevamo vse. Vseh 33 deskriptorjev je naštetih v tabeli 4.9. Podatki so povzeti po [5].

4.6.3 Rezultati

Začetni model z vsemi deskriptorji dobljen s PLS analizo (ni prikazan) ima slabe statistične parametre (tabela 4.10). Drugi model (enačba 4.19) vsebuje le 14 deskriptorjev (*V, B, As, dG, Ht, -Hp, Fr, Fb, F0, R, Pr, Es, El, D*) z vrednostjo VIP večjo od 1. Z zožitvijo nabora deskriptojev na 4 (*B, F0, R, D*) je bil dobljen končni model (enačba 4.20).

V	povprečni volumen notranjih atomskih skupin
B	obsežnost
As	dosegljiva površina v standardnem stanju
Af	dosegljiva površina v zgibanem stanju
Ar	razmerje znižanja dostopnosti za topilo
Bc	odstotek notranjih atomskih skupin
Bj	odstotek notranjih atomskih skupin
Ej	odstotek izpostavljenih atomskih skupin
dG	prosta energija transporta na površino
Ht	indeks hidrofobnosti
He	indeks hidrofobnosti
Hk	hidropatičnost
$-Hp$	indeks hidrofilnosti
Hf	hidrofobnost okolja v zgibanem stanju
Gh	razmerje povečanja hidrofobnosti
Ha	hidrofobnost okolja v stanju alfa vijačnice
Fr	lokalna fleksibilnost
Fb	indeks fleksibilnosti
$F0$	verižna fleksibilnost
Ca	normalizirana pogostost pojavljanja alfa vijačnice
Cb	normalizirana pogostost pojavljanja beta lista (beta-sheet)
Cr	normalizirana pogostost pojavljanja nasprotnega obrata (reverse turn)
Q	formalni naboj na stranski verigi
R	refraktivnost
Pe	faktor polarnosti
Pz	indeks polarnosti
Pr	rang polarnosti
pI	izoelektrična točka
pK	negativni logaritem ionizacijske konstante za α-karboksile
Et	celotna nevezna energija okolja proteina
Es	nevezna energija kratkega in srednjega dosega
El	nevezna energija dolgega dosega
D	aromatski obroč

Tabela 4.9: Deskriptorji aminokislin.

PLS modeli so podani v nadaljevanju.

• Model s 14. deskriptorji:

$$\begin{aligned} k &= 0.0003\,V - 0.1086\,B + 0.0007\,As - 0.1395\,dG + 0.1479\,Ht + \\ &\quad + 0.0255\,Hp + 0.0000\,Fr - 3.1907\,Fb - 0.3449\,F0 + 0.0352\,R + \\ &\quad + 0.0169\,Pr - 0.9454\,Es + 0.7118\,El + 3.2800\,D + 3.3703 \end{aligned} \tag{4.19}$$

• Model s štirimi deskriptorji:

$$k = -0.0688\,B - 1.7766\,F0 + 0.0590\,R + 3.4190\,D + 2.3347. \tag{4.20}$$

Model	PLS		Cubist							
			Privzeti parametri				Optimizirani parametri			
	R^2	Q^2	r	q	R^2	Q^2	r	q	R^2	Q^2
Model z vsemi deskriptorji	0.50	0.35	0.84	0.35	0.71	-0.08	0.84	0.35	0.71	-0.08
Model s 14. deskriptorji	0.86	0.60	–	–	–	–	–	–	–	–
Model s štirimi deskriptorji	0.84	0.75	0.84	0.46	0.71	0.07	0.84	0.48	0.71	0.04
Model	PLS		RETIS							
			Regresija (m=1)				Brez regresije (m=1)			
	R^2	Q^2	r	q	R^2	Q^2	r	q	R^2	Q^2
Model z vsemi deskriptorji	0.50	0.35	–	–	–	–	–	–	–	–
Model s 14. deskriptorji	0.86	0.60	1.00	0.14	1.00	-48	0.99	0.16	0.89	-0.30
Model s štirimi deskriptorji	0.84	0.75	1.00	0.74	0.99	0.49	0.99	0.32	0.89	0.01

Tabela 4.10: Korelacijski koeficienti modelov za mutante haloalkanske dehalogenaze.

S Cubistom so bili zgrajeni modeli z upoštevanjem vseh deskriptorjev ter z upoštevanjem le štrih deskriptorjev (*B, F0, R, D*). Noben model nima uporabne vrednosti. Zaradi omejitve RETISa na 30 deskriptorjev, so bili zgrajeni modeli z upoštevanjem 14. in 4. deskriptorjev. Korelacijski koeficienti (tabela 4.10) kažejo, da tudi ti modeli niso uporabni.

4.7 Aktivnost in stabilnost namensko spremenjenih proteinov

4.7.1 Uvod

Eden izmed končnih ciljev proteinskega inženiringa je načrtovanje in gradnja encimov z želenimi lastnostmi, za katalizo novih reakcij, ali za katalizo že znanih reakcij z večjo učinkovitostjo. Gradnja encimov z načrtovanimi zamenjavami aminokislin nam omogoča oceno vloge posamezne aminokisline v katalitski aktivnosti ali stabilnosti. Korak k temu cilju je QSFR (Quantitative Structure-Function Relationships) in QSSR (Quantitative Structure-Stability Relationships) modeliranje. S prvim smo se srečali že v prejšnjem razdelku, drugo pa opisuje povezave med strukturo proteinov in njihovo stabilnostjo. Rezultat QSFR/QSSR modeliranja je matematični opis vpliva strukturnih sprememb proteina na funkcijo/stabilnost proteina.

Obravnavani so štirje primeri analize podatkov o proteinih, dobljenih s pozicijsko usmerjenimi mutagenimi poskusi. Proučene so bile količinske strukturno–funkcijske povezave (QSFR) za 15 mutantov haloalkanske dehalogenaze na poziciji 172 (Dhla-Phe172) in za 19 mutantov subtilizina na poziciji 222 (Subt-Met222). Za 13 mutantov lizozima faga T4 na poziciji 157 (Lyso-Thr175) in 18 mutantov α-podenot triptofan sintaze na poziciji 49 (Synth-Glu49), so bile proučene strukturno–stabilnostne povezave (QSSR). Prva skupina proteinov je ista kot v prejšnjem razdelku (4.6), vendar so uporabljeni drugi deskriptorji za opis aminokislin. V drugi skupini so mutanti subtilizina, ki je eden najbolj proučevanih encimov (uporaben v industriji). Mutanti subtilizina so bili predmet prvega ameriškega patenta na področju proteinskega inženiringa. Lizozim faga T4 je drug primer proteina, množično uporabljanega v proučevanju vpliva zamenjave aminokislin na funkcionalnost in stabilnost proteinov; njegovi mutanti so v tretji skupini. V zadnji skupini so mutanti α-podenot triptofan sintaze, ki je proizvod bakterije *Escherichia coli*.

4.7.2 Množica podatkov

Baza podatkov o aminokislinah *AAindex* (dostopna na `www.genome.ad.jp`) vsebuje 402 deskriptorja. Iz te baze je bilo s PCA analizo določenih devet najpomembnejših deskriptorjev (tabela 4.11) za opis strukture in lastnosti 20. naravnih aminokislin. Aktivnost mutantov haloalkanske dehalogenaze je izražena kot dehalogenizacija 1,2-dibromoetana,

A362	glavna komponenta IV	*P159*	van der Waalsov parameter ε po Levittu
C144	fleksibilnostni parameter za dva toga soseda	*P214*	nevezna energija kratkega in srednjega dosega
H132	indeks hidrofobnosti	*P383*	R_f vrednost v visoko solni kromatografji (high salt cromatography)
H311	povprečna reducirana razdalja za C_α	*P399*	obsežnost
H377	relativna populacija prilagoditvenega stanja C		

Tabela 4.11: Deskriptorji aminokislin.

aktivnost mutantov subtilizina pa je bila ocenjena z oligopeptidno verigo (N-sukcinl-L-Ala-L-Ala-L-Pro-L-Phe-p-nitroanilid). Termodinamična stabilnost mutiranih lizozimov je

podana z Gibbsovo prosto energijo ($\Delta\Delta G$), stabilnost mutantov podenot triptofan sintaze pa z Gibbsovo prosto energijo razvijanja ($\Delta_d G$). Podatki so povzeti po [6], kjer je tudi več podrobnosti o načinu merjenja omenjenih lastnosti proteinov.

4.7.3 Rezultati

Kot že rečeno, je bil prvi korak PLS modeliranja izbira 9. najpomembnejših deskriptorjev iz obširne množice 402 deskriptorjev. Z nadaljnjo uporabo PCA analize so bili nato določeni deskriptorji uporabljeni v vseh štirih modelih (enačbe 4.21–4.24).

PLS modeli so podani v nadaljevanju.

