Računska orodja za odkrivanje prognostičnih bioloških označevalcev iz podatkov o genskih izrazih: opis raziskovalnega projekta

Predlagamo projekt za oblikovanje in razvoj interaktivne zbirke orodij za pomoč pri iskanju molekularnih prognostičnih bioloških označevalcev (biomarkerjev) iz molekularnih podatkov in podatkov preživetja, pridobljenih v kliničnih poskusih. V projektu bomo zasnovali računske metode in metode strojnega učenja za iskanje biomarkerjev, jih vključil v interaktivne komponente z grafičnim uporabniškim vmesnikom in zasnovali vizualno programiranje za povezavo teh komponent v cevovode. Razvite metode in zbirka orodij bodo podpirale sodelovanje med podatkovnimi znanstveniki in strokovnjaki s področja razvoja bioloških označevalcev - zdravniki, biomedicinski ali farmacevtski raziskovalci. Z njimi bo možno preiskati podatke o molekularnih odzivih tisočih genov, da bi našli tiste, ki najbolj korelirajo s preživetjem. Predlagano orodje bo dostopalo do obstoječih modelov, ontologij in podatkovnih baz, da bi tako pospešilo interpretacijo in ponudilo polavtomatske razlage rezultatov.

To je aplikativni projekt, pri katerem se povezujemo z Genialisom, specializiranim podjetjem za podatkovno znanost na področju k posamezniku usmerjene medicine (ang. *precision medicine*). Genialis je trenutno v postopku registracije pri FDA (Ameriška uprava za hrano in zdravila) prvega modela strojnega učenja, ki za napovedovanje odziva na zdravljenje bolnikov z rakom uporablja podatke o transkripciji. Genialis potrebuje metode, orodja in vizualizacije za pospešitev raziskav, da bo ostalo v vrhu raziskav biomarkerjev in izboljšalo komunikacijo rezultatov s strankami in regulatornimi agencijam. Po drugi strani pa bo projekt predlagatelju omogočil nadaljevanje naših raziskav interaktivnih vizualizacij in strojnega učenja ter uporabo naših novih pristopov k zahtevnemu področju odkrivanja biomarkerjev.

27.1 Znanstvena izhodišča ter predstavitev problema in ciljev raziskav

Znanstvena izhodišča

Projekt bo prispeval nove metode in praktične pristope k raziskovanju podatkov na področju analize preživetja in preučeval, koliko spremenljivk (genov s svojim izražanjem) skupaj vpliva na preživetje. Analiza preživetja je sklop statističnih metod, katerih namen je določiti pričakovano življenjsko dobo preiskovane populacije. Analiza preživetja proučuje pričakovano trajanje časa do nekega dogodka, recimo ponovne pojavitve raka po kemoterapiji [22]. Modeli preživetja, vključno z najbolj znanim modelom razmerja tveganja (ang. *hazard ratio*), se nanašajo na čas do dogodka za eno ali več kovariant. V biomedicini so spremenljivke, ki pomembno vplivajo na preživetje, potencialni označevalci, značilnost biološkega sistema, ki ga lahko merimo objektivno in uporabljamo kot kazalnik stanja sistema. Na primer, pri raku lahko označevalci razlikujejo med bolniki, ki se odzivajo na zdravljenje, in tiste, ki se ne.

Na podlagi markerjev lahko napovemo uspeh zdravljenja in izberemo pravo terapijo za posameznega pacienta. Identifikacija dobrih označevalcev je tako ključnega pomena za razvoj personalizirane medicine. Ker izboljšanje preživetja neposredno koristi bolniku, je nujno razumeti, kako se udeleženci odzivajo na različne oblike zdravljenja. Zato je treba zdravljenje izbrati glede na bolnikovo stanje in značilnosti, ki so definirane s pomočjo nabora markerjev. Označevalci so lahko klinični, povezani s pacientovimi simptomi, ali biološki, povezani z nekaterimi meritvami na molekularni ravni, kot je koncentracija določenega proteina ali izražanje določenih genov. Označevalci se lahko nanašajo na posamezno meritev ali skupino meritev, po možnosti povezanih s prognostičnim modelom ali omrežjem [27].

