UNIVERZA V LJUBLJANI
FAKULTETA ZA RAČUNALNIŠTVO IN INFORMATIKO

Črtomir Gorup

RAČUNSKE TEHNIKE NAPOVEDOVANJA VPLIVA UČINKOVIN NA FENOTIP MODELNIH ORGANIZMOV

Diplomska naloga
na univerzitetnem študiju

Mentor: prof. dr. Blaž Zupan

Ljubljana, 2009

Univerza
v Ljubljani

Fakulteta za računalništvo
in informatiko

Tržaška 25
1000 Ljubljana, Slovenija
telefon: 01 476 84 11
faks: 01 426 46 47
www.fri.uni-lj.si
e-mail: dekanat@fri.uni-lj.si

Št. naloge: 01573/2009
Datum: 01.09.2009

Univerza v Ljubljani, Fakulteta za računalništvo in informatiko izdaja naslednjo nalogo:

Kandidat: **ČRTOMIR GORUP**

Naslov: **RAČUNSKE TEHNIKE NAPOVEDOVANJA VPLIVA UČINKOVIN NA FENOTIP MODELNIH ORGANIZMOV**

COMPUTATIONAL TECHNIQUES FOR PREDICTION OF EFFECTS OF SMALL MOLECULES ON MODEL ORGANISMS

Vrsta naloge: Diplomsko delo univerzitetnega študija

Tematika naloge:

V nalogi razvijte računske tehnike, ki bodo namenjene napovedovanju kvantitativnega fenotipa modelna organizma, ko je ta izpostavljen delovanju različnih učinkovin. Pri tem uporabite tehnike strojnega učenja in bioinformatike. Preučite različne tipe atributnega opisa učinkovin (fragmenti, QSAR značilke, strukturne in farmakološke anotacije). Delovanje tehnik preskusite na podatkih iz poskusov na socialni amebi D. discoideum.

Mentor:

prof. dr. Blaž Zupan

Fakulteta za računalništvo in informatiko
2
UNIVERZA V LJUBLJANI

Dekan:

prof. dr. Franc Solina

V diplomskem delu je uporabljena črkovna vrsta Latin Modern velikosti 12pt. Besedilo je oblikovano z urejevalnikom besedil ConTEXt.

Namesto te strani **vstavite** original izdane teme diplomskega dela s podpisom mentorja in dekana ter žigom fakultete, ki ga diplomant dvigne v študentskem referatu, preden odda izdelek v vezavo!

Zahvala

Zahvaljujem se mentorju prof. dr. Blažu Zupanu za razvoj ideje tega dela, strokovno pomoč pri njegovi izdelavi ter vse koristne nasvete v času mojega študija.

Prav tako se zahvaljujem dr. Adamu Kuspa, ki nam je omogočil uporabo svojih laboratorijskih rezultatov ter dr. Gadu Shaulsky, ki me je pobližje seznanil z modelnim organizmom *D. discoideum* in mi omogočil udeležbo na šoli SMART 2008 ter poletno delo na Baylor College of Medicine v letu 2008.

Za pomoč pri uporabi programskega paketa Orange in veliko koristnih idej se zahvaljujem tudi dr. Tomažu Curku, Lanu Umeku, Mihi Štajdoharju, Alešu Erjavcu in vsem ostalim članom Laboratorija za umetno inteligenco.

Iskrena hvala pa gre tudi moji družini in Anji, ki so me med študijem podpirali in mi stali ob strani.

Kazalo

Seznam uporabljenih kratic in simbolov

CID CID (angl. Compound ID) identifikator je število, ki enolično določa kemijsko spojino v podatkovni bazi PubChem.

MeSH Ontologija MeSH (angl. Medical Subject Headings) je množica več kot 160000 povezanih pojmov s področja medicine in naravoslovja. Ureja jo National Library of Medicine, ZDA.

QSAR QSAR (angl. Quantitative structure-activity relationship) je relacija, ki povezuje kemijsko (pod)strukturo in biološko aktivnost.

REACH REACH (angl. Registration, Evaluation, Authorisation and restriction of CHemicals) je najnovejša uredba EU iz leta 2006, ki ureja uvoz, izvoz, prodajo in rokovanje s kemikalijami.

SMILES SMILES (angl. Simplified Molecular Input Line Entry Specification) je specifikacija za enoličen zapis kemijske strukture molekule v tekstovni obliki.

Povzetek

Preučevanje odziva modelnih organizmov na učinkovine, ki vključujejo obstoječa zdravila in potencialno zanimive kemikalije nam lahko pomaga razumeti aktivnosti učinkovin, odkriti relevantne biološke procese in nasploh razumeti celične mehanizme. V pričujočem delu smo se osredotočili na razvoj računskih tehnik napovedovanja vpliva učinkovin na fenotip modelnih organizmov. Posebnost našega pristopa je, da poleg značilk QSAR ter vektorja prisotnosti določenega kemijskega fragmenta uporabljamo tudi ekspertno določene označbe učinkovin iz ontologije MeSH. Razvili smo tehniko računanja obogatenih označb MeSH, ki omogoča iskanje podskupin učinkovin s statistično značilnim številom učinkovin s ciljnim fenotipom. Podali smo tudi metode za ocenjevanje podobnosti med učinkovinami. Slednje smo tudi uporabili v t.i. pristopu s semensko učinkovino, ki s stališča napovedovanja vpliva omogoča vrednotenje različnih tipov atributnih opisov učinkovin. Samega napovedovanje vpliva učinkovin smo se lotili z metodo podpornih vektorjev. Z razvitimi metodami in tehnikami smo nato analizirali podatke vplivu 1.045 različnih učinkovin na razvoj socialne amebe *D. discoideum*. Izkazalo se je, da napovedovanje vpliva učinkovin v splošnem ni možno. Kljub temu smo s pomočjo metode podpornih vektorjev in ontologije MeSH uspeli najti podskupine učinkovin, za katere je predikcija možna. Rezultati kažejo na večjo primernost opisa učinkovin z označbami MeSH kot z značilkami QSAR ali vektorjem prisotnosti posameznih kemijskih fragmentov. Odkrili smo tudi 27 obogatenih ($p < 0.02$) označb MeSH, ki določajo enako število podmnožic s statistično pomembnim številom učinkovin s ciljnim fenotipom na modelnem organizmu. Rezultati na modelnem organizmu *D. discoideum* potrjujejo uporabnost označb MeSH za napovedovanje vpliva učinkovin na fenotip.

Ključne besede:

bioinformatika
kemoinformatika
strojno učenje
odkrivanje znanj in podatkov
analiza kemijskih struktur
fenotipi
ontologija MeSH

Abstract

Studying the response of the model organisms exposed to chemicals can help us understand chemical activities, underlying biological processes and cell mechanisms. In the following disertation we have designed a set of computational techniques for predicting the effect of chemical compounds on model organisms. For chemical descriptors we have examined three different annotation systems, including QSAR-based descriptors, molecular fingerprints (presence of specific short fragments) and MeSH terms from the MeSH ontology. Use of MeSH terms is also the the distinctive feature of our approach. We have developed a technique for computing MeSH term enrichment which enabled us to identify enriched subsets of chemicals with statistical significant ratio of chemicals with the target effect on phenotype. In order to identify the most suitable chemical description we have also developed a method for evaluating different types of attribute-based chemical descriptions. We used the support vector machine for predicting the effect of chemical compounds. Using the developed methods we analyzed the data from the experiment where model organism *D. discoideum* was exposed to 1.045 different chemical compounds and relative growth inhibition was observed as a phenotype. In general we are not able to predict the effects. However, if we split the chemicals to groups sharing some MeSH annotation term, we were able to find the terms for which our predictive procedures worked well. Results from the chemical description evaluation show that attributes based on MeSH terms are more suitable than QSAR-based descriptors and molecular fingerprints. We have also indentified 27 enriched ($p < 0.02$) MeSH terms which determine the same number of subsets with statisticaly significant ratio of chemical compounds causing the observing phenotype on model organism. Results confirm that use of MeSH terms improves prediction of the chemical impact on model organism.

Keywords:

bioinformatics
chemoinformatics
machine learning
data mining
chemical structure analysis
phenotypes
MeSH ontology

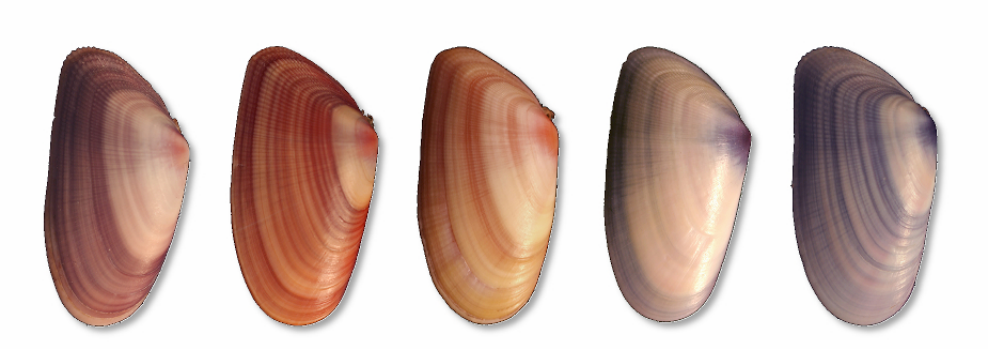
Poglavje 1

Uvod

> Nič v življenju ni takšnega, česar bi se morali bati, je le takšno, kar bi morali razumeti. Čas je, da razumemo več, da nas bo manj strah.
>
> *Marie Curie*

Fenotip je vsaka značilka ali karakteristika (npr. barva kril, oblika telesa, velikost ploda, obnašanje), ki omogoča razločevanje med osebki iste vrste. Danes je znano, da je fenotip odvisen od genotipa ali genetske zasnove ter okoljskih dejavnikov, ki vplivajo na razvoj organizma. Iskanje vzročne povezave med okoljskimi dejavniki, genotipom in fenotipom je osrednja raziskovalna tema v molekularni biologiji. Nove tehnologije (npr. za kvantitativno karakterizacijo kemijskih učinkovin in za hitro sekveniranje genoma) omogočajo analize, ki so bile še pred nekaj leti časovno neizvedljive. Velike količine podatkov so postale neprimerne za ročno analizo, zato se vedno večji del raziskav v molekularni biologiji seli iz biološkega laboratorija v nadaljno računalniško obdelavo.

V tem diplomskem delu se bomo osredotočili na računsko iskanje povezave med okoljskimi dejavniki ter fenotipom. Zaradi kompleksnosti problema in tehnoloških omejitev pri izvedbi laboratorijskega dela raziskav smo omejeni na en okoljski dejavnik ter eno fenotipno značilnost. V našem primeru je okoljski dejavnik izpostavljenost organizma določeni kemijski spojini. Opazovani fenotip pa je inhibicija rasti organizma v primerjavi z normalno pričakovano rastjo. Diplomska naloga se osredotoča na razvoj računskih tehnik, uporabljeni pa so eksperimentalni podatki, ki so jih o



Slika 1.1 Fenotip različnih osebkov školjke vrste *Donax variabilis*.

rasti socialne amebe *D. discoideum* zbrani na Baylor College of Medicine v Houston, ZDA.

1.1 Napovedovanje fenotipov na podlagi kemijske strukture

Napovedovanje lastnosti na podlagi kemijske strukture ima bogato zgodovino. Prvi delujoči modeli QSAR (angl. Quantitative structure-activity relationship) [1] so bili uporabljeni za napovedovanje fizičnih lastnosti kemikalije (npr. kislost, temperatura vrelišča). Naprednejši kemijski opisi so omogočili napovedovanje bolj kompleksnih pojavov, tudi na živih organizmih. Dandanes so aktualni modeli [2, 3, 4], ki napovedujejo fenotip organizma na podlagi kemijske strukture učinkovine, kateri je bil opazovani organizem izpostavljen. Večina takih modelov [3, 5] se osredotoča na napovedovanje toksičnosti učinkovine, pri tem pa uporablja različne tipe strukturnih značilk in različne tehnike strojnega učenja. Slednje vključujejo tehnike induktivnega logičnega programiranja [4], nevronskih mrež [6] in metode podpornih vektorjev [7, 8].

Modeli za napovedovanje toksičnosti učinkovin na podlagi kemijske strukture imajo veliko uporabno vrednost predvsem za farmacevtsko industrijo. Moderna biotehnologija danes omogoča paralelno izvajanje laboratorijskih eksperimentov z veliko kandidati za končno zdravilo. Na eksperimentalnih podatkih zgrajeni modeli omogočajo odkrivanje stranskih učinkov (npr. toksičnosti, kancerogenosti) že v zgodnji fazi zasnove strukture, ponavadi celo pred fizično sintezo učinkovine. Tako lahko podjetja v eksperimentalni del raziskav vključijo le tiste kandidate za zdravilo, za

katere se ne pričakuje toksičnega odziva. Tak pristop omogoča krajši čas razvoja zdravila in posledično prihranek pri raziskavah [5].

Modeli za napovedovanje toksičnosti so uporabni tudi za vladne agencije (npr. Evropska agencija za kemikalije), ki imajo opravka z novimi kemikalijami z nepoznanimi varnostnimi deklaracijami [5]. S pomočjo modelov jih lahko uredijo glede na pričakovano toksičnost in tiste z vrha seznama z laboratorijskimi eksperimenti prednostno testirajo. Tako lahko kljub omejenim laboratorijskim kapacitetam zagotavljajo posodobljeno zakonodajo (npr. uredba REACH [9]) glede ravnanja z nevarnimi snovmi.