• QSFR model za mutante haloalkanske dehalogenaze na poziciji 172:

$$k = 4.985\ \mathit{P159} - 5.033\ \mathit{P383} - 0.1397\ \mathit{P399} + 5.3994. \tag{4.21}$$

• QSFR model za mutante subtilizina na poziciji 222:

$$\log k = 1.4958\ \mathit{A362} - 3.6129\ \mathit{P214} - 3.4051. \tag{4.22}$$

• QSSR model za mutante faga T4 lizozima na poziciji 157:

$$\Delta\Delta G = -15.171\ \mathit{C144} + 0.3032\ \mathit{H132} - 1.2957\ \mathit{H377} + 15.4420. \tag{4.23}$$

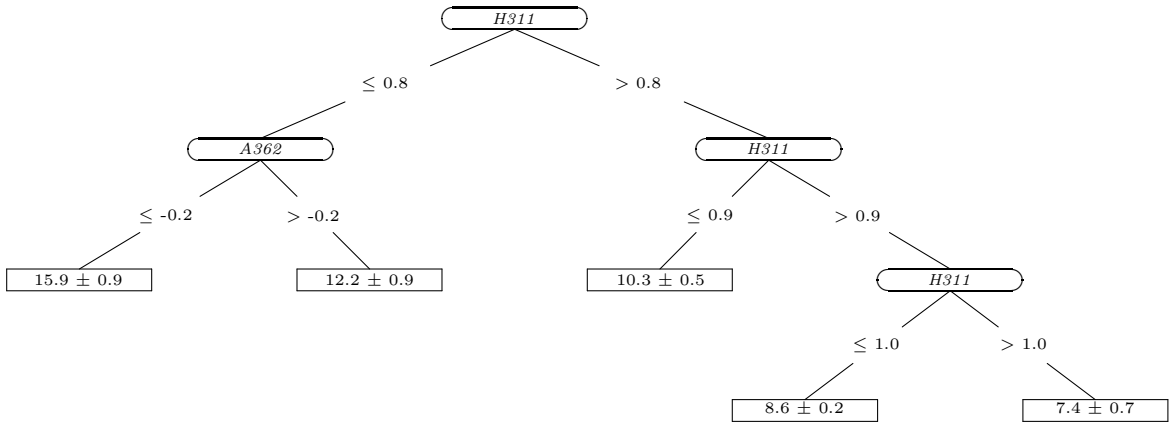
• QSSR model za mutante α-podenot triptofan sintaze na poziciji 49:

$$\Delta_d G = -24.036\ \mathit{H311} + 32.263. \tag{4.24}$$

Možnosti modeliranja z regresijskimi drevesi so bile omejene, ker je bila dostopna le množica podatkov z 9. deskriptorji. V teh pogojih je bil s Cubistom dobljen model (enačba 4.25) za Synth-Glu49 skoraj identičen PLS modelu (enačba 4.24):

$$\Delta_d G = 32.29 - 24.1\ H311. \tag{4.25}$$

Tudi model za Lyso-Thr175, zgrajen s Cubistom (ni prikazan) je zelo podoben PLS modelu (enačba 4.23). Med RETISovimi modeli omenimo model za Synth-Glu49 brez regresije (slika 4.7). Vsebuje deskriptor A362, ki je verjetno odveč. Ostali modeli niso presegli praga statistične pomembnosti. Napovedna moč vseh modelov z regresijskimi



Slika 4.7: RETISov model brez regresije za mutante α-podenot triptofan sintaze.

Model	PLS		Cubist							
			Privzeti parametri				Optimizirani parametri			
	R^2	Q^2	r	q	R^2	Q^2	r	q	R^2	Q^2
Dhla-Phe172	0.83	0.77	0.95	0.60	0.91	0.28	0.93	0.66	0.87	0.38
Subt-Met222	0.86	0.81	0.70	0.25	0.46	0.01	0.90	0.33	0.80	0.06
Lyso-Thr175	0.87	0.85	0.93	0.70	0.87	0.47	0.93	0.73	0.87	0.52
Synth-Glu49	0.76	0.71	0.87	0.81	0.76	0.65	0.89	0.85	0.79	0.71
Model	PLS		RETIS							
			Regresija (m=1)				Brez regresije (m=1)			
	R^2	Q^2	r	q	R^2	Q^2	r	q	R^2	Q^2
Dhla-Phe172	0.83	0.77	0.99	0.76	0.98	0.19	0.99	0.34	0.89	0.08
Subt-Met222	0.86	0.81	0.99	0.45	0.98	-0.13	0.96	0.14	0.76	-0.28
Lyso-Thr175	0.87	0.85	0.99	0.58	0.98	-0.21	0.94	0.38	0.81	0.05
Synth-Glu49	0.76	0.71	0.99	0.67	0.97	0.33	0.97	0.80	0.87	0.63

Tabela 4.12: Korelacijski koeficienti namensko spremenjenih proteinov.

drevesi je bistveno slabša od PLS modelov (tabela 4.12). Če upoštevamo tudi rezultate iz prejšnjega razdelka lahko sklepamo, da regresijska drevesa niso najbolj primerna za QSFR/QSSR modeliranje proteinskih mutantov, ker je največje možno število primerov enako 20. Vprašanje je, kako bi se obnesla na množici podatkov z vsemi 402. deskriptorji.

4.8 Biorazgradljivost haloalkenov

4.8.1 Uvod

Povezave med strukturo organskih spojin in njihovo mikrobno razgradnjo smo si ogledali že v razdelku 4.5. Tukaj pa je opisan primer QSBR modeliranja mikrobne hidrolitične dehalogenizacije preprostih kloro in bromoalkenov s hidrolitičnimi dehalogenazami. Področje je še posebej zanimivo (in zahtevno) zaradi različnih mehanizmov degradacije, ki pri tem nastopajo. Dober model lahko namreč dobimo le za en mehanizem. Nastopi problem določitve tistih spojin, pri katerih je proučevani mehanizem najboj izrazit ali vsaj prevladujoč.

4.8.2 Množica podatkov

Začetna množica je vsebovala 83 spojin. S PCA analizo jih je bilo nato 24 izbranih za testiranje dehalogenizacije. Merjenje je bilo izvedeno s celicami *R. erythropolis*, soj Y2, oz. haloalkansko dehalogenazo, ki so jo le te proizvedle. Izbranih je bilo le 13 spojin z znatno stopnjo dehalogenizacije. Vsaka spojina je opisana s 25. deskriptorji (tabela 4.13). Že v tej fazi je bilo število deskriptorjev zmanjšano na 18 (v tabeli 4.13 nimajo zvezdice). Podatke in podrobnosti o merjenju najdemo v [7].

4.8.3 Rezultati

Prvotni PLS model, ki je upošteval vseh 13 spojin (ni prikazan) ni imel uporabne vrednosti. Učno množico končnega modela (enačba 4.26) je sestavljalo le 8 izbranih spojin. Model temelji na 5. deskriptorjih, izbranih na osnovi PCA analize in VIP vrednosti:

$$\begin{aligned} \log k &= -0.004\, Mw - 0.2335\, DIP + 0.5352\, LUMO - 61.768\, AV1 + \\ &\quad + 4.5468\, q1 + 244.460. \end{aligned} \tag{4.26}$$

Modeli z regresijskimi drevesi so bili grajeni na osnovi množice 13. spojin, opisanih s 25. deskriptorji. Noben od modelov nima uporabne vrednosti. Omenimo le Cubistov model (slika 4.8), ki sicer vsebuje smiselne deskriptorje, vendar so vrednosti Q^2 in q zelo nizke (tabela 4.14).

Mw	molekulska teža
logP	logaritem porazdelitvenega koeficienta oktanol-voda
HF	toplota tvorbe
EE	* energija elektronov
DIP	dipolni moment
BCLU	* prispevek LUMO k vezi
BCHO	* prispevek HOMO k vezi
HARD	trdota
HOMO	enegija najvišje zasedene molekulske orbitale
LUMO	energija najnižje nezasedene molekulske orbitale
CX	dolžina vezi ogljik–halogen
AV1	* atomska valenca atoma C_1
BO1-X	red vezi C_1–halogen
BO1-2	red vezi C_1–C_2
BO2-3	red vezi C_2–C_3
DN1	nukleofilna krajevna nedoločenost atoma C_1
DN3	nukleofilna krajevna nedoločenost atoma C_3
Q1	* delni atomski naboj na atomu C_1
Q3	* delni atomski naboj na atomu C_3
av1	atomska valenca atoma C_1
av3	atomska valenca atoma C_3
bv1-x	* vezna valenca vezi med atomom C_1 in halogenom
q1	delni atomski naboj na atomu C_1
q3	delni atomski naboj na atomu C_3
qx	delni atomski naboj na atomu X

Tabela 4.13: Deskriptorji haloalkanov. Deskriptorji označeni z * niso bili uporabljeni pri razvoju PLS modela.

Pravilo 1:

če

$av1 > 3.9089$

potem

$\log k = 173.154 - 44.1\ av1$

Pravilo 2:

če

$av1 \leq 3.9089$

potem

$\log k = 146.787 + 7.05\ BCLU + 0.71\ \log P - 34\ av1$

Slika 4.8: Cubistov model za dehalogenizacijo haloalkenov.

Orodje	r	q	R^2	Q^2
PLS	–	–	0.92	0.81
Cubist (privzeti parametri)	0.88	-0.54	0.78	-1.37
Cubist (optimizirani parametri)	0.88	-0.54	0.78	-1.37
RETIS (regresija, m=1)	1.00	0.10	1.00	-177
RETIS (brez regresije, m=1)	0.93	-0.55	0.80	-1.39

Tabela 4.14: Korelacijski koeficienti modelov za dehalogenizacijo haloalkanov.