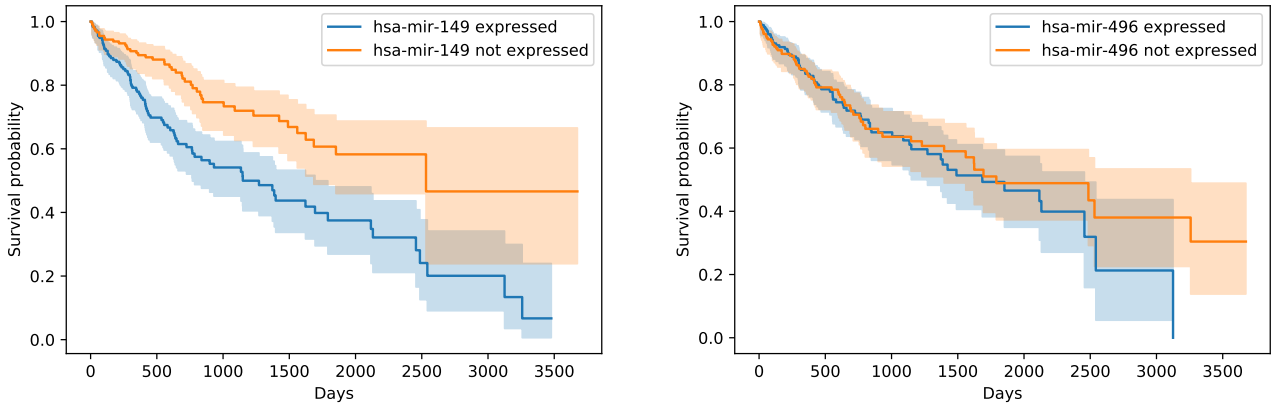

Fig. 1. Primer Kaplan-Meierjeve preživetvene krivulje za dve skupini, definirani z izražanjem genov. Preživetje je bistveno večje pri skupini bolnikov z močno izraženimi microRNA hsa-mir-149 (leva slika). Razlika pri mikroRNA hsa-mir-496 (desna slika) ni tako očitna. Lahko bi rekli, da je hsa-mir-149 torej boljši biomarker za preživetje. Pri odkrivanju biomarkerjev je ena od nalog razvrščanje genov glede na stopnjo ločevanja skupin z različnim preživetjem.

Z visokozmogljivim sekvenciranjem lahko merimo stopnjo aktivnosti genov v bioloških vzorcih. V zadnjih letih se pozornost usmerja s preprostih eno-genskih markerjev na nivoju DNA na kompleksne večgenske označevalce na nivoju izražanja genov. oličina mRNA v biološkem vzorcu, ki ustreza določenemu genu, je povezana z aktivnostjo gena in se imenuje izražanje genov. Visoko zmogljivo sekvenciranje omogoča določitev izražanje vseh genov v organizmu in tako predstavlja orodje za oceno stanja biološkega sistema. Gen lahko štejemo za biomarker preživetja, če je funkcija preživetja zelo drugačna v subpopulaciji, kjer je gen izražen, v primerjavi s tisto, kjer gen ni izražen (glej sliko 1). Ta opredelitev je nejasna, saj zahteva prag za izražanje genov, količinsko opredelitev in naknaden prag razlike med funkcijami preživetja. Poleg tega se skupine genov in ne posamezni geni običajno uporabljajo za zajem različnih biologij v vzorcu, npr. nagnjenosti k nastanku krvnih žil ali pripravljenosti na imunski odziv. Preverjen in učinkovit postopek odkrivanja biomarkerjev, ki temelji na izražanju genov, je lahko neverjetno dragoceno in nujno orodje pri odkrivanju, razvoju in diagnostičnih raziskavah zdravil [21].

V idealnem primeru bi torej podatkovno odkritje novih biomarkerjev zahtevalo le podatke o preživetju z ustreznimi profili izražanja genov. Algoritmi bi nato presejali vse gene in poiskali tiste, ki najbolje opredeljujejo skupine z različnimi funkcijami preživetja. Obstaja pa veliko težav in izzivov v tem postopku povezanih s šumom v podatkih, z majhnim številom preiskovanih vzorcev, z interakcijami med geni in vključevanjem razpoložljivega dodatnega znanja (podatkovnih baz). S tem se ukvarjamo podrobneje v nadaljevanju.