Izdelava modelov QSAR poteka v treh korakih:

- Cilj laboratorijskega dela je analiza vnaprej pripravljene množice učinkovin. Analiza je odvisna od lastnosti, ki jo želimo napovedovati. V primeru napovedovanja fizikalne lastnosti je analiza zgolj meritev fizikalno kemijskih lastnosti učinkovine. Za izdelavo kompleksnejših modelov, in sicer teh, ki napovedujejo toksičnost, so ponavadi potrebni eksperimenti na modelnih organizmih ali na tkivnih vzorcih.

- Drugi korak izdelave modela QSAR vključuje pridobivanje opisov in značilk učinkovin iz testne množice. Za napovedovanje željene lastnosti uporablja večina modelov logistično regresijo [5]. V zadnjem času pa se vedno bolj uporabljajo metode strojnega učenja [3, 5].

- V tretjem koraku izdelave modela QSAR je potrebno preveriti njegovo kvaliteto. To se lahko izvede računsko z uporabo učne množice podatkov ali z laboratorijsko analizo dodatne množice novih učinkovin.

Glede na kemijske značilke in uporabljene računske metode lahko modele razdelimo v tri skupine [5]:

- **Klasični modeli QSAR** so osnovani zgolj na kemijski strukturi učinkovine. Značilke učinkovin so izračunane na podlagi kemijske formule ter prisotnosti ali odsotnosti kemijskih fragmentov. Za napovedovanje novih vrednosti opazovanega parametra taki modeli ponavadi uporabljajo logistično regresijo [5].

- **3D modeli QSAR** uporabljajo značilke, izračunane na podlagi 3D strukture molekule. Te so ponavadi prostorska pozicija glavnih funkcionalnih skupin molekule, prostorska pozicija aktivnih mest in elektrostatski naboj. Taki modeli uporabljajo za napovedovanje metode, ki lahko hitro obdelajo večje število numeričnih značilk. Primer take metode je tehnika podpornih vektorjev [10].

- **Modeli, zasnovani s pomočjo strojnega učenja**, imajo to prednost, da lahko hkrati uporabljajo kemijske opise različnih zvrsti in tipov. Kemijski opis je tako lahko sestavljen iz besedila, numeričnih značilk, nominalnih značilk ali celo kombinacije različnih tipov značilk. Prav tako so lahko podatki, uporabljeni v kemijskem opisu, pridobljeni iz različnih virov (npr. internetna podatkovna zbirka, ontologija, programska oprema za analizo kemijskih struktur). Jedro tipičnega modela iz te skupine predstavlja ena izmed klasifikacijskih metod (npr. odločitveno drevo, naivni bayes, metoda podpornih vektorjev).

V diplomskem delu se bomo osredotočili na modele, ki jih gradimo s tehnikami strojnega učenja. Za karakterizacijo učinkovin bomo uporabili vire, ki vključujejo 2D QSAR značilke in označbe iz ontologije MeSH.

1.2 Razvita orodja in metode

V področju sistemske biologije se genske ontologije uspešno uporabljajo že vrsto let [11, 12]. Po tem vzoru smo se odločili, da raziščemo možnost uporabe ontologije MeSH pri napovedovanju fenotipov. V okviru diplomskega dela smo razvili orodje in več metod, primernih za vrednotenje kemijskih opisov in napovedovanje fenotipov. Metodam je skupno, da za napovedovanje poleg kemijske strukture uporabljajo tudi ontologijo MeSH. Slednja se je z eksperimenti pokazala za izjemno koristno.

V okviru diplomske naloge smo razvili:

1. **Programsko knjižnico "obiMeSH" ter grafični vmesnik "MeSH Ontology Browser"**

   Knjižnica in grafični vmesnik sta sestavna dela "Bioinformatics" modula programskega paketa Orange [13]. Grafični vmesnik omogoča hierarhično pregledovanje podatkov anotiranih z označbami MeSH. Najpomembnejša funkcija knjižnice obiMeSH je izračun obogatenih (statistično značilnih) označb MeSH za poljubni množici vhodnih podatkov.

2. **Metode za vrednotenje uporabnosti različnih tipov kemijskih opisov**

   Razvili smo tehniko, ki služi za vrednotenje tipov kemijskih opisov glede na njihovo uporabnost pri napovedovanju fenotipov modelnega organizma, ki je bil izpostavljen učinkovini. Metoda je opisana na strani 25, primer uporabe na realnem problemu je prikazan na strani 34 diplomskega dela.

3. **Metodo za klasifikacijo učinkovin s pomočjo ontologije MeSH**

   Razvili smo metodo, ki s pomočjo označb MeSH ter standardnih klasifikacijskih tehnik napoveduje fenotip. Njena teoretična podlaga je predstavljena na strani 26, rezultati njene uporabe pa na strani 36 diplomskega dela.

1.3 Struktura diplomskega dela

Poleg uvoda na začetku in zaključka na koncu, diplomsko delo sestavljajo še tri osrednja poglavja. Sledi poglavje, ki opisuje obstoječe označbe in opise značilk kemijskih spojin. V tretjem poglavju je podana teoretična osnova oziroma so opisane računske tehnike, iz katerih smo razvili metode za vrednotenje kemijskih opisov in tehnike za napovedovanje fenotipov na podlagi kemijske strukture. Razvite tehnike so uporabljene na podatkih, ki vključujejo rezultate laboratorijskih poizkusov na modelnem organizmu *Dictyostelium discoideum*. Zadnje poglavje povzema najbolj zanimive rezultate ter podaja predloge za nadaljnje delo.

Poglavje 2

Atributni opis kemijskih spojin

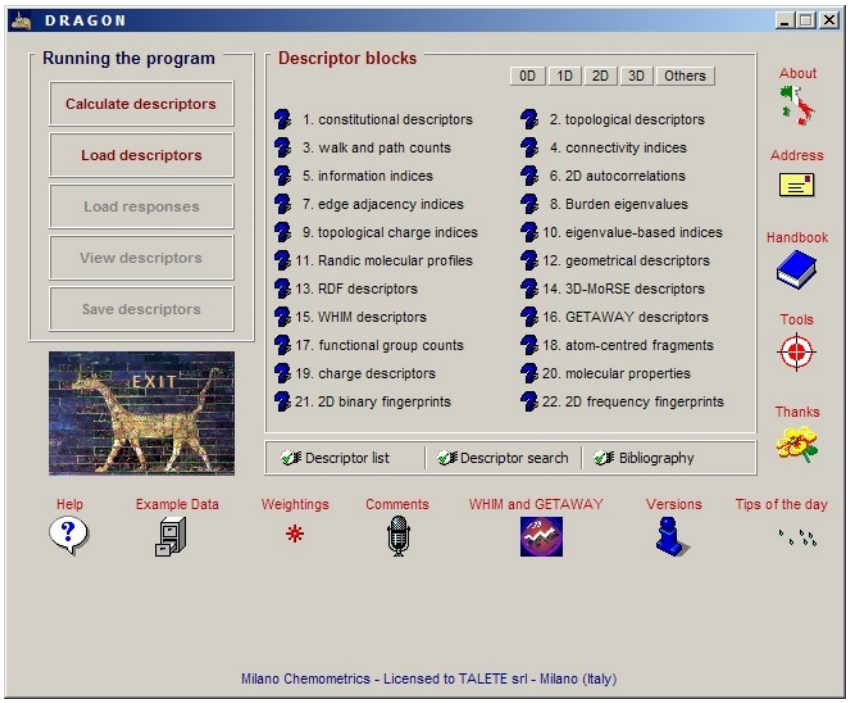
> Znanost je vse, kar razumemo dovolj, da pojasnimo računalniku. Umetnost je vse ostalo.
>
> *Donald Knuth*

Poznamo več načinov, kako kemijsko spojino oziroma učinkovino opišemo kvantitativno, z uporabo značilk (atributov). Značilke oziroma njihove vrednosti lahko najdemo v specializiranih podatkovnih bazah na internetu, druge pa lahko izračunamo s pomočjo specializiranih programskih orodij. V tem poglavju bomo predstavili tri načine opisa kemijskih spojin, ki smo jih, kot je opisano v poglavjih, ki sledijo, uporabili pri razvoju napovedovalnih modelov.

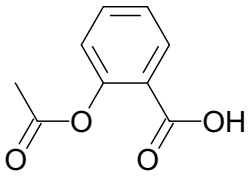
2.1 QSAR značilke

Za izračun QSAR značilk smo uporabili komercialni programski paket Dragon [14]. Na podlagi SMILES zapisa molekul program izračuna 1022 numeričnih in binarnih značilk razdeljenih v 22 skupin. Tipi značilk, ki jih program lahko izračuna, so razvidni iz slike 2.1.

Poglejmo si primer izračunanih značilk za acetil salicilno kislino ali Aspirin. Na podlagi strukturne formule oziroma njenega zapisa v SMILES obliki (slika 2.2) program Dragon izračuna deskriptorje, predstavljene v tabeli 2.1.



Slika 2.1 Vmesnik programa Dragon namenjen izračunu kemijskih značilk



strukturna formula

O=C(C)Oc1ccccc1C(=O)O

strukturna formula v SMILES zapisu

Slika 2.2 Strukturna formula in SMILES zapis učinkovine Aspirin.

2.2 Ontologija MeSH

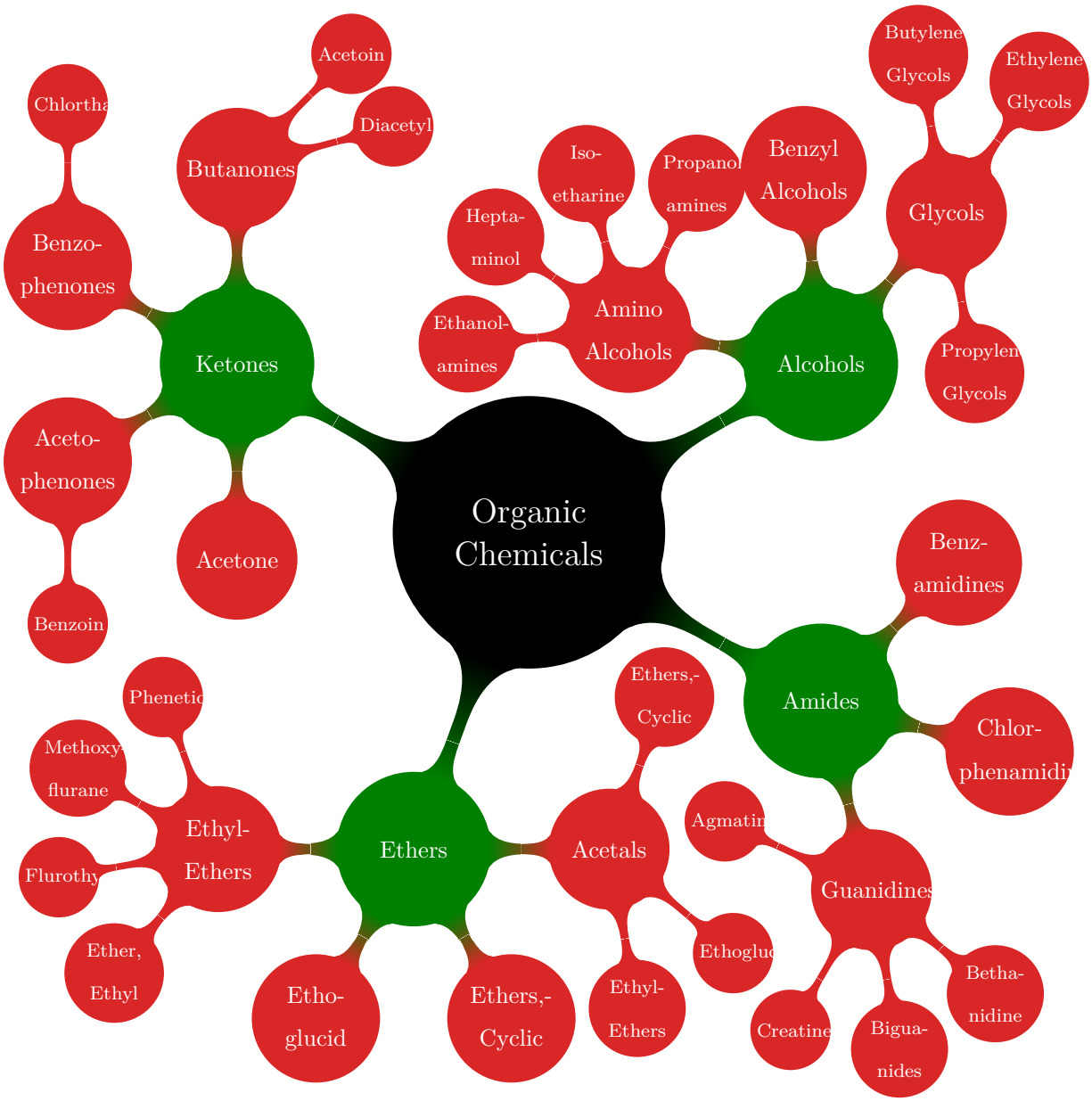
Ontologija MeSH (angl. Medical Subject Headings) [15] je hierarhična zbirka povezanih pojmov s področja medicine in biologije, ki jo ureja ameriški National Library of Medicine [15]. V verziji iz leta 2009 je zajetih 25.186 pojmov. Urejeni so v drevesni strukturi, v kateri so bolj splošni pojmi višje, bolj specializirani pa nižje v drevesu. Pojmi so organizirani v 15 glavnih vej. Ontologija MeSH je uporabljena na

Deskriptor	Vrednost	Opis
MW	180.17	molska masa [g/mol]
AMW	8.58	povprečna molska masa
Sv	13.44	vsota atomskih van der Valsovih prostorov
Se	21.84	vsota atomskih Sandersonovih elektronegativnosti
Sp	13.86	vsota atomske polarizacije
Ss	40.17	vsota Kier-Hall elektrotopoloških stanj
Mv	0.64	povprečje atomskih van der Valsovih prostorov
Me	1.04	povprečje atomskih Sandersonovih elektronegativnosti
Mp	0.66	povprečje atomske polarizacije
Ms	3.09	povprečje Kier-Hall elektrotopoloških stanj
nAT	21	število atomov
nSK	13	število nevodikov
nBT	21	število kemijskih vezi
nBO	13	število nevodikovih kemijskih vezi
nBM	8	število večkratnih kemijskih vezi

Tabela 2.1 Tabela prvih 15 izračunanih QSAR značilk za kemijsko spojino Aspirin.

veliko področjih. V svojih internetnih podatkovnih zbirkah jih uporablja predvsem ameriški National Center for Biotechnology Information. Tako lahko v podatkovni zbirki medicinskih člankov in objav PubMed [16] poleg povzetkov najdemo tudi relevantne označbe MeSH. Slednje omogočajo uporabnikom iskanje sorodnih člankov ter bolj strukturiran pregled nad iskalnim področjem. Na sliki 2.3 je prikazan vrhnji del ontologije MeSH, ki pokriva klasifikacijo dela organskih spojin.