4.9 Biorazgradljivost komercialnih kemičnih spojin

4.9.1 Uvod

Največkrat se QSAR modeliranje uporablja za napovedovanje aktivnosti ali lastnosti skupine spojin, ki pripadajo istemu razredu (ali manjšemu številu le teh) ali so si kako drugače sorodne (npr. fenoli). Ta pristop je bil uporabljen tudi v prejšnjih razdelkih (4.2–4.8). V tem razdelku obravnavana množica podatkov je sestavljena iz večih družin spojin. Vsebuje npr. alkohole, fenole, pesticide, klorove alifatske in aromatske ogljikovodike, kisline, ketone, etre in druge vrste spojin. Seveda je modeliranje tako raznolike skupine spojin precej zahtevnejša naloga. Tako dobljeni model naj bi omogočal napovedovanje stopnje biorazgradljivosti za široko paleto spojin, ne bo pa prinesel dodatnega znanja o posameznih mehanizmih biodegradacije, saj so ti pri posameznih tipih spojin različni in jih en sam model ne more opisati. V našem primeru je to pogojeno tudi z ocenami stopnje biorazgradljivosti v množici podatkov, ki ne ločijo med posameznimi mehanizmi.

4.9.2 Množica podatkov

Uporabljeno množico podatkov so sestavili avtorji [8]. Degradacijske stopnje 342. množično uporabljanih spojin so bile zbrane iz literature. Glavni vir je bil *Handbook of Environmental Degradation Rates* (Howard et al. 1991), ki za primere, kjer ni izmerjenih podatkov, upošteva ekspertne ocene, ki so lahko pristranske. Za vsako spojino najdemo tu degradacijsko stopnjo v obliki razpolovnega časa (spodnja in zgornja ocena) za skupno,

biotsko in abiotsko degradacijo v štirih okoljih (prst, zrak, površinske vode in podtalnica). Avtorji [8] so se osredotočili na biorazgradnjo v površinskih vodah, kjer živi organizmi niso prilagojeni onesnaženju s proučevano spojino. Prvotno so bili razpolovni časi podani v urah, dnevih, tednih in mesecih. Za nadaljnjo obdelavo so bila izračunana povprečja spodnjih in zgornjih ocen v urah, za vodno biorazgradnjo v aerobnoh pogojih; uporabljen je bil naravni logaritem teh vrednosti.

Za 328 od 342. spojin je bila dobljena reprezentacija v obliki atom–vez, na osnovi katere so bili določeni predikati, ki določajo podstrukture (funkcijske skupine) pomembne za biorazgradljivost (vsi deskriptorji v tabeli 4.15, razen zadnjih dveh). Funkcijske skupine

nitro	*carboxylic_acid*	*hetero_ar_6_ring*
sulfo	*ester*	*non_ar_6c_ring*
methyl	*amide*	*non_ar_hetero_6_ring*
methoxy	*imine*	*six_ring*
amine	*alkyl_halide*	*carbon_5_ar_ring*
aldehyde	*ar_halide*	*non_ar_5c_ring*
ketone	*epoxy*	*non_ar_hetero_5_ring*
ether	*n2n*	*five_ring*
sulfide	*c2n*	*logP*
alcohol	*benzene*	*mweight*
phenol		

Tabela 4.15: Deskriptorji množice podatkov P1.

so bile določene na podlagi poprejšnjega znanja o obravnavanem problemu. Tako določeni predikati, skupaj z molekulsko težo (*mweight*) in logaritmom porazdelitvenega koeficienta oktanol/voda (*logP*), so bili osnovna reprezentacija kemikalij. Iz nje sta bili dobljeni dve množici podatkov. Prva (P1) vsebuje poleg *logP* in *mweight* še število vsake od omenjenih funkcijskih skupin v spojini (tabela 4.15), skupno 31 deskriptorjev. Deskriptorji druge množice (P2) so bili dobljeni s štetjem vseh podstruktur z dvema ali tremi atomi ter podstruktur s štirimi atomi zvezdaste topologije (brez verig). Upoštevane so bile vse podstrukture (vsi deskriptorji v tabeli 4.16, razen zadnjih dveh), ki so bile prisotne v vsaj treh spojinah. Oznake v tabeli 4.16 pomenijo strukture oblike Atom1–Atom2, Atom1–Atom2–Atom3, Atom1=(Atom2,Atom4)–Atom3. Atom2 je vedno centralni atom, na katerega se vežejo vsi ostali. Številke v oznakah pomenijo vrsto vezi med atomi (1 – enojna vez,

br1c	*h1n*	*c1n1c*	*h1n1h*	
br1h	*h1o*	*c1n2c*	*o1p1s*	
c1cl	*n1o*	*c1n1h*	*o1p2s*	*c1n1h1h*
c1f	*n2o*	*c7n7c*	*o1p1o*	*c1s2o2o*
c1h	*o1p*	*c1o1p*	*o1p2o*	*o1p1s2s*
c1n	*o2p*	*c1o1s*	*o1s1o*	*o1p1o1s*
c2n	*o1s*	*c1o1c*	*o1s2o*	*o1p1o2s*
c3n	*o2s*	*c1o1h*	*o2s2o*	*o1s1o2o*
c7n	*p1s*	*c1o1n*	*br1c1h1h*	*o1s2o2o*
c1o	*p2s*	*c1s1c*	*br1c1br1h*	*mweight*
c2o	*br1c1h*	*c1s2o*	*c1n2o2o*	*logP*
c1s	*br1c1br*	*c1s1p*	*c1n1c1c*	
c2s	*c1n2o*	*h1n2c*	*c1n1c1h*	

Tabela 4.16: Deskriptorji množice podatkov P2.

2 – dvojna vez, 3 – trojna vez in 7 – aromatska vez). Deskriptorji množice P2 so število vsake od podstruktur ter *logP* in *mweight*, skupno torej 61 deskriptorjev.

4.9.3 Rezultati

Za obe množici (P1 in P2) je PLS analizo opravil J. Damborsky. Za vsako je zgradil dva modela. Prva vsebujeta vse deskriptorje (nista prikazana), druga dva pa upoštevata le deskriptorje izbrane na osnovi vrednosti VIP (enačba 4.27 za P1 in enačba 4.28 za P2).

$$\begin{aligned}\log HLT &= 5.651 + 0.101\ alkyl\_halide + 0.1286\ ar\_halide + 0.1530\ benzene + \\ &+ 0.1615\ six\_ring + 0.8462\ carbon\_5\_ar\_ring + 0.1679\ five\_ring + \\ &+ 0.1107\ \log P + 0.001746\ mweight; \end{aligned} \tag{4.27}$$

$$\begin{aligned}\log HLT &= 6.040 + 0.1542\ \log P + 0.002432\ mweight + 0.1797\ c1cl - \\ &- 0.1108\ c1o - 0.3223\ c2o - 0.3059\ h1o + 0.09968\ c1n2o - \\ &- 0.2960\ c1o1h. \end{aligned} \tag{4.28}$$

Korelacijski koeficienti vseh štirih modelov so v tabelah 4.17 in 4.18. Velja omeniti, da pri gradnji teh modelov ni bilo upoštevano dodatno strokovno znanje o problemu, tako da bi bilo modele mogoče še izboljšati.

Orodje	r	q	R^2	Q^2
Cubist (privzeti parametri)	0.755	0.665	0.569	0.439
RETIS (regresija, m=1)	0.838	0.592	0.702	0.292
RETIS (regresija, m=5)	0.776	0.579	0.601	0.296
RETIS (brez regresije, m=1)	0.948	0.562	0.859	0.276
RETIS (brez regresije, m=8)	0.650	0.575	0.411	0.331
PLS (pred izbiro deskriptorjev)	0.562	–	0.316	0.285
PLS (po izbiri deskriptorjev)	0.522	–	0.272	0.257
M5'	–	0.666*	–	–

Tabela 4.17: Korelacijski koeficienti modelov za množico P1.

Orodje	r	q	R^2	Q^2
Cubist (privzeti parametri)	0.767	0.630	0.587	0.381
RETIS (regresija)	Hrošč!	–	–	–
RETIS (brez regresije, m=1)	0.945	0.604	0.862	0.339
RETIS (brez regresije, m=11)	0.694	0.606	0.456	0.360
PLS (pred izbiro deskriptorjev)	0.585	–	0.343	0.296
PLS (po izbiri deskriptorjev)	0.601	–	0.361	0.349
M5'	–	0.693*	–	–

Tabela 4.18: Korelacijski koeficienti modelov za množico P2.

S Cubistom sta bila zgrajena dva modela: model na sliki 4.9 za množico P1 in model na sliki 4.10 za množico P2. Zaradi omejitve demonstracijske verzije paketa Cubist na 200 učnih primerov, je modeliranje na »polni« različici programa opravil A. Kobler. Uporabljeni so bili privzeti parametri. Zgrajena modela vsebujeta relativno malo pravil za tako raznolično množico spojin. Zato ju je strokovnjak ocenil kot zelo dobra. Med vsemi zgrajenimi modeli imata največjo napovedno moč (tabeli 4.17 in 4.18).