Predstavitev problema

Predlagani projekt obravnava tri kategorije problemov in izzivov, ki vplivajo na računske metode za ugotavljanje potencialnih biomarkerjev, pristopov k fuziji podatkov in izvedbe:

Računski izzivi, šum in prekomerno prilagajanje modela učni množici (ang. *overfitting*). Eksperimentalni podatki za preživetje, povezani z npr. vplivom novega zdravljenja ali zdravila, so pogosto dragi in zato so eksperimentalni vzorci majhni. V tipičnem kliničnem poskusu prve faze je pogosto manj kot 30 bolnikov, v 2. kliničnem poskusu pa je več kot sto bolnikov prej izjema kot pravilo. Podatkovni šum, povezan z načinom zbiranja, obdelavo vzorcev in merjenjem ekspresije genov, je lahko visok. Tako okolje lahko vodi do lažnih odkritij in prekomernega prilagajanja modela učni množici. Težava je še posebej očitna pri iskanju skupin ali mrež genov, ki bi lahko služili kot biomarkerji, saj število kandidatov (različnih skupin genov) raste eksponentno z želeno velikostjo nabora genov za biomarker. Na primer, za 20000 genov, ki kodi-

rajo beljakovine, obstaja več kot 1.3 bilijonov možnih genskih trojk. Tudi če bi nam uspelo vse računsko preučiti, bi to nujno povzročilo prekomerno prilagajanje modela učni množici, zato bi se rezultati dobro opisovali učne podatke, vendar bi se slabo posploševali na nove primere. Poleg šuma in prekomernega prilagajanja modela (ang. *overfitting*) računski izzivi vključujejo še iskanje pragov ekspresije genov (kdaj je gen izražen?) in agregacijske funkcije (kdaj je skupek genov aktiven?).

Vključevanje podatkovnih baz. Geni sodelujejo v molekularnih poteh, funkcionalnih skupinah in odzivih na kemikalije in zdravila. Znanje o tem in druge genske anotacije so shranjene v podatkovnih bazah, kot so GeneOntology[1], KEGG[2], CellMarker[3] in druge. Preučevanje genskih kompletov kot kandidatov za biomarkerje bi lahko in bi moralo uporabiti te dragocene vire informacij; tako za omejevanje prostora za iskanje biomarkerjev kot za razlago naborov najboljših kandidatnih genov (obogatitev genskega nabora). Takšno združevanje podatkov in baz znanja je že prineslo zanimive rezultate v bioinformatskih raziskavah [40, 39], vendar je bilo na področju odkrivanja biomarkerjev preživetja premalo raziskano.

Vmesnik za raziskovanje podatkov. V zadnjih dveh desetletjih so se pojavile različne metode, statistična orodja in orodja za strojno učenje za analizo visoko zmogljivih podatkov iz molekularne biologije. Analizi preživetja pa manjka elegantna zbirka orodij z intuitivnim uporabniškim vmesnikom, ki bi pomagala pri odkrivanju biomarkerjev, podpirala interaktivno analizo raziskovalnih podatkov v realnem času in ponudila enostavno izdelavo analitičnih podatkovnih cevovodov. Na voljo so dobre računalniške knjižnice za analizo preživetja v jezikih R in Python, vendar so to le nepovezani gradniki, ki za uporabo in integracijo zahtevajo napredne programersko znanje. Namesto tega končni uporabniki in znanstveniki s področja potrebujejo orodja za komunikacijo, raziskovanje in modeliranje podatkov ter intuitivna orodja s prilagodljivimi interaktivnimi vmesniki.

V projektu bomo te tri izzive napadli z razvojem novih tehnik in orodij za interaktivno raziskovanje potencialnih prognostičnih biomarkerjev. Naš cilj je demokratizirati to področje z razvojem metod in interaktivnega vmesnika za inteligentno analizo podatkov preživetja.

Cilji projekta

Projekt bo razvil in uporabil nabor računskih orodij za odkrivanje biomarkerjev iz podatkov genskega izražanja in kliničnega preživetja. Vključili bomo obstoječe pristope k ocenjevanju biomarkerjev na podlagi podatkov o preživetju, modeliranje preživetja in analize obogatitve genskega nabora. Predlagali bomo tudi nove tehnike za analizo preživetja glede na specifične genske interakcije, za izdelavo omrežij markerskih genov in interpretacijo vizualizacij. Razvili bomo metode za hevrističnega iskanje, ki bo uporabljalo objavljene podatkovne baze o genskih funkcijah in presnovnih poteh.