V povezavi z ontologijo MeSH je za tu opisano delo zanimiva predvsem internetna podatkovna zbirka kemijskih spojin PubChem [17]. V njej lahko za poljubno kemijsko spojino najdemo informacije:



Slika 2.3 Vrhnji del ontologije MeSH, ki zajema klasifikacijo dela organskih spojin.

- opis,
- osnovne kemijske lastnosti s strukturo,
- omembe v literaturi,
- seznam podobnih spojin,

- kemijsko klasifikacijo,
- farmakološko klasifikacijo.

Za nas sta bili najbolj zanimivi kemijska in farmakološka klasifikacija spojine. Obe sta predstavljeni z množico pojmov iz ontologije MeSH. S pomočjo označb MeSH je možno iskanje sorodnih učinkovin in iskanje objav s podobno tematiko. Slika 2.4 podaja primer kemične klasifikacije za acetil salicilno kislino, ki uporablja pojme iz ontologije MeSH. Klasifikacija je pridobljena iz podatkovne baze PubChem.

```
Organic Chemicals
  Carboxylic Acids
    Acids, Carbocyclic
      Benzoic Acids
        Hydroxybenzoic Acids
          Salicylic Acids
            Aspirin
      Hydroxy Acids
        Hydroxybenzoic Acids
          Salicylic Acids
            Aspirin
  Phenols
    Hydroxybenzoic Acids
      Salicylic Acids
        Aspirin
```

Slika 2.4 Hierarhija označb MeSH, ki opisujejo kemijsko klasifikacijo Aspirina.

Iz strukture označb MeSH aspirina je razvidno, da se v ontologiji nekateri pojmi (npr. Hydroxybenzoic Acids) pojavljajo v več različnih vejah.

2.3 Atributni opis učinkovin na podlagi prisotnosti kemijskih fragmentov

V pričujočem delu bomo uporabili prosto dostopni programski paket za kemoinformatiko Open Babel [18]. Večinoma se uporablja za pretvorbo med različnimi formati zapisa molekul, operacije nad kemijskimi formulami in za iskanje podstruktur oz. fragmentov v molekulskih formulah. Program omogoča izračun kemijskih profilov (angl. chemical fingerprint) poljubne množice molekul, podane v SMILES zapisu. Prvi korak izračuna je fragmentacija množice strukturnih formul na fragmente do dolžine 7 atomov. Nato program poišče 1024 najpogosteje zastopanih fragmentov. Slednji bodo predstavljali stolpce v binarnem kemijskem profilu enake dolžine. Prisotnost oziroma odsotnost posameznega fragmenta nastavlja vrednost v kemijskem profilu posamezne učinkovine.

Poglavje 3

Računske metode in tehnike

> On, ki ljubi prakso brez teorije, je kot mornar, ki pluje brez kompasa in krmila, in nikoli ne ve, kje bo pristal.
>
> *Leonardo da Vinci*

V nalogi smo uporabili pester nabor tehnik iz statistike, strojnega učenja in kemoinformatike. Povezali smo jih med seboj v učinkovit sistem za analizo vpliva učinkovin in njihovih kemijskih struktur in lastnosti na fenotip opazovanega modelnega organizma.

3.1 Metode ocenjevanja podobnosti med učinkovinami

Vsak opis kemijske spojine je možno pretvoriti v vektor. Vektorji označb MeSH in kemijski profili prisotnosti fragmentov so binarni, medtem ko vektor z značilkami QSAR, pridobljenimi s programom Dragon, vsebuje številčne in diskretne vrednosti. Za računanje podobnosti med učinkovinami smo uporabili dve različni funkciji, ena je primerna za binarne vektorje značilk, druga za številčne.

Za binarne vektorje smo uporabili razdaljo Tanimoto (formula 3.1), katere vrednosti se lahko nahajajo na območju $[0, 1]$:

$$\text{Tanimoto(A,B)} = \frac{c_{AB}}{c_A + c_B - c_{AB}} \tag{3.1}$$

Spremenljivki c_A in c_B predstavljata število enic v vektorju A oziroma B, spremenljivka c_{AB} pa predstavlja število enic v vektorju (A in B). Posebnost te funkcije

je, da pari istoležnih ničel v vektorjih A in B ne pripomorejo k višji podobnosti. To je zaželena lastnost pri računanju podobnosti med učinkovinami, saj na osnovi odsotnosti določenega fragmenta v obeh vektorjih še ne moremo sklepati na večjo medsebojno podobnost.

Za računanje podobnosti med učinkovinami, opisanimi z QSAR značilkami, smo uporabili negirano funkcijo evklidske razdalje (formula 3.2):

$$\text{Euclid(A,B)} = -\sqrt{\sum_{i=0}^{size(A)} (a_i - b_i)^2} \tag{3.2}$$

Spremenljivka a_i predstavlja vrednost na mestu i v vektorju A, spremenljivka b_i pa vrednost na mestu i v vektorju B.

3.2 Tehnike strojnega učenja

Strojno učenje je veja računalništva, ki se ukvarja z avtomatiziranim odkrivanjem znanja v podatkih. Strojno učenje sestavlja več metod, z njimi lahko rešujemo štiri vrste problemov [19]:

- klasifikacija

  Klasifikacija (angl. classification) zajema razvrščanje podatkov v znane skupine oziroma razred. Primer klasifikacije je razvrščanje elektronske pošte zaželeno in nezaželjeno (angl. spam) skupino.

- razvrščanje (angl. clustering)

  Razvrščanje (angl. clustering) se ukvarja z razvrščanjem podatkov v skupine, pri čemer število skupin ni naprej določeno. Cilj postopka je oblikovanje skupin, ki vsebujejo podatke z visoko medsebojno podobnostjo. Razvrščanja se lahko uporablja med predvolilno kampanijo za analizo volilcev ter njihovo razvrstitev v zaključene skupine.

- regresija Regresija (angl. regression) se ukvarja z napovedovanjem številčnih vrednosti. Primer uporabe je model za napovedovanje točke vrelišča kemijskih spojin.

- analiza asociacij in logičnih pravil

  Učenje asociacij (angl. association analysis) zajema iskanje zakonitosti in povezav znotraj podatkov. Rezultati analize asociacij so ponavadi pravila. Učenje asociacij se lahko uporablja za določevanje genov s podobnim izraznim profilom.

Cilj diplomskega dela je napovedovanje vpliva učinkovin na modelni organizem. Oblika problema ustreza klasifikaciji. Posamezne učinkovine oziroma primere bomo razporejali v dve skupini oziroma razreda. skupino tistih, ki nimajo vpliva in skupino tistih, ki imajo vpliv na modelni organizem. Izgradnja klasifikacijskega modela poteka v dveh korakih:

- učenje klasifikatorja

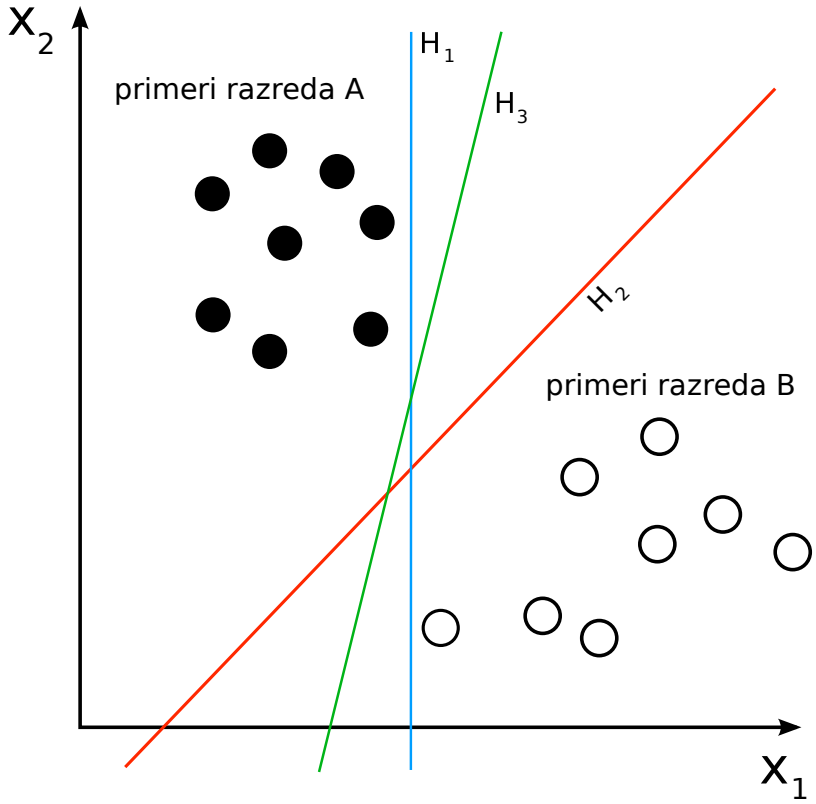
  Prvi korak izdelave klasifikacijskega modela predstavlja učenje klasifikatorja. V tem delu je nujno potrebna množica primerov, na primer kemikalij oziroma njihovih kemijskih opisov, za katere je razred znan. Izbrana klasifikacijska metoda iz množice primerov z znanim razredom izlušči ključne podatke, ki povezujejo primere z razredi. Ključni podatki sestavljajo model, s katerim je mogoče poljuben nov primer uvrstiti v razred.

- testiranje klasifikatorja

  Izračunanemu klasifikacijskemu modelu je potrebno preveriti njegovo kvaliteto napovedovanja. To najlažje storimo, če klasificiramo manjšo množico primerov, za katere poznamo razred. V izogib nerealni oceni klasifikatorja testna in učna množica primerov nikoli nista enaki. Praksa je, da se del vhodne množice primerov uporabi le za testiranje in ne tudi za učenje.

Za napovedovanje vpliva učinkovin na fenotip smo se odločili uporabiti eno izmed novejših klasifikacijskih metod. Imenuje se metoda podpornih vektorjev (angl. support vector machine) [10] ali na kratko metoda SVM. Ključni element metode je

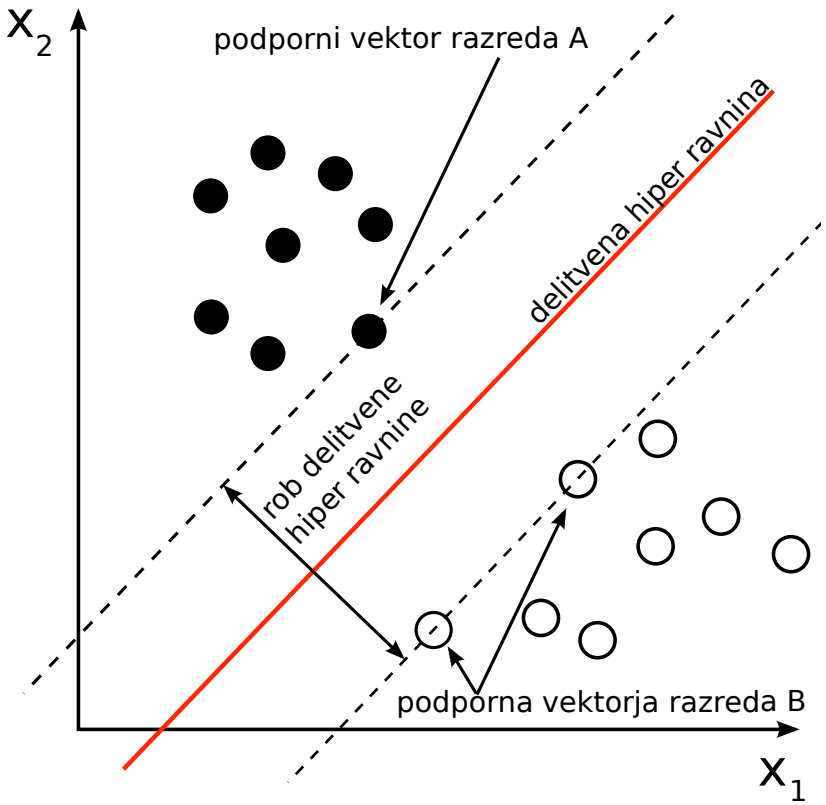
določitev enačbe delitvene hiper ravnine v n dimenzionalnem prostoru, ki ločuje primere iz enega razreda od primerov drugega razreda, pri čemer je n število značilk, ki opisujejo posamezni primer. Slika 3.1 prikazuje 16 primerov, od tega jih polovica pripada razredu A, preostala polovica pa razredu B. Vsak primer je opisan samo z dvema atributoma (X_1 in X_2). Posledica tega je, da se n dimenzionalni poenostavi v dvodimenzinalni prostor oziroma ploskev, delitvena hiper ravnina pa se poenostavi v premico. Iz slike 3.1 vidimo, množico primerov lahko deli neskončno mnogo premic. Princin SRM (angl. structural risk minimization) [20] zagotavlja, da ima najboljša delitvena hiper ravnina tudi najširši možni rob. Rob delitvene hiper ravnine je območje med dvema vzporednima hiper ravninama. Zanju velja, da sekata mejne primere, hkrati pa sta enako oddaljeni od delitvene hiper ravnine.