Zaradi omejitve sistema RETIS na primere z največ 30 deskriptorji, modeliranje celotnih množic P1 in P2 ni bilo mogoče. Modeli so bili zgrajeni z deskriptorji, izbranimi na naslednji način. Iz vsake množice je bilo 10-krat naključno izbranih po 197 primerov

```
Rule 1: [95 cases, mean 5.630266, range 2.140066 to 7.822445, est err 1.210264]
  if  alkyl_halide <= 1
      mweight <= 110.971
  then  target = 5.29 - 0.947 benzene + 0.00915 mweight - 1.22 ester + 0.15 logP
                - 0.76 phenol - 0.64 alcohol - 1.28 ketone - 0.75 aldehyde
                + 0.108 ar_halide + 0.052 alkyl_halide + 0.14 nitro
                + 0.046 six_ring + 0.07 amine - 0.09 non_ar_6c_ring
                - 0.09 carboxylic_acid + 0.05 methoxy + 0.08 imine
                + 0.013 methyl + 0.01 five_ring

Rule 2: [7 cases, mean 6.079297, range 6.040255 to 6.313548, est err 0.000741]
  if  alkyl_halide > 1
      logP <= 1.43
  then  target = 5.494 + 0.273 alkyl_halide

Rule 3: [139 cases, mean 6.484813, range 4.533674 to 9.054388, est err 1.078450]
  if  alkyl_halide <= 1
      logP <= 4.84
      mweight > 117.107
  then  target = 6.248 + 0.43 ar_halide + 0.42 amine + 0.54 nitro
                - 0.00195 mweight + 0.052 alkyl_halide + 0.046 six_ring
                + 0.021 logP - 0.11 phenol - 0.08 ester - 0.09 non_ar_6c_ring
                + 0.05 methoxy - 0.018 benzene + 0.08 imine
                - 0.06 carboxylic_acid - 0.08 aldehyde + 0.013 methyl
                + 0.01 five_ring

Rule 4: [41 cases, mean 7.551500, range 4.564348 to 9.768354, est err 1.767965]
  if  alkyl_halide > 1
      logP > 1.43
      logP <= 4.84
  then  target = 7.515 - 0.00647 mweight + 0.34 alkyl_halide + 0.142 ar_halide
                + 0.129 six_ring + 0.059 logP - 0.098 benzene + 0.2 nitro
                - 0.26 non_ar_6c_ring + 0.12 amine - 0.15 phenol + 0.12 methoxy
                + 0.21 imine - 0.1 ester - 0.21 aldehyde + 0.032 methyl
                + 0.01 five_ring

Rule 5: [5 cases, mean 7.831836, range 7.822445 to 7.869402, est err 0.035290]
  if  alkyl_halide <= 1
      mweight > 110.971
      mweight <= 117.107
  then  target = 7.832

Rule 6: [41 cases, mean 8.624575, range 5.81413 to 11.27416, est err 2.209445]
  if  logP > 4.84
  then  target = 8.385 - 0.549 methyl + 0.145 alkyl_halide + 0.144 ar_halide
                - 0.0014 mweight + 0.101 six_ring + 0.049 logP + 0.17 nitro
                - 0.2 non_ar_6c_ring + 0.08 amine + 0.046 benzene - 0.15 phenol
                - 0.1 ester + 0.11 methoxy + 0.19 imine + 0.07 five_ring
                - 0.19 aldehyde
```

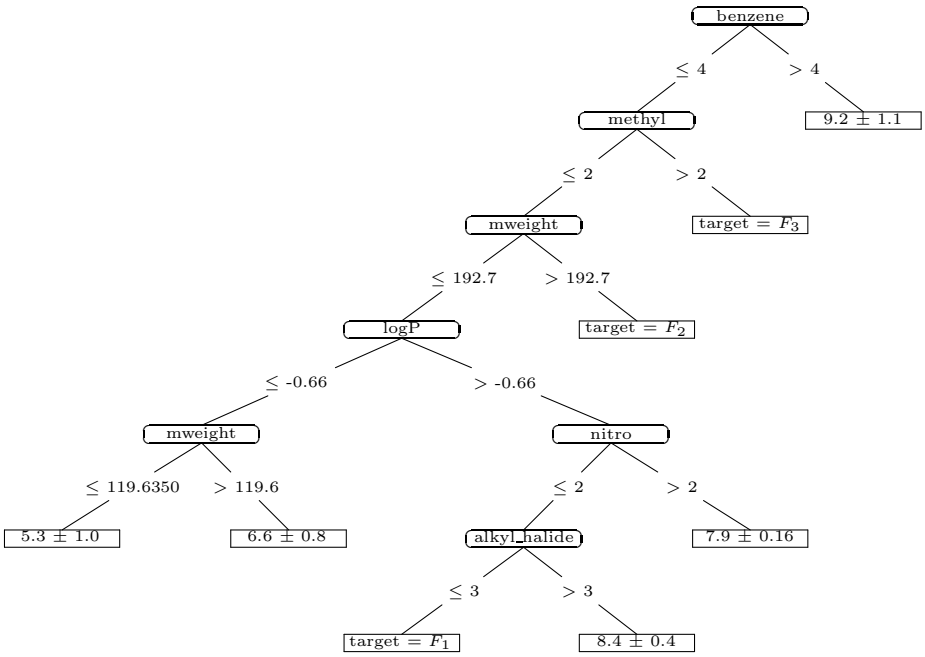
Slika 4.9: Cubistov model za množico P1.

(spojin). Za vsakih 10 (pod)množic so bili zgrajeni modeli s Cubistom. Vsi deskriptorji, ki so se vsaj enkrat pojavili v teh modelih, so bili nato uporabljeni za gradnjo regresijskih dreves z RETISom. Drevesa, dobljena z minimalnim naknadnim rezanjem (m=1), so bila zelo obsežna. Ta drevesa so bila naknadno toliko porezana, da so imela največ 8 listov. S tem smo se izognili preveliki prilagojenosti modela učnim primerom in omogočili strokovno interpretacijo modelov. Zgrajeni so bili modeli z in brez uporabe regresijskih

```
Rule 1: [81 cases, mean 5.697934, range 4.102643 to 8.427706, est err 0.872088]
  if  logP <= 4.88
      c2o > 0
  then  target = 4.982 + 0.449 c1n - 0.381 c1n1c + 0.00367 mweight - 0.094 c1o
                + 0.2 c1o1c + 0.057 c1cl - 0.12 c2o - 0.037 h1n

Rule 2: [13 cases, mean 5.947746, range 4.564348 to 6.313548, est err 0.394327]
  if  logP <= 1.43
      c1cl > 0
  then  target = 6.248 - 0.178 logP - 0.12 c1o1c

Rule 3: [115 cases, mean 6.054388, range 2.140066 to 8.785692, est err 1.144239]
  if  logP <= 4.88
      c1cl <= 0
      c2o <= 0
      n2o <= 0
  then  target = 6.019 + 0.36 c1o1c - 0.141 c1o + 0.085 c1n + 0.061 c1cl - 0.07 c2o
                - 0.044 h1n + 0.00041 mweight + 0.038 n2o - 0.025 c1n1c
                - 0.04 c1o1h

Rule 4: [13 cases, mean 6.270995, range 4.533674 to 7.822445, est err 15.848075]
  if  mweight > 135.637
      c1cl <= 0
      n2o > 0
      c1n2o <= 2
  then  target = 6.419 + 0.004 mweight - 0.265 c1o - 0.275 n2o - 0.116 logP
                + 0.102 c1n2o + 0.12 c1o1c + 0.03 c1cl + 0.035 c1n - 0.07 c2o
                - 0.025 c1n1c - 0.016 h1n

Rule 5: [9 cases, mean 6.970921, range 5.81413 to 10.0903, est err 1.378385]
  if  logP > 4.88
      c1h > 15
  then  target = 11.779 - 0.0136 mweight - 0.015 c1h

Rule 6: [65 cases, mean 7.331427, range 4.564348 to 9.768354, est err 1.229769]
  if  logP > 1.43
      logP <= 4.88
      c1cl > 0
  then  target = 7.263 + 0.41 c1o1c - 0.173 c1o + 0.098 c1cl + 0.08 c1n - 0.07 c2o
                - 0.043 h1n + 0.00041 mweight + 0.039 n2o - 0.025 c1n1c

Rule 7: [13 cases, mean 7.555413, range 4.564348 to 7.833996, est err 0.373306]
  if  mweight <= 135.637
      n2o > 0
  then  target = 7.801

Rule 8: [11 cases, mean 7.661757, range 5.817111 to 8.285766, est err 7.385338]
  if  logP <= 4.88
      c1n2o > 2
  then  target = 6.61 - 0.314 c1o + 0.52 c1o1c + 0.0026 mweight + 0.142 c1n
                - 0.076 logP + 0.14 c1n2o + 0.054 c1cl - 0.074 h1n - 0.11 c1o1h
                - 0.027 c1n1c - 0.04 c2o - 0.014 n2o

Rule 9: [29 cases, mean 9.405121, range 6.870573 to 11.27416, est err 1.013474]
  if  logP > 4.88
      c1h <= 15
      c1n <= 1
  then  target = 9.774 - 0.147 c1n - 0.029 c1h - 0.0006 mweight
```