Projekt bo opolnomočil domenske strokovnjake za uporabo teh orodij v resničnih aplikacijah, v realnem času in brez potrebe po pisanju računalniške kode. Projekt bo vključeval računske metode v komponente z grafičnim uporabniškim vmesnikom. Izboljšali bomo lastno odprtokodno platformo za rudarjenje podatkov Orange[4] [8, 7, 11] (slika 2) z novimi metodami za analizo preživetja. Pokazali bomo, da nastala platforma za vizualno programiranje ne samo bistveno zmanjša zapletenost in čas, porabljen za analizo podatkov, ampak tudi izboljša sodelovanje različnih strokovnjakov z informativnimi vizualizacijami in avtomatsko uporabo domenskega znanja.

[1] `http://geneontology.org`
[2] `https://www.genome.jp/kegg`
[3] `http://biocc.hrbmu.edu.cn/CellMarker`
[4] `http://orangedatamining.com`

Na koncu bomo predstavili še uporabnost izdelanega orodja. Uporabili bomo nabor javnih in zasebnih (od sodelujočega podjetja Genialis) podatkov genske ekspresije in ustreznih kliničnih izidov. O uspešnosti projekta se bo presojalo glede na predstavljene primere uporabe in našo zmožnostjo, da nove uporabnike usposobimo za samostojno uporabo razvite tehnologije.

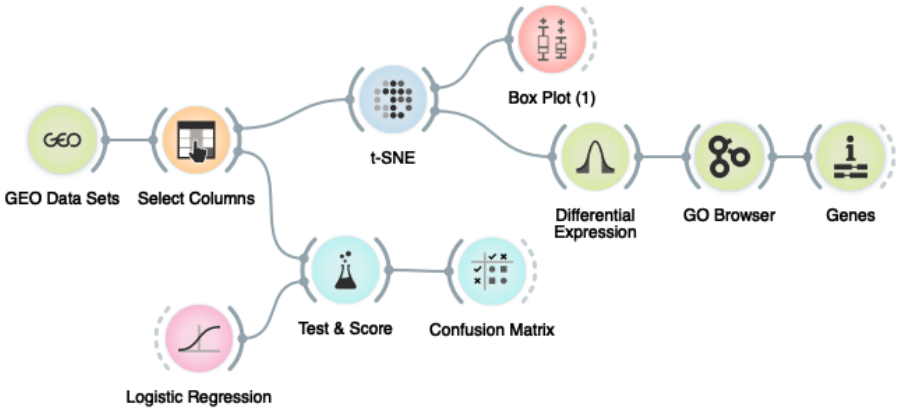

Fig. 2. Tipičen potek dela v programu Orange. Slika prikazuje primer, ko smo ponovno analizirali gensko izražanje v mononuklearnih celicah periferne krvi. V raziskavi (GDS5363) so avtorji ugotavljali, ali je profiliranje ekspresije genov lahko zaznalo pojav osteoartritisa [24]. Potek dela naloži podatke iz baze Gene Expression Omnibus in definira odvisne spremenljivke (stanje bolezni, izbrane komponente *Select Columns*). Zgornja veja postopka preveri strukturo vzorcev tkiva (komponenta *t-SNE*), za izbrane vzorce najde različno izražene gene in analizira njihove skupne značilnosti z obogatitvijo izrazov genske ontologije. V spodnji veji preverimo hipotezo preiskovalcev neposredno in ocenimo natančnost napovedi logističnega regresijskega modela s pomočjo navzkrižne validacije (komponenta *Test & Score*). V predlaganem projektu bomo razvili podoben potek dela, vendar s komponentami, ki bodo naložile, obdelale, analizirale in prikazale podatke o preživetju ter predlagale nove biomarkerje.