Slika 3.1 Primeri, razporejeni v atributnem prostoru.

Slika 3.2 prikazuje rob ter optimalno hiper ravnino za primer iz slike 3.1. Primeri, ki ležijo na robnih hiper ravninah, se imenujejo podporni vektorji. Hiper ravnino se lahko v prostoru zapiše z enačbo $\vec{w} \cdot \vec{x} + b = 0$ pri čemer sta $\vec{w}$ in b parametra

modela, $\vec{x}$ pa točka v n dimenzionalnem prostoru. Iskanje optimalne hiper ravnine zajema iskanje vrednosti vektorja $\vec{w}$ in skalarja b pod pogojem, da se noben primer na nahaja znotraj roba odločitvene hiper ravnine. To je optimizacijski problem, ki se ga rešuje z Lagrangeovo metodo [10]. Klasifikacija novih primerov je enostavna. Z vstavljanjem vektorja $\vec{x}$ v izraz $\vec{w} \cdot \vec{x} + b$ lahko glede na predznak rezultata določimo stran delitvene hiper ravnine, na kateri leži primer $\vec{x}$.



Slika 3.2 Primeri, razporejeni v prostoru, skupaj z delitveno hiper ravnino in njenim robom.

Za klasifikacijo učinkovin glede na vpliv na modelne organizme smo uporabili programsko knjižnico LIBLINEAR [21], ki je vgrajena v programski paket Orange [13].

3.3 Tehnika računanja obogatenih označb MeSH

Pri računanju obogatenih označb MeSH smo se zgledovali po metodi iz sistemske biologije, ki omogoča izračun obogatenih označb iz genske ontologije [11, 12]. Skupna

točka obeh metod je diskretna hipergeometrijska porazdelitev, opisana z enačbo 3.3. Slednja opisuje porazdelitev verjetnosti pojavitve k pozitivnih primerov v množici n izbranih primerov iz končne populacije velikosti N, kjer je pozitivnih primerov m. Za razliko od binomske hipergeometrijska porazdelitev upošteva izbiranje primerov brez ponovnega vračanja:

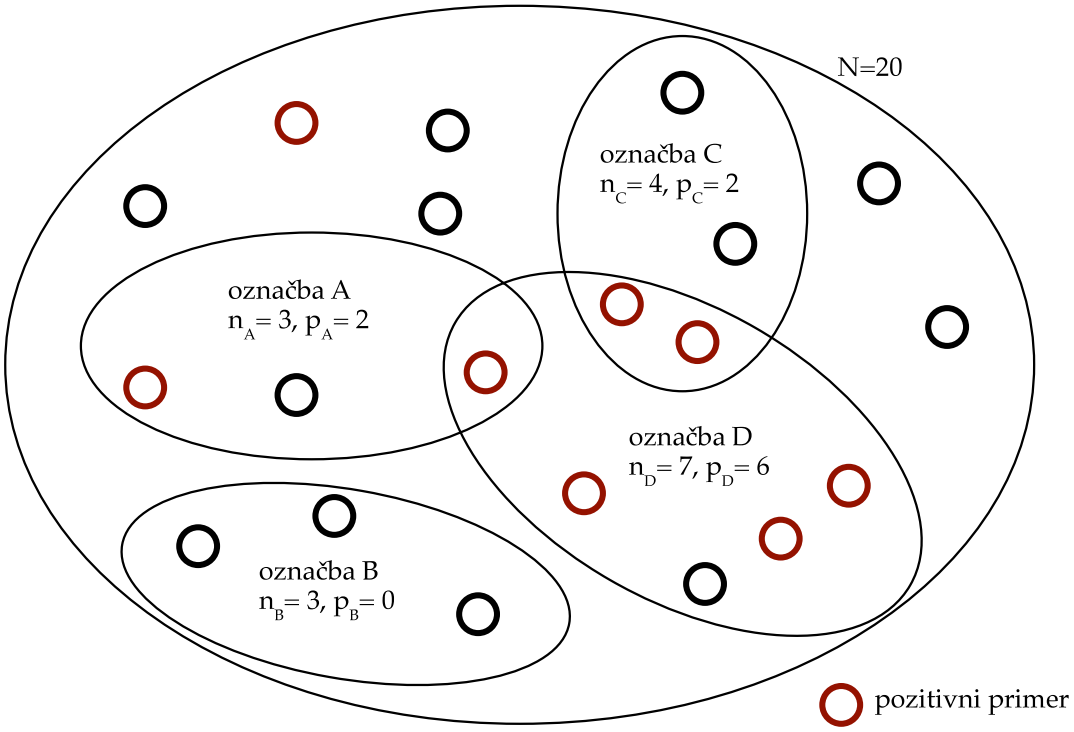
$$P(X = k) = \frac{\binom{m}{k}\binom{N-m}{n-k}}{\binom{N}{n}} \tag{3.3}$$

Primeri so pri eksperimentih, ki smo jih obravnavali v tej diplomi, učinkovine, pozitivni primeri pa le tiste učinkovine, ki povzročijo pri modelnem organizmu opazovani fenotip. Množico kemikalij se lahko glede na njihove označbe MeSH v podatkovni bazi PubChem razvrsti v podmnožice. S pomočjo enačbe 3.3 lahko za vsako označbo MeSH (oziroma podskupino učinkovin) izračunamo obogatitev. Vrednost $P(X = k)$ predstavlja verjetnost pojavitve k kemikalij iz pozitivne skupine v vzorcu velikosti n, pri čemer je vseh kemikalij n, velikost pozitivne skupine pa je m:

$$Enrichment(A) = 1 - \sum_{k=0}^{p_A - 1} P(X = k) \tag{3.4}$$

Obogatitev označbe MeSH A se izračuna po formuli 3.4, pri čemer je p_A število pozitivnih primerov iz podmnožice, določene z označbo A. Obogatitev skupine kemikalij (ali p-vrednost) si lahko razlagamo kot verjetnost, da bi enako velika, naključno izbrana podmnožica, vsebovala enako ali večje število učinkovin iz pozitivnih primerov. Cilj postopka je odkrivanje označb MeSH z nizkimi p-vrednostmi, torej takih označb, ki v svojih množicah vsebujejo statistično značilno število pozitivnih primerov.

Oglejmo si primer na sliki 3.3. Celotna množica vsebuje 20 primerov oz. v našem primeru učinkovin ($N = 20$). Slednje so označene s krogci, pri čemer rdeči označujejo primere iz pozitivne skupine ($m = 8$), črni pa vse ostale ($N - m = 12$). Primer sestavljajo tudi štiri označbe (A, B, C in D), ki določajo štiri podmnožice različnih velikosti ($n_A = 3, n_B = 3, n_C = 4$ in $n_D = 7$). Obogatitev označbe A se izračuna z naslednjo enačbo:



Slika 3.3 Grafična predstavitev primera računanja obogatenih označb MeSH.

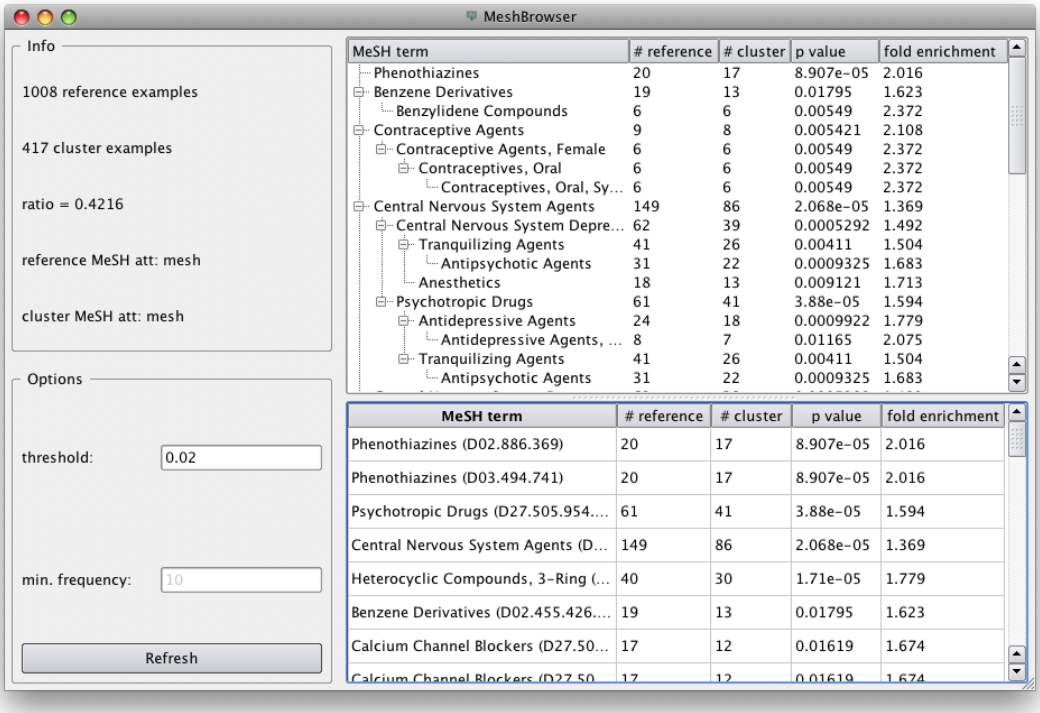
$$
\begin{aligned}
Enrichment(A) &= 1 - \sum_{k=0}^{1} P(X = k) = \\
&= 1 - (P(X = 0) + P(X = 1)) = \\
&= 1 - \frac{\binom{8}{0}\binom{12}{3} + \binom{8}{1}\binom{12}{2}}{\binom{20}{3}} = \\
&= 1 - \frac{\binom{12}{3} + 8 \cdot \binom{12}{2}}{\binom{20}{3}} = \\
&= 1 - \frac{220 + 528}{1140} = \\
&= 1 - 0.656 = \\
&= 0.344
\end{aligned}
$$

Tabela 3.1 podaja p-vrednosti za preostale označbe iz slike 3.3. V našem primeru je statistično značilna ($p < 0.05$) le označba D. Njena množica vsebuje 7 primerov, od tega je kar 6 pozitivnih.

označba MeSH	n	p	p-vrednost
D	7	6	0.0044
A	3	2	0.3440
C	4	2	0.5300
B	3	0	1.0000

Tabela 3.1 Tabela obogatenih označb MeSH s pripadajočimi p-vrednostmi.

Računanje obogatenih označb MeSH smo realizirali v obliki knjižnice "obiMeSH" za programski paket Orange. Slika 3.4 prikazuje razvito komponento "MeSH Term Browser" ki služi kot grafični vmesnik do programske knjižnice.



Slika 3.4 Grafični vmesnik "MeSH Term Browser".

3.4 Tehnike vrednotenja atributnih opisov in modelov

V tem delu smo uporabili različne vrste atributnega opisa učinkovin. Zanimalo nas je, kateri od opisov je bolj informativen s stališča napovedovanja odziva modelnega organizma na dano učinkovino.

3.4.1 Pristop s semensko učinkovino

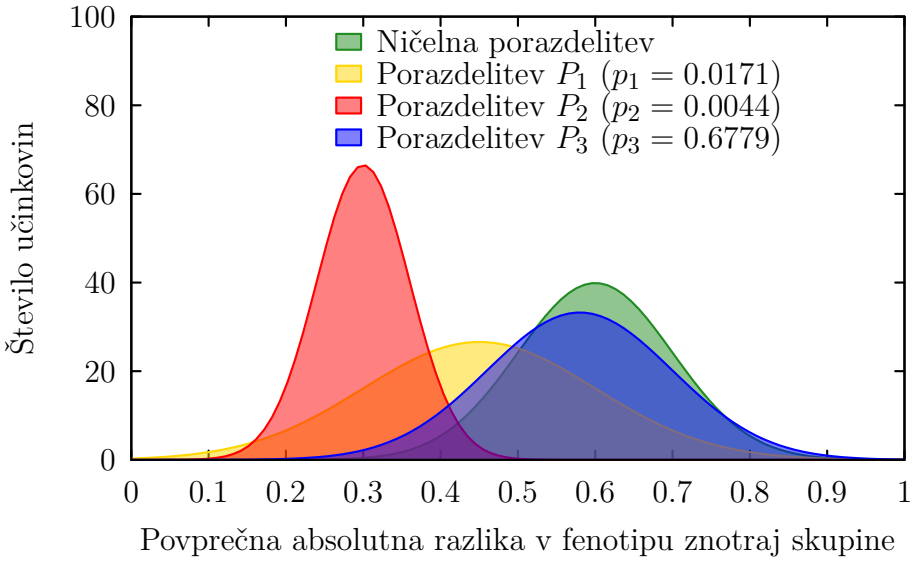
Pričakovati je, da učinkovine s podobnim kemijskim opisom tudi podobno delujejo na modelni organizem, oziroma pri tem lahko opazimo podoben fenotip. Da bi to res lahko ocenili ter hkrati razvrstili kemijske opise glede na njihovo moč napovedovanja, smo razvili pristop s semensko učinkovino.