Slika 4.10: Cubistov model za množico P2.

modelov v listih, razen v primeru množice P2, ko regresijski model ni bil zgrajen zaradi programskega hrošča v paketu RETIS. Minimalno porezani modeli (m=1) so preveč pri-

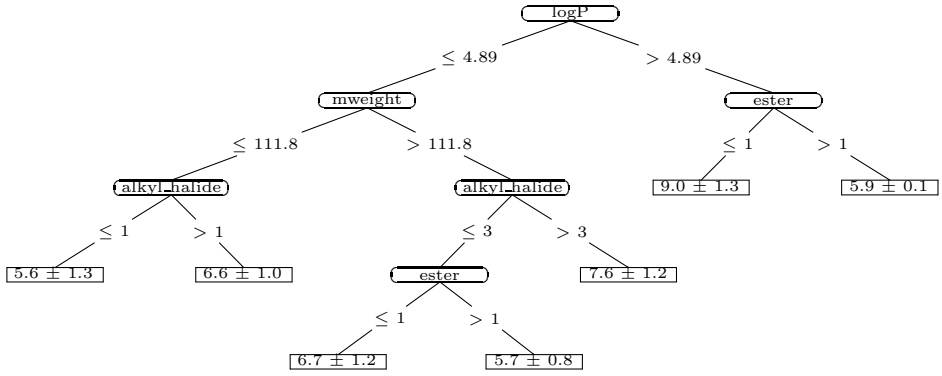


$$
\begin{aligned}
F_1 &= 5.94 + 0.056\ methyl + amine - aldehyde - alcohol - phenol - ester + 0.52\ alkyl\_halide + ar\_halide \\
&\quad -0.86\ benzene - non\_ar\_6c\_ring + 0.50\ six\_ring + carbon\_5\_ar\_ring + 0.26\ \log P - 0.0033\ mweight \\
F_2 &= 5.42 - 0.22\ methyl + 65385884\ aldehyde + 0.25\ alkyl\_halide + 0.045\ benzene + 0.25\ six\_ring \\
&\quad +2\ carbon\_5\_ar\_ring + 0.25\ \log P - 0.0012\ mweight \\
F_3 &= 5.47 + 0.14\ methyl - aldehyde - ketone - alcohol - ester + 0.14\ alkyl\_halide - ar\_halide \\
&\quad -41.59\ benzene - 42\ non\_ar\_6c\_ring + 41.43\ six\_ring + 17326544\ carbon\_5\_ar\_ring + 0.0038\ mweight
\end{aligned}
$$

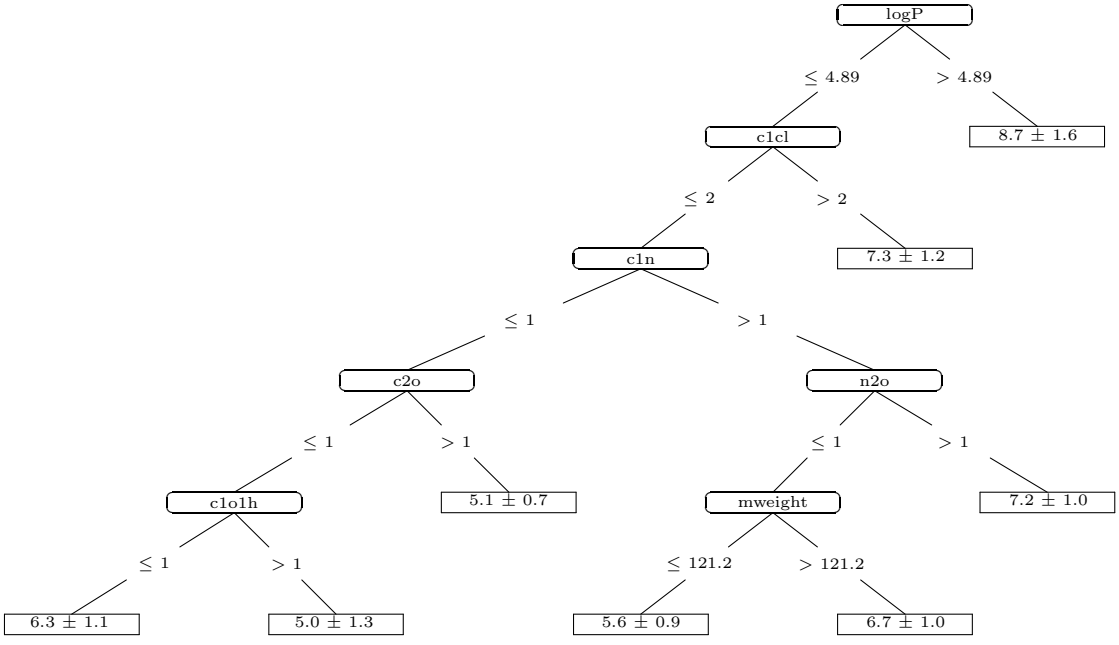
Slika 4.11: Porezano RETISovo drevo z regresijo za množico P1.

lagojeni učnim primerom, poleg tega so drevesa preveč obsežna (tukaj niso prikazana), da bi jih strokovnjak lahko komentiral. Porezana drevesa ($m > 1$) so bistveno boljša. Izbrani deskriptorji so večinoma skladni s poprejšnjim znanjem o biorazgradljivosti. Njihova napovedna moč je sicer slabša od Cubistovih modelov, a boljša od PLS modelov.

V tabelah korelacijskih koeficientov 4.17 in 4.18 je podan tudi korelacijski koeficient modela iz [8], zgrajenega z orodjem za gradnjo regresijskih dreves M5', ki je reimplementacija sistema M5 opisanega v razdelku 3.2.1. Žal ni direktno primerljiv z ostalimi vrednostmi v stolpcu q. Izračunan je kot povprečje petih »10 delnih« križnih validacij. Glede na to, da je Cubist naslednik M5, sta si sistema zelo podobna, še posebej če Cubist uporablja privzete parametre.



Slika 4.12: Porezano RETISovo drevo brez regresije za množico P1



Slika 4.13: Obrezano RETIS-ovo drevo brez regresije za množico P2

5. Zaključki

Na večih realnih problemih smo preizkusili uporabnost regresijskih dreves za napovedovanje biorazgradljivosti spojin iz njihovih strukturnih lastnosti. Uporabili smo metodi gradnje regresijskih dreves, implementirani v programskih orodjih Cubist in RETIS.

Uporabili smo več različnih množic podatkov, na osnovi katerih smo zgradili večje število modelov z različnimi parametri algoritmov za njihovo gradnjo. Iz te množice modelov smo izbrali tiste z največjo napovedno močjo ter jih dali v oceno strokovnjakoma na področju biorazgradljivosti. Naše modele sta primerjala z modeli, zgrajenimi s klasičnimi metodami linearne regresije, oziroma z modeli zgrajenimi z drugimi metodami strojnega učenja. Vsi uporabni modeli so predstavljeni v prejšnjem poglavju.

Ugotovili smo, da je kvaliteta zgrajenega modela močno odvisna od števila učnih primerov, ki smo jih uporabili za njegovo gradnjo. Za velike učne množice dobimo z orodji za gradnjo regresijskih dreves natančnejše modele kot s klasično linearno regresijo. Njihova prednost pride še posebej do izraza pri modeliranju množice različnih vrst spojin, ker nam drevo omogoča za vsako vrsto spojin svoj model. Pri modeliranju manjših množic podatkov so regresijska drevesa primerljive ali slabše natančnosti od linearne regresije. Kljub temu regresijska drevesa zgrajena iz majhnih učnih množic niso neuporabna. Dobili smo namreč nekaj modelov, ki so se ob slabši natančnosti odlikovali po svoji preprostosti in razumljivosti, zaradi česar so bili zelo dobro ocenjeni. Pri tem ne gre pozabiti, da je gradnja regresijskih dreves z omenjenimi orodji, s stališča uporabnika, bistveno hitrejša in preprostejša kot uporaba linearne regresije. Pri slednji gre za interaktiven postopek modeliranja, ki zahteva veliko strokovnega znanja, medtem ko strojno učenje poteka avtomatično, ko imamo že pripravljene podatke.

Primerjava natančnosti modelov, zgrajenih z obema orodjema pokaže, da je Cubist v rahli prednosti. Po drugi strani pa imamo pri RETISu večjo kontrolo nad gradnjo in

predvsem nad naknadnim rezanjem dreves.

Na osnovi ugotovljenega sklepamo, da so bili cilji zadane naloge izpolnjeni. Na koncu naj omenimo še nekaj idej za nadaljnje izboljšave predstavljenih modelov:

1. namesto »surovih« bi lahko uporabili centrirane in skalirane podatke;
2. preizkusili bi lahko razne nelinearne funkcije deskriptorjev (log, exp, $1/x$, itn.), pri čemer bi bila nujna pomoč strokovnjaka;
3. namesto uporabljenih deskriptorjev bi lahko s pomočjo PCA metode izračunali nove, med seboj ortogonalne deskriptorje.

Na osnovi teh in drugih predlogov se nam ponuja še obilo nadaljnjega dela.

Literatura

[1] Damborsky, J., Schultz, T.W., Comparison of the QSAR models for toxicity and biodegradability of anilines and phenols. V *Chemosphere*, št. 34, strani 429–446, 1997.