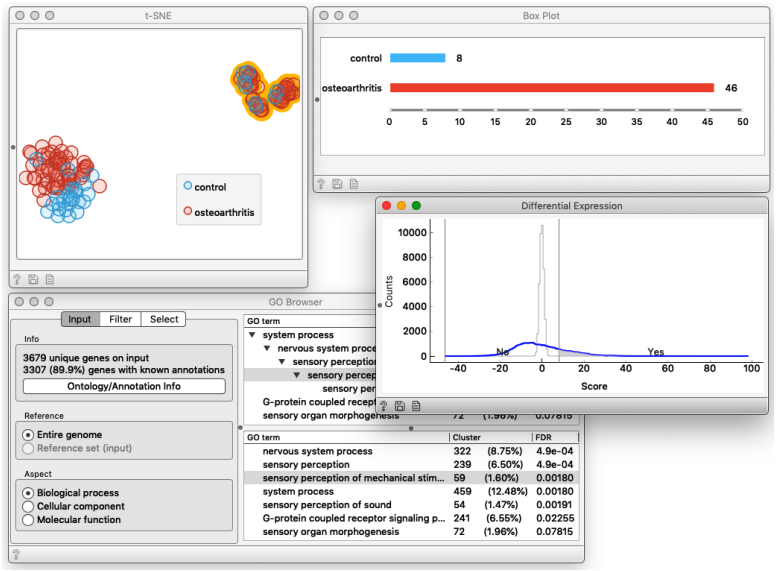

Fig. 3. Večina grafičnih komponent v programu Orange je interaktivnih. Tukaj prikazujemo vsebino več komponent iz poteka dela v Sliki 2. Uporabnik lahko na primer izbere podmnožico podatkovnih točk iz vizualizacije *t-SNE* (podatkovne točke z rumenim obrisom zgoraj desno) ali ustrezne podatke v *Box Plot* ali nabor genov, ki so povezani z izbranim izrazom iz modula *GO Browser*. V pripomočku *Differential Expression* lahko izberemo nabor različno izraženih genov v repih porazdelitve. Večina modulov omogoča različne ravni interakcije in medsebojnega povezovanja. Projekt bo nekatere vizualizacijske komponente, vključno s *t-SNE*, ponovno uporabil in razvil specializirane, visoko interaktivne module za analizo preživetja in odkrivanje kompleksnih biomarkerjev.

Pričakovani rezultati

Najpomembnejši pričakovani rezultati tega projekta so:

1. **Novi pristopi za odkrivanje biomarkerjev**, ki vključujejo računske pristope za analizo interakcij biomarkerjev iz podatkov o preživetju, vizualizacijske pristope za pomoč pri interpretaciji, in izboljšano preisko-

vanje prostora skupin biomarkerjev z vključevanjem drugih podatkov in baz.

2. **Knjižnica računalniške kode z metodami za odkrivanje biomarkerjev iz podatkov o preživetju in genskem izražanju.** Knjižnica bo razvita v jeziku Python in bo objavljena v odprtokodni različici na GitHub skupaj z dokumentacijo, testi in primeri uporabe.

3. **Nabor orodij za odkrivanje biomarkerjev z vmesnikom za vizualno programiranje, interaktivnimi vizualizacijami in razlago rezultatov.** Nabor orodij bomo združili z javnimi podatkovnimi bazami in tako omogočili gradnjo analitičnih cevovodov v realnem času, in interaktivno vizualizacijo podatkov in modelov (glej sliko 3), podprto z domenskim znanjem.

4. **Nabor primerov uporabe, razvit v tesnem sodelovanju z Genialis.** Primeri uporabe bodo prikazali uporabnost naše programske opreme, moč intuitivnega grafičnega vmesnika in zagotoviti učno gradivo za razširjanje rezultatov projekta.

27.2 Pregled literature na področju odkrivanja prognostičnih biomarkerjev

V literaturi se za primerjavo preživetje med dvema različnima skupinama najpogosteje uporabljata log-rank test in Coxov model sorazmernega tveganja [26]. Ti dve metodi lahko štejemo za izhodišče pri oblikovanju strategij za odkrivanje prognostičnih biomarkerjev. Na kratko, iskanje markerskih genov lahko razdelimo na dva pod-problema: združevanje odzivnih spremenljivk (tu izražanja genov) in iskanje obetavnih skupin genov s prognostično močjo.

V kliničnih študijah so biološki markerji običajno zvezne spremenljivke pridobljene z različnimi meritvami. Vzpostavitev mejne točke, ki predstavlja mejo med visoko in nizko ekspresijo genov ali bolj splošno, ki razlikuje med skupinami z visokim in nizkim tveganjem, je lahko bistvenega pomena za njihovo uporabo pri kliničnih odločitvah [20]. Budczies et al. [3] predlagajo več pristopov k izbiri mejnih vrednosti: glede na porazdelitev biološkega markerja z optimizacijo soodvisnosti ciljne spremenljivke, kot je odziv na zdravljenje, ali z iskanjem najmanjše vrednosti p. Slednji je najpogostejši in izbere mejno vrednost glede na optimalno razliko v napovedi izida preživetja med skupinami [33]. Na splošno pa je iskanje optimalne vrednosti težaven problem, ki je odvisen tudi od zasnove študije. Pravilni postopki za določitev mejne vrednosti so zelo pomembni, saj lahko precenimo dejanski učinek biološkega markerja [1].