Postopek se začne z naključnim izborom semenske učinkovine C_s ter množice desetih najbolj podobnih učinkovin $C_i \in C_1, C_2, \ldots, C_{10}$. Nato smo izračunali povprečno absolutno razliko med fenotipom učinkovine C_s in fenotipom učinkovine C_i. Povprečno absolutno razliko smo označili z Δ_S. Izračuna se jo po formuli 3.5, pri čemer funkcija $gi(C_i)$ vrača vrednost fenotipa pri učinkovini C_i:

$$\Delta_S = \frac{\sum_{i=1}^{10} abs(gi(C_s) - gi(C_i))}{10} \tag{3.5}$$

Postopek smo ponovili za 1000 naključno izbranih semenskih učinkovin za vsak kemijski opis ter, odvisno od uporabe enega od treh načinov atributnega zapisa lastnosti učinkovin, dobili tri porazdelitve spremenljivke Δ_S. Porazdelitve smo primerjali z ničelno porazdelitvijo, kjer smo sosedske učinkovine C_i izbrali naključno. S Kolmogorov-Smirnov statističnim testom smo testirali hipotezo o enakosti parov porazdelitev. Tako pridobljene p-vrednosti lahko uporabimo tudi za razvrstitev kemijskih opisov glede na moč napovedovanja.

Pričakovati je, da se izračunane porazdelitve pomaknejo proti nižjim vrednostim od ničelne porazdelitve. Boljši kot je kemijski opis, manjše absolutne razlike v fenotipu lahko pričakujemo med podobnimi učinkovinami. Slika 3.5 prikazuje hipotetično ničelno porazdelitev, skupaj s tremi porazdelitvami P_1, P_2 in P_3. Iz slike lahko razberemo največje odstopanje med ničelno in P_2 porazdelitvijo. Kolmogorov-Smirnov test s 50 uporabljenimi vzorci iz vsake porazdelitve to potrjuje, saj velja $p_2 < p_1 < p_3$.



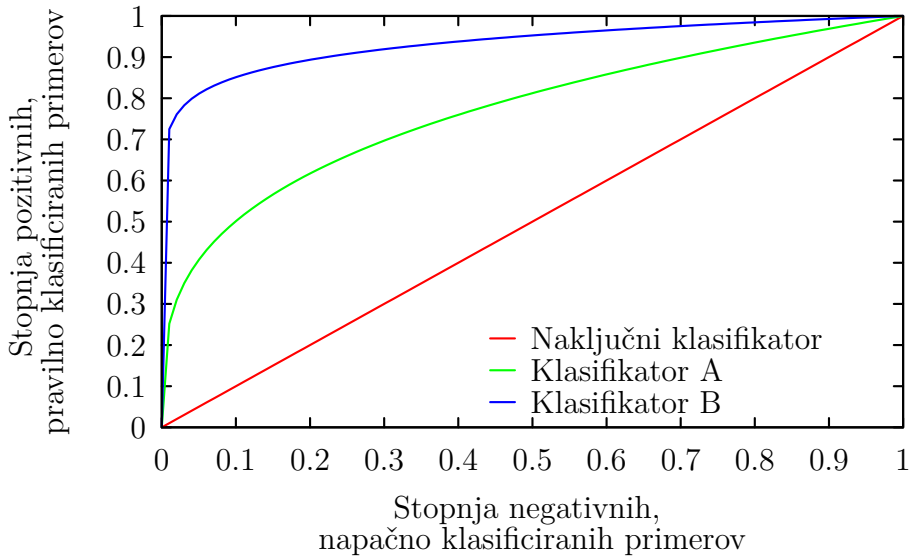
Slika 3.5 Primer ničelne porazdelitve skupaj s tremi testnimi porazdelitvami s pripadajočimi p-vrednostmi.

Razvili smo tudi različico metode, ki poleg vrednotenja kemijskih opisov na celotni zbirki učinkovin omogoča tudi vrednotenje opisov učinkovin, ki so lastne skupini učinkovin z določeno označbo MeSH. Postopek se razlikuje pri izboru semenske učinkovine, ki smo jo tokrat izbrali iz podmnožice kemikalij, označenih z določeno označbo MeSH. Tako smo za množico označb MeSH dobili specifične p-vrednosti. Slednje lahko uporabljamo kot mero za kvaliteto napovedi za posamezne podmnožice učinkovin.

3.4.2 Vrednotenje napovednih točnosti postopka s strojnim učenjem

Za testiranje napovednih kvalitet modela, ki za dani izbor kemijskih značilk napovejo fenotip, smo uporabili 5-kratno prečno preverjanje [19]. Množica podatkov je bila razdeljena na pet podmnožic s približno enako razporeditvijo razredov. Vedno je bila uporabljenih $\frac{4}{5}$ podatkov za učenje, preostanek pa za testiranje. Učenje in testiranje modela se je izvedlo 5 krat, vsakič je bila za testiranje uporabljena druga petina podatkov.

Kot mero za ocenjevanje uspešnosti klasifikatorja smo uporabili oceno AUC (angl. Area Under Curve) [22]. Osnova AUC ocene je krivulja ROC (angl. Receiver Operating Charateristic) [23]. Krivulja ROC se uporablja za vizualizacijo kvalitete oz. lastnosti klasifikatorjev. Vsaka krivulja pripada enemu klasifikatorju, vsaka točka na krivulji pa predstavlja ocene modela pri določenem verjetnostnem pragu klasifikacije v ciljni razred. Vodoravno os prostora ROC predstavlja stopnja negativnih, napačno klasificiranih primerov (angl. FP rate). Na navpični osi pa je nanešena stopnja pozitivnih, pravilno klasificiranih primerov. Primer krivulje ROC je predstavljen na sliki 3.6.



Slika 3.6 Primer krivulje ROC.

Naključni klasifikator ima linearno krivuljo ROC, ki deli prostor ROC po diagonali. Boljši kot je klasifikator (npr. Klasifikator B na sliki 3.6), bolj se njegova krivulja približuje levemu zgornjemu kotu. Krivulja ROC idealnega klasifikatorja (pravilno klasificira vse pozitivne in negativne primere) gre skozi točko (0,1). Oceno AUC dobimo tako, da izračunamo površino pod krivuljo ROC. Zavzema lahko območje od 0.5 pri naključnem klasifikatorju, pa do 1 pri idealnem klasifikatorju. Oceno AUC lahko interpretiramo kot verjetnost, da bo klasifikator naključni pozitivni primer ocenil z višjo verjetnostjo pripadnosti razredu kot naključni negativni primer. Možno je, da se klasifikator z višjo oceno AUC v nekaterih delih območja ROC obnese slabše od klasifikatorja z nižjo oceno AUC. V splošnem pa je ocena AUC dober pokazatelj napovednih sposobnosti klasifikatorja.

Poglavje 4

Analiza podatkov o amebi *D.discoideum*

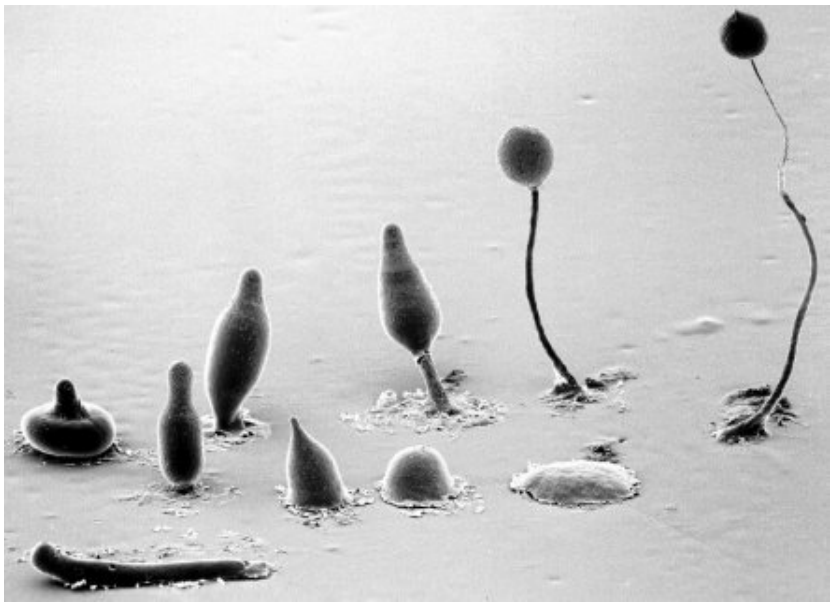
> Kljub lepi strategiji občasno poglej rezultate.
>
> *Winston Churchill*

V eksperimentalnem delu pričujoče diplomske naloge smo analizirali eksperimentalne podatke, ki so bili pridobljeni v okviru študije vpliva različnih učinkovin na rast modelnega organizma *D. discoideum*.

4.1 Opis problema

Organizem *Dictyostelium discoideum* je ameba, ki živi v zgornjih plasteh zemlje [24]. Posebnost organizma je, da ga lahko najdemo v enocelični kot v večcelični obliki. V normalnih pogojih se *Dictyostelium discoideum* nahaja v enocelični obliki, ko pa nastopi pomanjkanje hrane, se začnejo celice združevati v tvorbo, sestavljeno iz večih celic. Tvorba je sestavljena iz debla in spore, le celice v spori preživijo pomanjkanje hrane. Ko je le te v izobilju, spora razpade in življenski cikel (slika 4.1) se nadaljuje.

Zaradi svojih lastnosti (enostavno in poceni gojenje, kratek razvojni cikel, sekveniran genom [25]) je *Dictyostelium discoideum* zelo priljubljen modelni organizem, primeren za veliko vrst raziskav. Raziskava tega organizma, pri kateri smo sodelovali v fazi analize podatkov, se odvija na Baylor College of Medicine (ZDA) pod vodstvom dr. Adam Kuspa. Cilj raziskave je ugotoviti vpliv različnih učinkovin na fenotip organizma, ki je v našem primeru rast. Laboratorijski del raziskav je zajemal testiranje 1.040 različnih učinkovin na gojiščih celične kulture *Dictyostelium discoideum*. Vsako gojišče je bilo 48 ur izpostavljeno eni izmed učinkovin ($1\mu g$). Celice



Slika 4.1 *Dictyostelium discoideum* v vseh stopnjah svojega življenskega cikla.

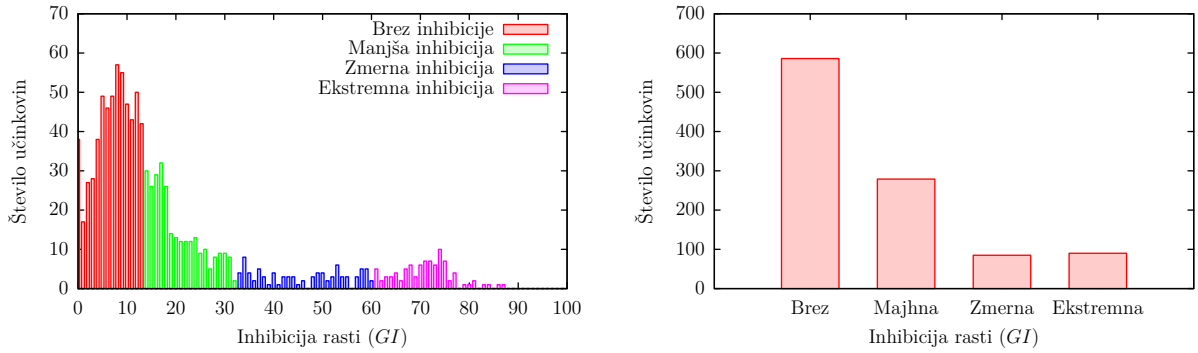
so bile po 48 urah preštete in primerjane s številom celic neizpostavljenega vzorca. V izogib napakam so bili za vsako učinkovino izvedeni trije neodvisni eksperimenti. Analiza rezultatov iz biološkega laboratorija se je nadaljevala z računalniško obdelavo podatkov in iskanjem zakonitosti.

4.2 Predstavitev in obdelava vhodnih podatkov

Vhodna množica vsebuje 1.045 testiranih učinkovin. Za vsako so bile podane tri meritve relativne inhibicije rasti GI (v primerjavi z neizpostavljeno celično kulturo *D. discoideum*) v razponu od 0% do 100%. Za vsako učinkovino smo izračunali povprečno vrednost GI. Glede na velikost GI smo na podlagi ekspertnega mnenja učinkovine razporedili v štiri razrede:

- brez inhibicije, $0 < GI \leq 13$, 586 učinkovin
- manjša inhibicija, $13 < GI \leq 32$, 279 učinkovin
- zmerna inhibicija, $32 < GI \leq 60$, 85 učinkovin
- ekstremna inhibicija, $60 < GI \leq 100$, 90 učinkovin.