[2] Blaha, L., Damborsky, J., Nemec, M., QSAR for acute toxicity of saturated and unsaturated halogenated aliphatic compounds. V *Chemosphere*, št. 36, strani 1345–1365, 1998.

[3] Damborsky, J., Lynam, M., Kuty, M., Structure-biodegradability relationships for chlorinated dibenzo-p-dioxins and dibenzofurans. V Wittich, R.-M., *Biodegradation of dioxins and furans*. R.G. Landes Company, Austin, 1998.

[4] Damborsky, J., Manova, K. and Kuty, M., A mechanistic approach to deriving QSBR - A case study: dehalogenation of haloaliphatic compounds. V Peijnenburg, W.J.G.M., Damborsky, J., *Biodegradability Prediction*. Kluwer Academic Publishers, Dordrecht, 1996.

[5] Damborsky, J., Quantitative structure-function relationships of the single-point mutants of haloalkane dehalogenase: A multivariate approach. V *Quantitative Structure-Activity Relationships*, št. 16, strani 126-135, 1997.

[6] Damborsky, J., Quantitative structure-function and structure-stability relationships of purposely modified proteins. V *Protein Engineering*, št. 11, strani 21-30, 1998.

[7] Damborsky, J., Berglund, A., Kuty, M., Ansorgova, A., Nagata, Y., Sjostrom, M., Mechanism-based Quantitative Structure-Biodegradability Relationships for hydrolytic dehalogenation of chloro- and bromo-alkenes. V *Quantitative Structure-Activity Relationships*, št. 17, strani 450-458, 1998.

[8] Džeroski, S., Blockeel, H., Kompare, B., Kramer, S., Pfahringer, B., Van Laer, W., Experiments in predicting biodegradability. V *Proceedings of the 9th International Workshop on Inductive Logic Programming*, strani 80–91. Springer, Berlin, 1999.

[9] Matko, D., *Identifikacije*, Fakulteta za elektrotehniko, Ljubljana, 1998.

[10] Geladi, P., Kowalski, B. R., Partial Least-Squares Regression: A Tutorial. V *Analytica Chimica Acta*, št. 185, strani 1–17, 1986.

[11] Wold, S., Johansson, E., Cocchi, M., PLS Analysis. V Kubinyi, N. (ured.), *3D QSAR in Drug Design*. ESCOM, Leiden, 1993.

[12] Wold, S., Validation of QSAR's. V *Quantitative Structure-Activity Relationships*, št. 10, strani 191–193, 1991.

[13] Kononenko, I., *Strojno učenje*. Fakulteta za računalništvo in informatiko, Ljubljana, 1997.

[14] Kubat, M., Bratko, I., Michalski, R. S., A Rewiew of Machine Learning Methods. V Michalski, R. S., Bratko, I., Kubat, M. (ured.), *Machine Learning and Data Mining: Methods and Applications*. John Wiley&Sons Ltd., New York, 1997.

[15] Quinlan, J, R., Learning with continuous classes. V Adams, Sterling (ured.), *Proceedings AI'92*, strani 343–348. World Scientific, Singapore, 1992.

[16] Quinlan, J. R., Combining instance-based and model-based learning. V *Proceedings of the Tenth International Conference on Machine Learning*, strani 236–243. Morgan Kaufmann Publishers, San Fransisco, 1993.

[17] Karalič, A., *Avtomatsko učenje regresijskih dreves iz nepopolnih podatkov*. Magistrsko delo, Fakulteta za elektrotehniko in računalništvo, Ljubljana, 1991.

Izjava

Izjavljam, da sem diplomsko delo izdelal samostojno pod vodstvom mentorja prof. dr. Nikole Pavešića in somentorja doc. dr. Saša Džeroskega. Izkazano pomoč drugih sodelavcev sem v celoti navedel v zahvali.

Bernard Ženko

Priloge

A Primer podatkov za QSAR modeliranje

Spojina	$\log K_{ow}$	HOMO $[ev]$	LUMO $[ev]$	r_w $[10nm]$	V_w $[\frac{cm^3}{Mol}]$	μ $[Debye]$	M_w	σ	pK_a	$\log IGC_{50}^{-1}$ $[mol.l^{-1}]$	$\log k_b$ $[L.org.^{-1}.h^{-1}]$
anilin	0.90	-8.61	0.42	0.12	56.38	1.5833	93.13	-0.16	4.58	0.109	-12.68
3-metilanilin	1.40	-8.57	0.42	0.17	70.05	1.4153	107.16	-0.23	4.75	-0.28	-13.77
3-metoksianilin	0.95	-8.69	0.26	0.26	73.75	2.3423	123.15	-0.04	4.26	0.085	-14.30
3-kloroanilin	1.88	-8.76	0.12	0.175	63.38	2.6026	127.57	0.21	3.51	0.092	-13.64
3-bromoanilin	2.10	-8.83	-0.01	0.185	71.50	2.7165	172.02	0.23	3.43	0.517	-13.43
3-cianoanilin	1.05	-9.03	-0.51	0.32	61.08	4.4301	118.14	0.40	2.79	-0.465	-15.67
3-nitroanilin	1.37	-9.28	-1.06	0.259	73.18	6.3025	138.13	0.55	2.45	0.026	-14.92
fenol	1.48	-9.17	0.29	0.12	53.88	1.2338	89.07	-0.16	9.92	-0.241	-11.16
4-metilfenol	2.12	-8.95	0.33	0.17	67.55	1.3602	108.14	-0.31	10.1	-0.162	-11.33
4-metoksifenol	1.57	-9.11	0.17	0.26	71.25	2.4059	124.14	-0.32	10.2	-0.143	-12.70
4-klorofenol	2.48	-9.01	0.05	0.175	65.88	1.4767	128.56	0.11	9.38	0.402	-11.77
4-bromofenol	2.63	-9.31	-0.03	0.185	69.00	1.5930	173.01	0.12	9.45	0.500	-11.80
4-cianofenol	1.60	-9.56	-0.54	0.32	69.58	3.3106	119.12	0.84	7.96	0.516	-13.82
4-nitrofenol	1.85	-10.71	-1.08	0.259	70.68	5.2640	139.11	1.08	7.15	1.420	-13.00
4-acetilfenol	1.45	-9.45	-0.4	0.25	78.25	3.8245	136.15	0.34	8.05	-0.093	-12.51

Podatki o toksičnosti in biorazgradljivosti anilinov ter fenolov iz razdelka 4.2.

B Primer vhodnih in izhodnih datotek sistema Cubist

```
logkb.      | logkb - biodegradation

compound:   ignore.
logKow:     continuous.
HOMO:       continuous.
LUMO:       continuous.
rw:         continuous.
Vw:         continuous.
Dip:        continuous.
Mw:         continuous.
sigma:      continuous.
pKa:        continuous.
logIGC50:   ignore.
logkb:      continuous.
```

Vhodna datoteka `QSAR_An_Phe_Bio.names` z opisom deskriptorjev anilinov in fenolov iz razdelka 4.2.

```
ani,  0.90,  -8.61,  0.42, 0.12,  56.38, 1.5833,  93.13, -0.16,  4.58,  0.109, -12.68.
CH3a, 1.40,  -8.57,  0.42, 0.17,  70.05, 1.4153, 107.16, -0.23,  4.75, -0.28,  -13.77.
OCHa, 0.95,  -8.69,  0.26, 0.26,  73.75, 2.3423, 123.15, -0.04,  4.26,  0.085, -14.3.
Cla,  1.88,  -8.76,  0.12, 0.175, 63.38, 2.6026, 127.57,  0.21,  3.51,  0.092, -13.64.
Bra,  2.10,  -8.83, -0.01, 0.185, 71.5,  2.7165, 172.02,  0.23,  3.43,  0.517, -13.43.
CNa,  1.05,  -9.03, -0.51, 0.32,  61.08, 4.4301, 118.14,  0.40,  2.79, -0.465, -15.67.
NO2a, 1.37,  -9.28, -1.06, 0.259, 73.18, 6.3025, 138.13,  0.55,  2.45,  0.026, -14.92.
phe,  1.48,  -9.17,  0.29, 0.12,  53.88, 1.2338,  89.07, -0.16,  9.92, -0.241, -11.16.
CH3p, 2.12,  -8.95,  0.33, 0.17,  67.55, 1.3602, 108.14, -0.31, 10.1,  -0.162, -11.33.
OCHp, 1.57,  -9.11,  0.17, 0.26,  71.25, 2.4059, 124.14, -0.32, 10.2,  -0.143, -12.7.
Clp,  2.48,  -9.01,  0.05, 0.175, 65.88, 1.4767, 128.56,  0.11,  9.38,  0.402, -11.77.
Brp,  2.63,  -9.31, -0.03, 0.185, 69,    1.593,  173.01,  0.12,  9.45,  0.5,   -11.8.
CNp,  1.60,  -9.56, -0.54, 0.32,  69.58, 3.3106, 119.12,  0.84,  7.96,  0.516, -13.82.
NO2p, 1.85, -10.71, -1.08, 0.259, 70.68, 5.264,  139.11,  1.08,  7.15,  1.42,  -13.
ACp,  1.45,  -9.45, -0.40, 0.25,  78.25, 3.8245, 136.15,  0.34,  8.05, -0.093, -12.51.
```

Vhodna datoteka `QSAR_An_Phe_Bio.data` s podatki o anilinih in fenolih iz razdelka 4.2.

```
Cubist [Release 1.07]  Thu Jun 01 15:26:12 2000
---------------------
    Target attribute ‘logkb’

Read 15 cases (12 attributes) from QSAR_An_Phe_Bio.data

Model:

  Rule 1: [15 cases, mean -13.100, range -15.67 to -11.16, est err 0.332]

        logkb = -12.983 + 0.328 pKa - 10.5 rw

Evaluation on training data (15 cases):

    Average  |error|              0.192
    Relative |error|               0.18
    Correlation coefficient        0.98

Time: 0.1 secs
```

Izhodna datoteka `QSAR_An_Phe_Bio.out` z zgrajenim modelom biorazgradljivosti anilinov in fenolov iz razdelka 4.2.