Sliki 4.2 prikazuje numerično in diskretno porazdelitev vrednosti GI.



Slika 4.2 Realna in diskretna porazdelitev vrednosti GI.

V naslednjem koraku je sledilo pridobivanje podatkov o posamezni učinkovini iz podatkovne baze PubChem. Preko imena učinkovine smo najprej poiskali pripadajočo identifikacijsko številko CID (angl. Compound ID). Od 1045 danih učinkovin smo jih v podatkovni bazi našli 1008 (97%). Za učinkovine s CID številko smo v podatkovni bazi poiskali označbe MeSH (farmakološko in kemijsko klasifikacijo) in kemijsko formulo v zapisu SMILES. Označbe MeSH smo našli za 887 učinkovin. Podatke smo združili v tabelo s stolpci, opisanimi v tabeli 4.1.

Za vsako izmed učinkovin smo s pomočjo formule v zapisu SMILES in programskega paketa Open Babel izračunali kemijski odtis, s programskim paketom Dragon pa smo pridobili značilke QSAR. Za vsako učinkovino smo izračunali binarni vektor označb MeSH, kjer vrednost v posameznem polju pomeni prisotnost oz. odsotnost določene označb MeSH. Uporabili smo samo označbe MeSH, ki so uporabljene pri več kot 10 in manj kot 600 učinkovinah. Takih označb je bilo 170. Na enak način smo izdelali vektor prisotnosti posameznih fragmentov, le da vrednost polja v tem vektorju nakazuje prisotnost oziroma odsotnost določene kemijske podstrukture.

Za vhodno množico kemijskih spojin in vse tri kemijske opise smo izračunali matrike podobnosti, ki smo jih uporabili v naslednjih analizah.

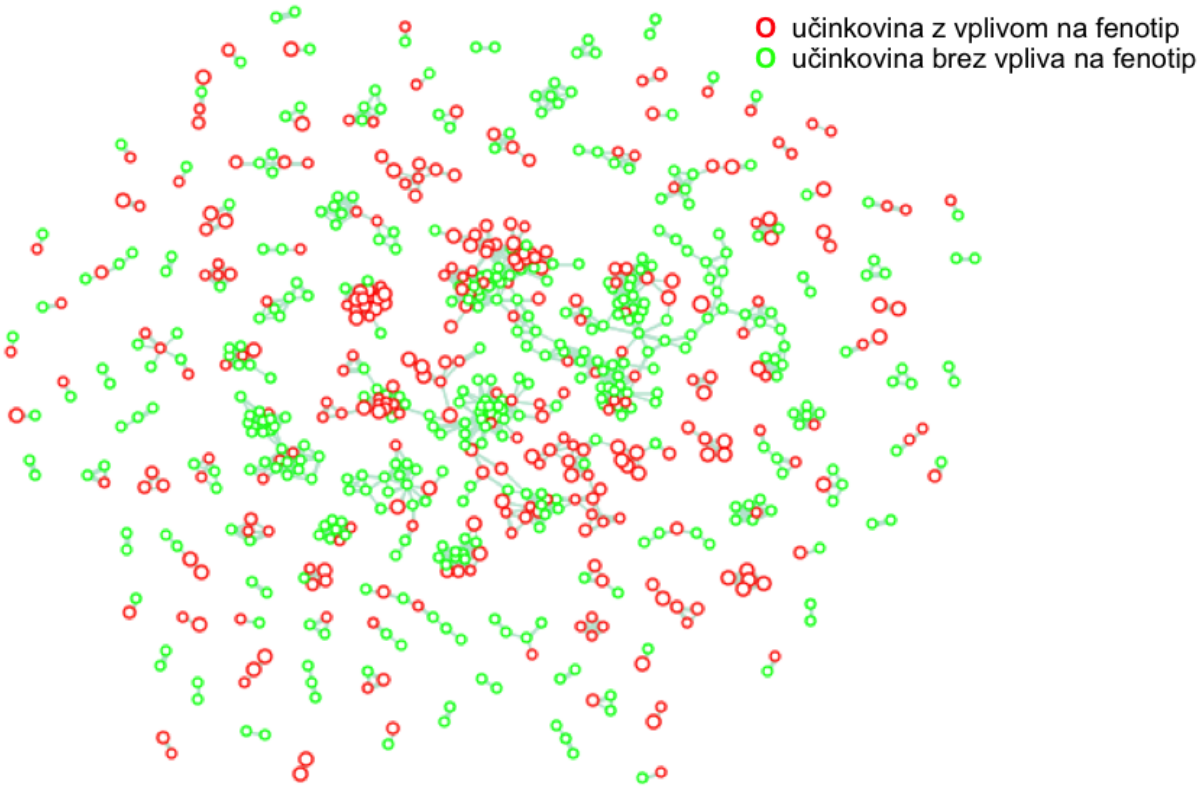
Matrike podobnosti lahko tudi izkoristimo za vizualizacijo učinkovin s pomočjo mrež (sliki 4.3). V tem primeru vozlišča mreže predstavljajo posamezne učinkovine, povezave med vozlišči pa določeno stopnjo podobnosti. Osnova za izdelavo mreže je

Ime stolpca	**Opis**
Plate ID	Identifikator laboratorijskega poizkusa. Nahaja se v območju od 01A01 do 14G10 (npr. 01A02).
CID	Številka CID iz podatkovne zbirke PubChem (npr. 8646).
Name	Ime učinkovine (npr. Tetrahydroxy).
Formula	Kemijska formula (npr. $C_4H_4N_6O$).
Weight	Molska masa [g/mol] (npr. 152.12).
MeSH annotation	Označbe MeSH (kemijske in farmakološke, npr. "Organic Chemicals, Hypoxanthines, Guanine, Aza Compounds, Azaguanine, Heterocyclic Compounds").
Smiles	Kemijska struktura v SMILES zapisu (npr. "C12=NNNC1=NC(=NC2=O)N").
Growth	Relativna inhibicija rasti (npr. "8.82%").
Class	Fenotipni razred; enak "no", če vrednost GI pripada razredu brez inhibicije, drugače enak "yes".

Tabela 4.1 Stolpci v tabeli učinkovin.

matrika podobnosti. Slednja je bila izračunana na podlagi vektorjev prisotnosti posameznih fragmentov in s funkcijo razdalje Tanimoto. Povezane so le tiste učinkovine, ki imajo medsebojno razdaljo Tanimoto večjo ali enako 0.6. Z zeleno barvo so označene učinkovine brez učinka na rast modelnega organizma ($GI \leq 13$). Z rdečo barvo so označene učinkovine, ki imajo učinek na rast organizma $GI > 13$. Velikost vozlišča je premo sorazmerna z vrednostjo GI. Nepovezana vozlišča so izpuščena.Pri razporejanju vozlišč je bila uporabljena metoda Fruchterman Reingold [26], za vizualizacijo pa je bil uporabljen programski paket Orange [13].

Pojavitev povezanih podskupin učinkovin z enakim fenotipnim razredom nakazuje, da so podobne učinkovine povzročitelj podobnega fenotipa. Opazimo lahko tudi, da se te podskupine pojavljajo zelo lokalno, kar daje slutiti, da bo učinkovite globalne modele napovedovanja fenotipov težko pridobiti.

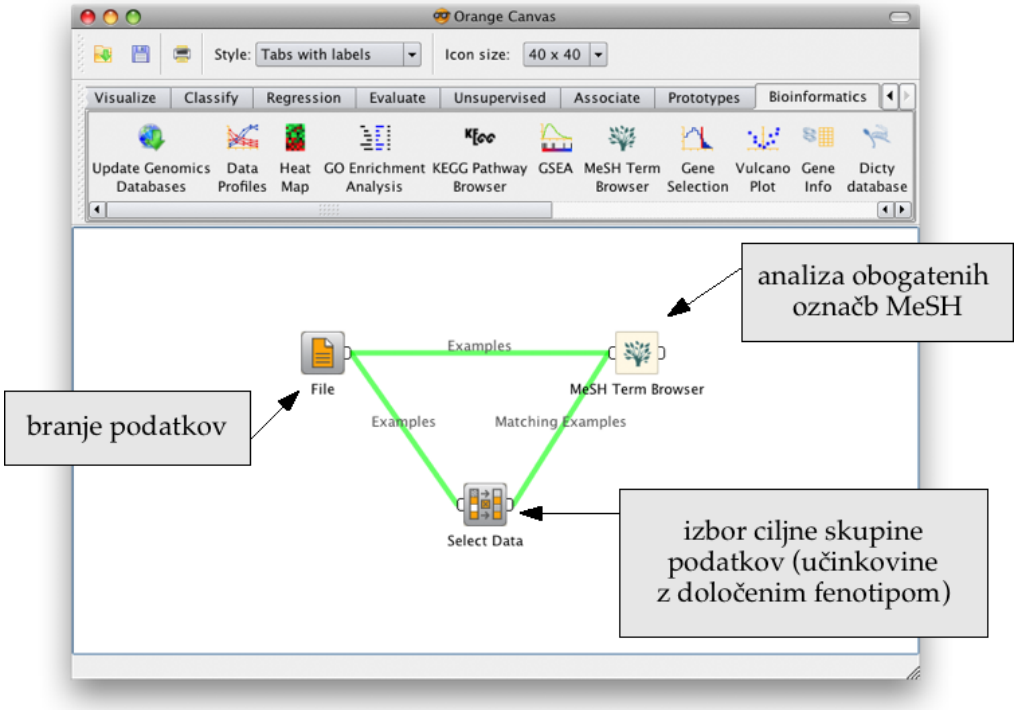


Slika 4.3 Mreža podobnosti učinkovin, kjer so podobne učinkovine (razdalja Tanimoto ≥ 0.6) v mreži povezane.

4.3 Iskanje obogatenih označb MeSH

Z uporabo metode, opisane v poglavju »Tehnika računanja obogatenih označb MeSH« ter programskim paketom Orange, smo glede na vhodne učinkovine izračunali obogatene označbe MeSH. Slika 4.4 ponazarja realizacijo sheme komponent za izračun obogatenih označb MeSH v programskemu paketu Orange. Komponenta “File” služi le za uvoz podatkovne tabele v okolje Orange, komponenta “Select data” iz vhodnih podatkov glede na vrednost GI oblikuje ciljno množico učinkovin. Komponenta “MeSH Term Browser” (glej sliko 3.4 na strani 24) pa na podlagi celotne in ciljne množice poišče obogatene označbe MeSH.

V izračunu smo lahko uporabili le učinkovine z razpoložljivimi označbami MeSH, teh je 887. Pozitivna skupina zajema 417 učinkovin z vrednostjo $GI > 13$. Pri izračunu smo se omejili le na statistično najbolj značilne označbe MeSH ($p < 0.02$), slednjih



Slika 4.4 Shema komponent za izračun obogatenih označb MeSH v programskemu paketu Orange.

smo našli 27. Deset statistično najbolj značilnih je prikazanih v tabeli 4.2, ostale se nahajajo v »Priloga B: Tabela obogatenih označb MeSH« na strani 49.

4.4 Vrednotenje tipov atributnih opisov učinkovin

Z uporabo pristopa s semensko učinkovino smo najprej primerjali med sabo kemijske opise. Izračunane p-vrednosti so zbrane v tabeli 4.3. Zaradi velikega števila vzorcev porazdelitev, razlik v porazdelitvah in moči Kolmogorov Smirnov testa so p-vrednosti manjše. Kljub temu obstajajo razlike med posameznimi kemijskimi opisi. Manjša kot je p-vrednost, manjša je tudi razlika v stopnji inhibicije rasti podobnih učinkovin.

Porazdelitve, s katerimi smo s Kolmogorov Smirnov testom izračunali p-vrednosti, so izrisane na sliki 4.5.

označba MeSH	n	p	p-vrednost
Phenothiazines	20	18	0.000127
Central Nervous System Agents	149	93	0.000231
Heterocyclic Compounds, 3-Ring	40	30	0.000576
Antifungal Agents	19	16	0.001548
Psychotropic Drugs	61	41	0.002343
Dopamine Antagonists	29	22	0.002607
Dopamine Agents	42	29	0.005889
Analgesics, Non-Narcotic	56	37	0.006053
Central Nervous System Depressants	62	40	0.008073
Antidepressive Agents	24	18	0.008098

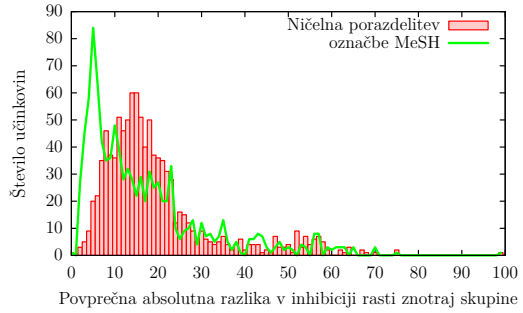
Tabela 4.2 Tabela desetih obogatenih označb MeSH z najnižjimi p-vrednostmi.