C Primer vhodnih in izhodnih datotek sistema RETIS

```
logkb
5
2
9
logKow
continuous
5
2
HOMO
continuous
5
2
LUMO
continuous
5
2
rw
continuous
5
3
Vw
continuous
5
```

```
2
Dip
continuous
5
4
Mw
continuous
5
2
sigma
continuous
5
2
pKa
continuous
5
3
```

Vhodna datoteka `ds1_apb.rdo` z opisom deskriptorjev anilinov in fenolov iz razdelka 4.2.

```
15
-12.68  0.9    -8.61   0.42  0.12   56.38  1.5833   93.13  -0.16   4.58   0.109
-13.77  1.4    -8.57   0.42  0.17   70.05  1.4153  107.16  -0.23   4.75  -0.28
-14.3   0.95   -8.69   0.26  0.26   73.75  2.3423  123.15  -0.04   4.26   0.085
-13.64  1.88   -8.76   0.12  0.175  63.38  2.6026  127.57   0.21   3.51   0.092
-13.43  2.1    -8.83  -0.01  0.185  71.5   2.7165  172.02   0.23   3.43   0.517
-15.67  1.05   -9.03  -0.51  0.32   61.08  4.4301  118.14   0.4    2.79  -0.465
-14.92  1.37   -9.28  -1.06  0.259  73.18  6.3025  138.13   0.55   2.45   0.026
-11.16  1.48   -9.17   0.29  0.12   53.88  1.2338   89.07  -0.16   9.92  -0.241
-11.33  2.12   -8.95   0.33  0.17   67.55  1.3602  108.14  -0.31  10.1   -0.162
-12.7   1.57   -9.11   0.17  0.26   71.25  2.4059  124.14  -0.32  10.2   -0.143
-11.77  2.48   -9.01   0.05  0.175  65.88  1.4767  128.56   0.11   9.38   0.402
-11.8   2.63   -9.31  -0.03  0.185  69     1.593   173.01   0.12   9.45   0.5
-13.82  1.6    -9.56  -0.54  0.32   69.58  3.3106  119.12   0.84   7.96   0.516
-13     1.85  -10.71  -1.08  0.259  70.68  5.264   139.11   1.08   7.15   1.42
-12.51  1.45   -9.45  -0.4   0.25   78.25  3.8245  136.15   0.34   8.05  -0.093
```

Vhodna datoteka `ds1_apb.rda` s podatki o anilinih in fenolih iz razdelka 4.2.

```
--------------------------------------------------------------------------------
Regression tree generated by RETIS, Version V2.16.6.d/modified

Date   : 01-Jun-2000
Time   : 17:26:15

Domain : DS1_APB
Class  : logkb
Nodes  : 1
Leaves : 2
Created with m = 0.00
Pruned  with m = 1.00

--------------------------------------------------------------------------------
NODE: Weight=15.00  y = -12.60+-0.19010699*x1+0.29401568*x2+
                        2*x3-14*x4+0.00237019*x5+0.50480235*x6+
                        0.00799948*x7+1*x8+0.45475623*x9          Error= 0.14
[ logKow <=  1.47 ]
    LEAF: Weight=6.00  y = -11.37+-0.44346055*x1+0.49067312*x2+
                           3*x3-14*x4+-0.01361180*x5+0.73674935*x6+
                           0.01464552*x7+2*x8+0.53297234*x9       Error= 0.02
[  1.47 < logKow ]
    LEAF: Weight=9.00  y = -14.77+-0.04929547*x1+0.01503376*x2+
                           0*x3-16*x4+0.04028730*x5+0.12072358*x6+
                           -0.00075751*x7+1*x8+0.34700477*x9      Error= 0.01
ENDNODE
--------------------------------------------------------------------------------
```

Izhodna datoteka `ds1_apb.txt` z zgrajenim modelom biorazgradljivosti anilinov in fenolov iz razdelka 4.2.

D Preglednica vseh zgrajenih modelov

Model	PLS		Cubist										RETIS							
			Privzeti par.				Optimizirani par.				Z regresijo									
											$m = 0$				$m = 0.5$					
	R^2	Q^2	r	q	R^2	Q^2	r	q	R^2	Q^2	r	q	R^2	Q^2	r	q	R^2	Q^2		
Toksičnost ter biorazgradljivost anilinov in fenolov (razdelek 4.2)																				
Toksičnost anilinov	–	–	0.00	-0.39	0.00	-0.96	0.00	-0.39	0.00	-0.96	1.00	0.11	1.00	-78.2	1.00	0.57	1.00	-7.78		
Toksičnost fenolov	0.96	0.83	0.83	-0.15	0.69	-0.44	0.83	-0.15	0.69	-0.44	1.00	-0.07	1.00	-1.52	1.00	0.55	1.00	-6.64		
Toksičnost anilinov in fenolov	–	–	0.51	0.05	0.26	-0.24	0.51	0.05	0.26	-0.24	1.00	0.10	1.00	-12.4	0.98	-0.39	0.96	-4.29		
Biorazgradljivost anilinov	0.95	0.89	0.97	-0.40	0.93	-0.77	0.97	-0.40	0.93	-0.77	1.00	-0.23	1.00	-1199	1.00	0.12	1.00	-164		
Biorazgradljivost fenolov	0.99	0.93	0.98	0.48	0.96	0.16	0.98	0.48	0.96	0.16	1.00	0.13	1.00	-109	1.00	0.76	1.00	-11.2		
Biorazgradljivost anilinov in fenolov	0.96	0.95	**0.98**	**0.91**	**0.96**	**0.82**	0.98	0.91	0.96	0.82	1.00	0.37	1.00	-43.0	1.00	0.92	1.00	0.84		
Akutna toksičnost nasičenih in nenasičenih alifatskih ogljikovodikov (razdelek 4.3)																				
Haloakani (vsi desk.)	0.90	0.77	**0.92**	**0.79**	**0.84**	**0.61**	0.92	0.84	0.84	0.71	1.00	0.39	1.00	-0.17	0.99	0.41	0.92	0.07		
Haloakani (desk. MR in EE)	0.90	0.88	**0.93**	**0.83**	**0.86**	**0.65**	0.93	0.88	0.86	0.75	1.00	0.85	1.00	0.60	0.98	0.92	0.96	0.85		
Haloakani in haloalkeni (desk. MR in EE)	0.42	0.30	0.71	0.62	0.51	0.38	0.71	0.62	0.51	0.38	1.00	0.23	1.00	-2.46	0.91	0.36	0.83	-0.28		
Haloakani in haloalkeni brez dveh spojin (desk. MR in EE)	0.89	0.88	**0.94**	**0.92**	**0.88**	**0.84**	0.98	0.91	0.96	0.82	1.00	0.40	1.00	-4.21	0.99	0.94	0.97	0.88		
Haloakani in haloalkeni (desk. MR, EE, BO, Hf in CR)	0.85	0.68	0.71	0.62	0.51	0.38	0.71	0.62	0.51	0.38	1.00	-0.75	1.00	-54.2	0.99	0.76	0.99	0.56		
Biorazgradljivost dioksinov in furanov (razdelek 4.4)																				
Model z vsemi deskriptorji	0.94	0.78	**0.98**	**0.78**	**0.97**	**0.60**	0.98	0.88	0.96	0.73	–	–	–	–	–	–	–	–		
Model s 15. deskriptorji	0.95	0.88	**0.93**	**0.82**	**0.85**	**0.64**	0.92	0.88	0.83	0.75	1.00	-0.10	1.00	-22619	1.00	-0.24	1.00	-17.0		
Model z 9. deskriptorji	0.94	0.92	**0.89**	**0.87**	**0.79**	**0.76**	0.88	0.89	0.77	0.79	1.00	-0.04	1.00	-226	1.00	0.32	1.00	-2.72		
Biorazgradljivost haloalifatskih spojin (razdelek 4.5)																				
Model z vsemi spojinami	0.34	0.20	0.55	0.12	0.30	-0.38	0.65	0.60	0.42	0.33	1.00	0.26	1.00	-0.99	0.96	0.05	0.91	-2.88		
Model brez dveh spojin	0.92	0.87	**0.89**	**0.80**	**0.80**	**0.63**	0.97	0.85	0.94	0.71	1.00	0.20	1.00	-13.8	1.00	0.83	0.99	0.58		
Biorazgradljivost mutantov haloalkanske dehalogenaze (razdelek 4.6)																				
Model z vsemi deskriptorji	0.50	0.35	0.84	0.35	0.71	-0.08	0.84	0.35	0.71	-0.08	–	–	–	–	–	–	–	–		
Model s 14. deskriptorji	0.86	0.60	–	–	–	–	–	–	–	–	0.99	0.29	0.98	-2.73	1.00	0.23	1.00	-28.3		
Model s 4. deskriptorji	0.84	0.75	0.84	0.46	0.71	0.07	0.84	0.48	0.71	0.04	1.00	0.64	1.00	0.12	1.00	0.76	1.00	0.43		
Aktivnost in stabilnost namensko spremenjenih proteinov (razdelek 4.7)																				
Dhla-Phe172	0.83	0.77	0.95	0.60	0.91	0.28	0.93	0.66	0.87	0.38	1.00	0.55	1.00	-1.54	1.00	0.72	0.99	-0.02		
Subt-Met222	0.86	0.81	0.70	0.25	0.46	0.01	0.90	0.33	0.80	0.06	1.00	0.62	1.00	-7.36	1.00	0.46	0.99	-0.20		
Lyso-Thr175	0.87	0.85	0.93	0.70	0.87	0.47	0.93	0.73	0.87	0.52	1.00	-0.14	1.00	-12.0	0.99	0.58	0.98	-0.23		
Synth-Glu49	0.76	0.71	**0.87**	**0.81**	**0.76**	**0.65**	0.89	0.85	0.79	0.71	1.00	0.23	1.00	-178	0.99	0.61	0.97	0.23		
Biorazgradljivost haloalkenov (razdelek 4.8)																				
Edini model	0.92	0.81	0.88	-0.54	0.78	-1.37	0.88	-0.54	0.78	-1.37	1.00	-0.16	1.00	-44.9	1.00	-0.19	1.00	-675		