Kemijski opis	p-vrednost
označbe MeSH	$1.87 \cdot 10^{-34}$
vektor fragmentov	$2.87 \cdot 10^{-18}$
značilke QSAR	$8.35 \cdot 10^{-7}$

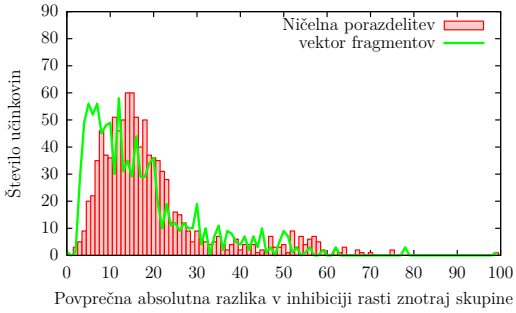
Tabela 4.3 Izračunane p-vrednosti za vse tri kemijske opise.

4.5 Vrednotenje kemijskih opisov na podskupinah učinkovin glede na označbo MeSH

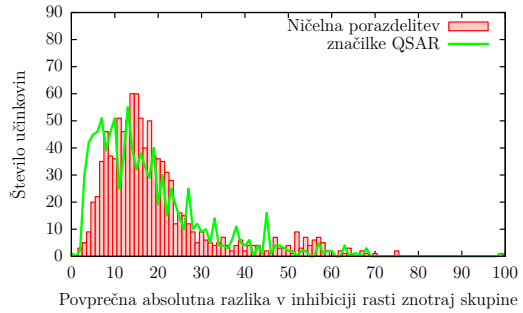
V tem primeru izbor semenske učinkovine ni bil naključen, saj smo jo izbrali iz skupine učinkovin, zaznamovanih z določeno označbo MeSH. Uporabili smo le označbe MeSH, ki označujejo več kot 30 in manj kot 600 učinkovin. Takih označb je 61 (glej »Priloga C: Tabela ovrednotenih označb MeSH«, stran 50), najboljših 10 pa je predstavljenih v tabeli 4.4. Nizka p-vrednost nakazuje višjo stopnjo korelacije med podobnostjo učinkovine znotraj skupine, določene z označbo MeSH ter fenotipom.



porazdelitev MeSH ($p = 1.87 \cdot 10^{-34}$)



porazdelitev vektorja fragmentov ($p = 2.87 \cdot 10^{-18}$)



porazdelitev QSAR ($p = 8.35 \cdot 10^{-7}$)

Slika 4.5 Porazdelitve vrednosti Δ_S za posamezne kemijske opise skupaj z ničelno porazdelitvijo.

4.6 Napovedovanje fenotipa z uporabo strojnega učenja

Kemikalije smo razdelili v dva razreda. V prvem so bile učinkovine brez inhibicije rasti ($GI < 13$), v drugem pa vse ostale. Z metodo podpornih vektorjev smo poskušali napovedati fenotipni razred. V tabeli 4.5 so predstavljene izračunane ocene AUC za vsakega izmed kemijskih opisov. Ocene so razmeroma nizke, kar pomeni, da napovedovanje fenotipa na celotni množici ni bilo uspešno.

Klasifikacijski postopek smo ponovili, le da smo tokrat učinkovine razdelili v skupine glede na njihovo MeSH anotacijo. Tokrat smo uporabili le tiste označbe MeSH, ki anotirajo manj kot 600 učinkovin in vsaj 5 učinkovin iz vsakega fenotipnega razreda. Takih označb MeSH je 135. Ocene AUC desetih najboljših skupin so zbrane v tabeli 4.6. Ocene AUC z vrednostjo 1.0 pomenijo, da lahko klasifikacijski algoritem za to skupino razlikuje med kemikalijami iz obeh razredov.

označba MeSH	št. učinkovin	označbe MeSH	vektor fragmentov	QSAR značilke
Amino Alcohols	33	$1.08 \cdot 10^{-287}$	$3.38 \cdot 10^{-163}$	$2.79 \cdot 10^{-51}$
Sulfones	39	$4.38 \cdot 10^{-251}$	$1.17 \cdot 10^{-180}$	$1.93 \cdot 10^{-28}$
Sulfonamides	38	$8.92 \cdot 10^{-248}$	$1.15 \cdot 10^{-170}$	$7.88 \cdot 10^{-16}$
Bridged Compounds	31	$3.93 \cdot 10^{-189}$	$8.19 \cdot 10^{-235}$	$1.59 \cdot 10^{-147}$
Alcohols	38	$8.00 \cdot 10^{-23}$	$5.31 \cdot 10^{-143}$	$7.01 \cdot 10^{-52}$
Bicyclo Compounds, Heterocyclic	31	$1.82 \cdot 10^{-173}$	$3.13 \cdot 10^{-209}$	$6.29 \cdot 10^{-129}$
Antipsychotic Agents	31	$3.98 \cdot 10^{-195}$	$6.31 \cdot 10^{-26}$	$8.79 \cdot 10^{-78}$
Tranquilizing Agents	41	$3.85 \cdot 10^{-169}$	$3.38 \cdot 10^{-16}$	$2.72 \cdot 10^{-56}$
Anti-Bacterial Agents	78	$3.01 \cdot 10^{-155}$	$1.49 \cdot 10^{-149}$	$2.42 \cdot 10^{-121}$
Adrenergic Agents	71	$2.34 \cdot 10^{-128}$	$5.34 \cdot 10^{-36}$	$4.63 \cdot 10^{-20}$

Tabela 4.4 Rezultati vrednotenja tipov označb učinkovin s semenskim pristopom. Semenske u;inkovine smo izbirali iz množice, ki jo opredeljuje označba MeSH.

	značilke QSAR	vektor fragmentov	označbe MeSH
AUC	0.62	0.62	0.68

Tabela 4.5 Ocene AUC za vse tri kemijske opise.

V naslednjem koraku smo zmanjšali vhodno množico učinkovin. Učinkovine brez inhibicije rasti so ostale nespremenjene, v drugem razredu pa smo uporabili le učinkovin z ekstremno inhibicijo rasti ($GI > 60$). V tem primeru so rezultati (tabela 4.7) klasifikacije na celotni množici boljši, kar nakazuje na dejstvo, da je napoved ekstremne inhibicije enostavnejša. Najboljših deset označb MeSH oziroma podskupin učinkovin je predstavljenih v tabeli 4.8. Izbor označb MeSH je enak kot v prejšnjem primeru, le da tokrat omejitvam ustreza 41 označb MeSH.

označba MeSH	**značilke QSAR**	**vektor fragmentov**	**inhibicija rasti**	**brez inhibicije rasti**
Adrenergic alpha-Antagonists	0.80	1.00	8	7
Neuromuscular Agents	0.55	1.00	8	8
Glucocorticoids	0.40	1.00	8	7
Antiparkinson Agents	1.00	0.20	7	5
Antiemetics	0.73	0.93	12	5
Antiviral Agents	0.90	0.90	5	7
Purines	0.70	0.90	5	8
Acids, Acyclic	0.70	0.90	12	7
Cyclooxygenase Inhibitors	0.62	0.90	16	7
Gastrointestinal Agents	0.67	0.86	16	18

Tabela 4.6 Ocene AUC napovedovanja fenotipa za posamezne skupine kemikalij oziroma označbe MeSH. Zadnja dva stolpca podajata distribucijo razreda.

	značilke QSAR	**vektor fragmentov**	**označbe MeSH**
AUC	0.86	0.87	0.69

Tabela 4.7 Ocene AUC za vse tri kemijske opise.

označba MeSH	**značilke QSAR**	**vektor fragmentov**	**ekstremna inhibicija rasti**	**brez inhibicije rasti**
Gastrointestinal Agents	0.82	1.00	6	18
Imidazoles	0.67	1.00	6	13
Antiemetics	1.00	1.00	5	5
Hydrocarbons, Cyclic	0.85	1.00	5	18
Sulfur Compounds	0.86	0.97	14	60
Heterocyclic Compounds, 3-Ring	0.75	0.97	12	10
Dopamine Antagonists	0.55	0.95	11	7
Autonomic Agents	0.94	0.90	7	50
Polycyclic Compounds	0.80	0.94	6	57
Hydrocarbons	0.78	0.94	6	22

Tabela 4.8 Ocene AUC napovedovanja fenotipa za posamezne skupine učinkovin oziroma označbe MeSH.

Poglavje 5

Zaključek

> Razmišljam in razmišljam mesece in leta. Devetindevedesetkrat je sklep napačen, stotič pa je pravilen.
>
> *Albert Einstein*

Eksperimentalni rezultati kažejo na to, da napovedovanje fenotipa modelnega organizma izpostavljenega poljubni učinkovini v splošnem ni uspešno (glej tabelo 4.5 na strani 37). Kljub temu smo uspeli z uporabo tehnik strojnega učenja in označb MeSH izluščiti podskupine učinkovin, za katere je napoved možna. Primer take podskupina je "Adrenergic alpha-Antagonists" v tabeli 4.6 na strani 38. Razlogov za ta pojav je več. Eden izmed pomembnejših je, da je opazovana inhibicija rasti zelo širok pojem. Različne učinkovine imajo zelo različne vplive na celični mehanizem. Tako lahko nekatere prodrejo v celico, druge povzročijo celično smrt, tretje reguliran celični odziv, nekatere pa celo ostanejo zunaj celice. Vsi našteti vplivi se lahko odražajo v zmanjšani rasti modelnega organizma v primerjavi z normalno rastjo.

Omenjeni problem smo poskušali rešiti z uporabo dodatnega ekspertnega znanja, skritega v označbah MeSH različnih učinkovin. Učinkovine smo razporedili v podskupine glede njihove označbe MeSH. Predpostavljali smo, da imajo učinkovine znotraj skupine podoben vpliv na celični mehanizem modelnega organizma. Višje ocene AUC (glej tabeli 4.5 in 4.6 straneh 37 in 38) modelov nakazujejo uspešnost tega pristopa.

Razvita metoda za izračun obogatenih označb MeSH omogoča domenskemu ekspertu (npr. biologu) strukturiran pregled nad celotnim naborom testiranih učinkovin.

Cilj metode je določitev obogatenih označb MeSH oz. podmnožic (tabela 4.2 na strani 35), ki vsebujejo statistično značilno število učinkovin iz skupine, ki povzroča opazovani fenotip na modelnem organizmu. Domenski ekspert lahko tako z uporabo svojega predznanja in seznama obogatenih označb MeSH pride novih spoznanj.

Z uporabo pristopa s semensko učinkovino smo pokazali, da vsi tipi atributnih opisov učinkovin niso enako primerni za napovedovanje fenotipa modelnega organizma. Rezultati (tabela 4.3 na strani 35) kažejo, da ontologija MeSH, kljub svoji enostavnosti, boljše ocenjuje fenotip modelnega organizma kot QSAR značilke ali vektor prisotnosti posameznega fragmenta.

Kljub temu, da so se zgornje metode izkazale za koristne na primeru *D. discoideum*, imajo tudi nekaj slabosti. Neugodna je predvsem velika odvisnost od PubChem podatkovne zbirke. Zaradi velikega števila učinkovin zbirka ni popolna, z označbami MeSH pa so anotirane le bolj aktualne učinkovine. Učinkovin z manjkajočimi označbami MeSH ne moremo analizirati z razvitimi metodami.

Skoraj z gotovostjo lahko trdimo, da bo napovedovanje vpliva učinkovin na fenotip modelnih organizmov v prihodnosti zelo napredovalo. O posameznih učinkovinah je dostopnih vedno več uporabnih informacij, pričakovati pa je tudi, da bo tehnologija [27] za merjenje izraženosti genov tako napredovala, da bo časovno in cenovno primerna za masovne eksperimente. Takrat bomo lahko posamezne učinkovine, namesto z inhibicije rasti modelnega organizma, povezovali kar z izrazom posameznih genov.

Literatura

[1] K. G. Joback in R. C. Reid, Estimation of pure-component properties from group-contributions v *Chemical Engineering Communications*, št. 1, zv. 57, str. 233-243, 1987.

[2] L. Umek, P. Kaferle, M. Mattiazzi, A. Erjavec in C. Gorup et al., A subgroup discovery approach for relating chemical structure and phenotype data in chemical genomics v *Third International Workshop on Machine Learning in Systems Biology*, sprejeto v objavo, 2009.

[3] J. Klekota in F. P. Roth, Chemical substructures that enrich for biological activity. v *Bioinformatics*, št. 21, zv. 24, str. 2518-25, 2008.

[4] R. D. King, S. H. Muggleton, A. Srinivasan in M. J. Sternberg, Structure-activity relationships derived by machine learning: the use of atoms and their bond connectivities to predict mutagenicity by inductive logic programming. v *Proceedings of the National Academy of Sciences USA*, št. 1, zv. 93, str. 438-42, 1996.

[5] C. Helma, *Predictive toxicology*, Taylor in Francis Group, 2005.

[6] A. L. Milac, S. Avram in A. J. Petrescu, Evaluation of a neural networks QSAR method based on ligand representation using substituent descriptors application to hiv-1 protease inhibitors v *Journal of Molecular Graphics and Modelling*, št. 1, zv. 25, str. 37-45, 2006.

[7] C. Y. Zhao, H. X. Zhang, X. Y. Zhang, M. C. Liu in Z. D. Hu et al., Application of support vector machine (SVM) for prediction toxic activity of different data sets. v *Toxicology*, št. 2-3, zv. 217, str. 105-19, 2006.

[8] I. Massarelli, M. Imbriani, A. Coi, M. Saraceno in N. Carli et al., Development of QSAR models for predicting hepatocarcinogenic toxicity of chemicals. v *European Journal of Medicinal Chemistry*, št. 9, zv. 44, str. 3658-64, 2009.