Korelacijski koeficienti vseh zgrajenih modelov, 1. del. Poudarjene modele je pregledal strokovnjak.

Model	PLS		RETIS															
			Z regresijo				Brez regresije											
			$m = 1$				$m = 0$				$m = 0.5$				$m = 1$			
	R^2	Q^2	r	q	R^2	Q^2	r	q	R^2	Q^2	r	q	R^2	Q^2	r	q	R^2	Q^2
Toksičnost ter biorazgradljivost anilinov in fenolov (razdelek 4.2)																		
Toksičnost anilinov	–	–	1.00	0.24	1.00	-9.02	1.00	0.04	1.00	-0.92	1.00	-0.07	0.89	-0.65	0.99	-0.11	0.75	-0.47
Toksičnost fenolov	0.96	0.83	1.00	0.96	1.00	0.18	1.00	0.11	1.00	-0.19	0.99	0.01	0.91	-0.24	0.98	-0.08	0.81	-0.25
Toksičnost anilinov in fenolov	–	–	0.97	-0.39	0.94	-4.05	1.00	0.15	1.00	-0.28	0.98	0.13	0.91	-0.19	0.94	0.18	0.79	-0.10
Biorazgradljivost anilinov	0.95	0.89	1.00	-0.59	1.00	-135	1.00	0.31	1.00	-0.31	0.97	0.24	0.87	-0.10	0.85	0.55	0.68	0.28
Biorazgradljivost fenolov	0.99	0.93	1.00	-0.43	1.00	-670	1.00	-0.29	1.00	-1.74	0.98	-0.26	0.91	-1.04	0.95	-0.24	0.80	-0.91
Biorazgradljivost anilinov in fenolov	0.96	0.95	**1.00**	**0.92**	**1.00**	**0.84**	1.00	0.58	1.00	0.14	0.98	0.59	0.94	0.28	0.94	0.72	0.84	0.49
Akutna toksičnost nasičenih in nenasičenih alifatskih ogljikovodikov (razdelek 4.3)																		
Haloakani (vsi desk.)	0.90	0.77	0.97	0.36	0.83	0.06	1.00	0.39	1.00	-0.17	0.99	0.41	0.92	0.07	0.97	0.36	0.83	0.06
Haloakani (desk. MR in EE)	0.90	0.88	**0.98**	**0.93**	**0.95**	**0.85**	1.00	0.77	1.00	0.52	0.99	0.73	0.92	0.47	0.95	0.59	0.78	0.31
Haloakani in haloalkeni (desk. MR in EE)	0.42	0.30	0.87	0.35	0.75	-0.23	1.00	0.21	1.00	-0.52	0.99	0.18	0.92	-0.23	0.91	0.22	0.73	-0.10
Haloakani in haloalkeni brez dveh spojin (desk. MR in EE)	0.89	0.88	**0.98**	**0.94**	**0.97**	**0.88**	1.00	0.87	1.00	0.72	0.99	0.84	0.95	0.70	**0.96**	**0.81**	**0.88**	**0.65**
Haloakani in haloalkeni (desk. MR, EE, BO, Hf in CR)	0.85	0.68	**0.99**	**0.78**	**0.97**	**0.59**	1.00	0.50	1.00	0.14	0.98	0.48	0.91	0.19	0.95	0.36	0.79	0.07
Biorazgradljivost dioksinov in furanov (razdelek 4.4)																		
Model z vsemi deskriptorji	0.94	0.78	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
Model s 15. deskriptorji	0.95	0.88	**1.00**	**0.89**	**1.00**	**0.71**	1.00	0.74	1.00	0.46	0.98	0.76	0.94	0.58	**0.96**	**0.75**	**0.88**	**0.55**
Model z 9. deskriptorji	0.94	0.92	1.00	0.55	1.00	-0.01	1.00	0.78	1.00	0.57	0.98	0.79	0.94	0.61	**0.96**	**0.75**	**0.88**	**0.55**
Biorazgradljivost haloalifatskih spojin (razdelek 4.5)																		
Model z vsemi spojinami	0.34	0.20	0.94	0.01	0.88	-2.60	1.00	0.30	1.00	-0.61	0.99	0.21	0.95	-0.41	0.97	0.19	0.89	-0.13
Model brez dveh spojin	0.92	0.87	**0.99**	**0.88**	**0.98**	**0.74**	1.00	0.83	1.00	0.67	0.99	0.79	0.95	0.62	0.96	0.74	0.88	0.54
Biorazgradljivost mutantov haloalkanske dehalogenaze (razdelek 4.6)																		
Model z vsemi deskriptorji	0.50	0.35	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
Model s 14. deskriptorji	0.86	0.60	1.00	0.14	1.00	-48.0	1.00	0.19	1.00	-0.73	0.99	0.17	0.95	-0.41	0.99	0.16	0.89	-0.30
Model s 4. deskriptorji	0.84	0.75	1.00	0.74	0.99	0.49	1.00	0.35	1.00	-0.12	0.99	0.36	0.95	0.04	0.99	0.32	0.89	0.01
Aktivnost in stabilnost namensko spremenjenih proteinov (razdelek 4.7)																		
Dhla-Phe172	0.83	0.77	0.99	0.76	0.98	0.19	1.00	0.33	1.00	-0.12	0.99	0.31	0.94	0.01	0.99	0.34	0.89	0.08
Subt-Met222	0.86	0.81	0.99	0.45	0.98	-0.13	1.00	0.15	1.00	-0.68	0.99	0.15	0.89	-0.27	0.96	0.14	0.76	-0.28
Lyso-Thr175	0.87	0.85	0.99	0.58	0.98	-0.21	1.00	0.51	1.00	0.01	0.98	0.52	0.91	0.22	0.94	0.38	0.81	0.05
Synth-Glu49	0.76	0.71	0.99	0.67	0.97	0.33	1.00	0.84	1.00	0.63	0.99	0.80	0.93	0.64	**0.97**	**0.80**	**0.87**	**0.63**
Biorazgradljivost haloalkenov (razdelek 4.8)																		
Edini model	0.92	0.81	1.00	0.10	1.00	-177	1.00	-0.52	1.00	-2.09	0.99	-0.55	0.94	-1.61	0.93	-0.55	0.80	-1.39

Korelacijski koeficienti vseh zgrajenih modelov, 2. del. Poudarjene modele je pregledal strokovnjak.

Množica (razdelek 4.9)	P1				P1			
Orodje	r	q	R^2	Q^2	r	q	R^2	Q^2
Cubist (privzeti parametri)	0.755	0.665	0.569	0.439	0.767	0.630	0.587	0.381
RETIS (z regresijo, m=1)	0.838	0.592	0.702	0.292	–	–	–	–
RETIS (z regresijo, m=5)	0.776	0.579	0.601	0.296	–	–	–	–
RETIS (brez regresije, m=1)	0.948	0.562	0.859	0.276	0.945	0.604	0.862	0.339
RETIS (brez regresije, m(P1)=8,m(P2)=11)	0.650	0.575	0.411	0.331	0.694	0.606	0.456	0.360
PLS (pred izbiro deskriptorjev)	0.562	–	0.316	0.285	0.585	–	0.343	0.296
PLS (po izbiri deskriptorjev)	0.522	–	0.272	0.257	0.601	–	0.361	0.349

Korelacijski koeficienti vseh zgrajenih modelov, 3. del.