[9] U. Lahl in U. Gundert-Remy, The use of (Q)SAR methods in the context of reach v *Toxicology Mechanisms and Methods*, št. 2-3, zv. 18, str. 149-158, 2008.

[10] N. Cristianini in J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*, Cambridge University Press, 2000.

[11] Q. Xu in G. Shaulsky, Goat: An R tool for analysing gene ontology term enrichment. v *Applied Bioinformatics*, št. 4, zv. 4, str. 281, 2005.

[12] Q. Zheng in Xiu-Jie J. Wang, GOEAST: a web-based software toolkit for gene ontology enrichment analysis. v *Nucleic Acids Research*, št. Web Server issue, zv. 36, str. W358-63, 2008.

[13] (2009) *Programski paket Orange*. Dostopno na: http://www.ailab.si/orange

[14] (2009) *Programski paket Dragon*. Dostopno na: http://talete.mi.it/

[15] (2009) *MeSH ontologija*. Dostopno na: http://www.nlm.nih.gov/mesh/

[16] (2009) *Podatkovna zbirka PubMed*. Dostopno na: http://www.ncbi.nlm.nih.gov/pubmed/

[17] (2009) *Podatkovna zbirka PubChem*. Dostopno na: http://pubchem.ncbi.nlm.nih.gov/

[18] (2009) *Programski paket Open Babel*. Dostopno na: http://openbabel.org

[19] I. H. Witten in F. Eibe, *Data Mining: practical machine learning tools and techniques: practical machine learning tools and techniques*, San Francisco: Morgan Kaufman, Elsevier, 2005.

[20] V. Vapnik in A. Chervonenkis, Theory of pattern recognition: Statistical problems of learning v *Nauka, Moscow. Math. Review*, 1974.

[21] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang in Chih-Jen Lin, LIBLINEAR: A library for large linear classification v *Journal of Machine Learning Research*, 2008.

[22] J. A. Hanley in B. J. McNeil, The meaning and use of the area under a receiver operating characteristic (ROC) curve v *Radiology*, št. 1, zv. 143, str. 29-36, 1982.

[23] T. Fawcett, An introduction to ROC analysis v *Pattern recognition letters*, št. 8, zv. 27, str. 861-874, 2006.

[24] G. Shaulsky, The cheatin' amoeba v *The Scientist*, št. 7, zv. 22, str. 30-36, 2008.

[25] L. Eichinger, J. A. Pachebat, G. Glőckner, M. A. Rajandream in R. Sucgang et al., The genome of the social amoeba Dictyostelium discoideum v *Nature*, 2005.

[26] T. M. J. Fruchterman in E. M. Reingold, Graph drawing by force-directed placement v *Software: Practice and Experience*, št. 11, zv. 21, str. 1129-1164, 1991.

[27] J. Shendure in H. Ji, Next-generation DNA sequencing. v *Nature Biotechnology*, št. 10, zv. 26, str. 1135-45, 2008.

Priloge

Priloga A: Algoritmi pristopa s semensko učinkovino

Algoritem za izračun ničelne porazdelitve

1. $D = []$.
2. Ponovi 1000 krat.
   - Naključno izberi semensko učinkovino C_S ter poišči njeno inhibicijo rasti $gi(C_S)$.
   - Naključno izberi 10 učinkovin $(C_1, C_2, \cdots, C_{10})$.
   - Izračunaj povprečno absolutno razliko δ_{GI} med inhibicijo semenske in naključno izbrane učinkovine C_i. $\delta_{GI} = \frac{\sum_{i=1}^{10} abs(gi(S) - gi(C_i))}{10}$.
   - Dodaj rezultat δ_{GI} na seznam D.

Algoritem za vrednotenje atributnih opisov učinkovin

1. $A = []$.
2. Ponovi 1000 krat.
   - Naključno izberi semensko učinkovino C_S ter poišči njeno inhibicijo rasti $gi(C_S)$.
   - Glede na izbrani kemijski opis in funkcijo podobnosti izberi 10 učinkovin $(C_1, C_2, \cdots, C_{10})$, ki so najbolj podobne C_S.
   - Izračunaj povprečno absolutno razliko δ_{GI} med inhibicijo semenske in izbrane učinkovine C_i. $\delta_{GI} = \frac{\sum_{i=1}^{10} abs(gi(S) - gi(C_i))}{10}$.
   - Dodaj rezultat δ_{GI} na seznam A.
3. Z uporabo Kolmogorov-Smirnov testa izračunaj p-vrednost med seznamoma A in D.

Algoritem za vrednotenje atributnih opisov učinkovin glede na označbe MeSH

1. Za vsako označbo MeSH M:
   - Poišči podmnožico učinkovin T, ki so označene z označbo MeSH M.
   - $A = []$.
   - Ponovi 100 krat.
     1. Naključno izberi semensko učinkovino C_S iz množice T ter poišči njeno inhibicijo rasti $gi(C_S)$.
     2. Glede na izbrani kemijski opis in funkcijo podobnosti izberi 10 učinkovin $(C_1, C_2, \cdots, C_{10})$, ki so najbolj podobne C_S.
     3. Izračunaj povprečno absolutno razliko δ_{GI} med inhibicijo semenske učinkovine in naključno izbrane učinkovine C_i. $\delta_{GI} = \frac{\sum_{i=1}^{10} abs(gi(S) - gi(C_i))}{10}$.
     4. Dodaj rezultat δ_{GI} na seznam A.
   - Za označbo MeSH M izračunaj z uporabo Kolmogorov-Smirnov testa p-vrednost med seznamoma A in D.

Priloga B: Tabela obogatenih označb MeSH

označba MeSH	**n**	**p**	**p-vrednost**
Phenothiazines	20	18	0.000127
Central Nervous System Agents	149	93	0.000231
Heterocyclic Compounds, 3-Ring	40	30	0.000576
Antifungal Agents	19	16	0.001548
Psychotropic Drugs	61	41	0.002343
Dopamine Antagonists	29	22	0.002607
Dopamine Agents	42	29	0.005889
Analgesics, Non-Narcotic	56	37	0.006053
Central Nervous System Depressants	62	40	0.008073
Antidepressive Agents	24	18	0.008098
Anti-Inflammatory Agents	71	45	0.008101
Sensory System Agents	77	48	0.009915
Antipsychotic Agents	31	22	0.010060
Hydrocarbons, Cyclic	52	34	0.010800
Hormones, Hormone Substitutes, and Hormone Antagonists	45	30	0.011100
Anesthetics	18	14	0.011850
Antirheumatic Agents	56	36	0.012940
Norsteroids	6	6	0.013670
Norpregnanes	6	6	0.013670
Contraceptive Agents, Female	6	6	0.013670
Benzylidene Compounds	6	6	0.013670
Estrogens	6	6	0.013670
Contraceptives, Oral	6	6	0.013670
Contraceptives, Oral, Synthetic	6	6	0.013670
Anti-Inflammatory Agents, Non-Steroidal	49	32	0.013700
Contraceptive Agents	9	8	0.016550
Calcium Channel Blockers	17	13	0.019390

Priloga C: Tabela ovrednotenih označb MeSH

označba MeSH	**št. učinkovin**	**označbe MeSH**	**vektor fragmentov**	**značilke QSAR**
Amino Alcohols	33	1.08E-287	3.38E-163	2.79E-051
Sulfones	39	4.38E-251	1.17E-180	1.93E-028
Sulfonamides	38	8.92E-248	1.15E-170	7.88E-016
Azabicyclo Compounds	30	4.02E-185	2.81E-241	1.21E-148
Bicyclo Compounds	30	1.02E-181	2.50E-239	1.75E-141
Bridged Compounds	31	3.93E-189	8.19E-235	1.59E-147
Alcohols	38	8.00E-230	5.31E-143	7.01E-052
Bicyclo Compounds, Heterocyclic	31	1.82E-173	3.13E-209	6.29E-129
Antipsychotic Agents	31	3.98E-195	6.31E-026	8.79E-078
Tranquilizing Agents	41	3.85E-169	3.38E-016	2.72E-056
Anti-Bacterial Agents	78	3.01E-155	1.49E-149	2.42E-121
Adrenergic Antagonists	33	6.43E-140	2.93E-037	1.27E-020
Amides	72	2.17E-110	1.60E-132	1.07E-025
Adrenergic Agents	71	2.34E-128	5.34E-036	4.63E-020
Respiratory System Agents	38	5.45E-128	4.15E-073	8.42E-030
Amines	94	2.29E-125	3.70E-054	3.04E-026
Central Nervous System Depressants	62	3.53E-125	6.11E-013	6.47E-041
Psychotropic Drugs	61	9.82E-092	9.55E-013	2.07E-031
Sulfur Compounds	108	9.96E-081	1.06E-067	8.40E-014
Heterocyclic Compounds. 3-Ring	40	3.15E-051	8.27E-064	3.07E-080
Antihypertensive Agents	60	1.68E-074	1.03E-019	1.94E-018
Autonomic Agents	88	2.73E-071	1.22E-052	2.23E-015
Dopamine Agents	42	1.73E-069	9.96E-003	8.48E-030
Cholinergic Agents	30	1.07E-053	1.46E-030	2.69E-045
Anti-Infective Agents	176	3.94E-048	5.76E-052	1.47E-029

označba MeSH	**št. učinkovin**	**označbe MeSH**	**vektor fragmentov**	**značilke QSAR**
Toxic Actions	30	2.27E-051	3.19E-035	1.02E-026
Azoles	48	3.62E-050	2.07E-015	5.64E-022
Alkaloids	44	2.41E-045	3.86E-047	3.12E-007
Cardiovascular Agents	131	7.47E-046	8.52E-028	3.22E-015
Heterocyclic Compounds, 2-Ring	113	1.54E-037	6.65E-031	7.54E-010
Membrane Transport Modulators	31	1.26E-022	1.60E-034	1.37E-031
Acids, Carbocyclic	32	1.91E-029	1.61E-015	1.78E-007
Carboxylic Acids	55	3.19E-029	3.37E-005	2.42E-009
Polycyclic Compounds	114	1.16E-028	7.64E-026	2.90E-026
Organic Chemicals	341	3.35E-026	5.31E-023	4.40E-020
Sensory System Agents	77	4.52E-025	7.56E-009	9.90E-010
Central Nervous System Agents	149	1.90E-024	5.76E-006	1.61E-010
Gastrointestinal Agents	34	3.33E-024	3.65E-012	3.16E-008
Specialty Uses of Chemicals	54	1.84E-003	2.24E-022	4.79E-016
Hydrocarbons	59	6.79E-007	4.54E-011	1.55E-021
Polycyclic Hydrocarbons, Aromatic	31	2.16E-017	1.58E-010	1.56E-021
Neurotransmitter Agents	189	1.90E-020	1.02E-006	1.25E-010
Antineoplastic Agents	39	1.99E-020	2.24E-010	9.33E-016
Hydrocarbons, Cyclic	52	1.12E-005	6.44E-006	5.05E-020
Peripheral Nervous System Agents	176	1.04E-019	2.08E-015	9.01E-005
Molecular Mechanisms of Pharmacological Action	364	1.58E-019	4.24E-009	2.28E-016
Hydrocarbons, Aromatic	45	1.16E-007	4.65E-008	1.80E-019
Hormones, Hormone Substitutes, and Hormone Antagonists	45	2.16E-017	8.75E-019	4.31E-010

označba MeSH	**št. učinkovin**	**označbe MeSH**	**vektor fragmentov**	**značilke QSAR**
Physiological Effects of Drugs	405	1.53E-016	7.44E-010	3.52E-003
Anti-Inflammatory Agents	71	3.24E-016	2.98E-014	8.41E-006
Analgesics	61	1.96E-015	3.71E-004	2.71E-003
Anti-Inflammatory Agents, Non-Steroidal	49	1.96E-015	7.40E-005	2.92E-009
Enzyme Inhibitors	114	3.89E-010	3.09E-010	2.01E-014
Heterocyclic Compounds	315	5.84E-014	1.18E-013	2.20E-006
Heterocyclic Compounds, 1-Ring	159	2.34E-012	2.93E-007	1.35E-009
Analgesics, Non-Narcotic	56	4.30E-011	8.68E-008	1.72E-007
Antirheumatic Agents	56	2.75E-010	2.40E-004	6.54E-007
Imidazoles	30	5.66E-010	7.42E-009	1.07E-007
Vasodilator Agents	46	5.81E-006	2.39E-006	4.30E-009
Antiparasitic Agents	44	9.70E-007	4.49E-005	2.76E-005
Steroids	43	1.05E-005	7.09E-006	5.92E-003

Izjava o samostojnosti dela

Podpisani Črtomir Gorup, z vpisno številko 63030160, sem avtor diplomskega dela z naslovom:

RAČUNSKE TEHNIKE NAPOVEDOVANJA VPLIVA UČINKOVIN NA FENOTIP MODELNIH ORGANIZMOV

S svojim podpisom zagotavljam, da:

- sem diplomsko delo izdelal samostojno pod mentorstvom prof. dr. Blaža Zupana,
- so elektronska oblika diplomskega dela, naslov (slov., angl.), povzetek (slov., angl.) ter ključne besede (slov., angl.) identični s tiskano obliko diplomskega dela,
- soglašam z javno objavo elektronske oblike diplomskega dela v zbirki "Dela FRI".

Ljubljana, 3.9.2009 Črtomir Gorup