Witten et al. [32] izpostavljajo problem iskanja napovednih lastnosti v visokodimenzionalnih podatkih. Ko je število spremenljivk veliko večje od števila primerov, običajni statistični pristopi k analizi preživetja niso več zadostni. Obstaja veliko različnih objavljenih pristopov k iskanju marker genov. Nekateri priporočajo dvostopenjsko filtriranje: najprej filtriranje različno izraženih genov (genov, ki se dobro ločijo med izbranimi skupinami) in nato nadaljnje zožitev nabora možnih kandidatov na podlagi statistične pomembnosti pri analizi preživetja [30, 17, 38, 15]. Relator et al. [25] so kritični do takšnih pristopov, ker lahko pustijo številne možne kombinacije genov nepreizkušene. Predlagajo rešitev, ki lahko zazna interakcije potencialnih označevalcev, ki bi jih običajni pristopi izpustili. Pomembna pomanjkljivost njihovega pristopa, kot priznavajo sami, je računska zapletenost. Predlagajo razdelitev podatkov na manjše vzorce, uporabo predlagane rešitve na podvzorcih, in nato kombinacijo rezultatov. Vendar taka rešitev znova lahko izpusti pomembne genske interakcije.

Pri zapletenih boleznih, kot je rak, so učinki izražanja genov na preživetje praviloma nelinearni. Za odkrivanje nelinearnih genskih odnosov so se v zadnjem času pojavili različni pristopi s tehnikami globokega učenja [12]. Za napovedovanje preživetja je bilo predlaganih več različnih modelov globokega učenja, vključno s standardnim Coxovovim modelom sorazmernega tveganja (Cox-nnet [6], SurvivalNet [35], DeepSurv [14]). Kljub naprednim tehnikam in povečanju števila potencialnih bioloških markerjev jih je bilo zelo malo klinično uporabljenih [4]. Če je odkrite označevalce težko razložiti, so ti premalo raziskani ali brez znanih bioloških funkcij, so za kliniko neuporabni. Z modeli globokega učenja je ta izziv toliko večji. Hao et al. [12] naredijo korak k

iskanju razložljivih skupin genskih markerjev z globokimi nevronskimi mrežami z vključevanjem genomskih in kliničnih podatkov.

Poleg računskih metod se bo predlagani projekt nanašal tudi na obstoječe objavljene prognostične biomarkerje. Na primer, Xiwen et al. [17] in Wang et al. [16] so preučevali korelacije med miRNA in prognozo bolnikov s hepatocelularnim karcinomom (HCC). Oba sta vzpostavila model s podpisi petih miRNA, ki bi lahko služil kot potencialni biomarker pri prognozi bolnikov s HCC. Podobno so Guodong et al. [34] identificirali nov model podpisa s petimi miRNA kot prognostični biomarker pri bolnikih z rakom debelega črevesa in danke. Poleg tega so Martinez-Ledesma et al. [19] raziskali omrežni pristop in identificirali biomarker na osnovi genske ekspresije, ki lahko uspešno napove klinični izid 12 različnih vrst raka. Navedene študije so odličen primer pomena projektov, kot je TCGA. Izčrpni in strukturirani podatki iz baze podatkov TCGA lahko drastično pospešijo razvoj tehnik za odkrivanje (Di et al., [13]) in potrjevanje (Chen et al. [5]) obetavnih genskih biomarkerjev.

27.3 Podroben opis delovnega programa

27.3.1 Projektne naloge

Projekt bo organiziran v okviru naslednjih sklopov nalog:

T0 Vzpostavitev sodelovalnega okolja. Vso kodo in dokumentacijo bomo odložili na GitHub[5]. Repozitorij bo hranil projektno dokumentacijo in zapisnike sestankov, sledenje spremembam in težavam, knjižnico kode v jeziku Python, teste računalniške kode in primere uporabe. Podatkovne datoteke bodo shranjene na ločen spletni strežnik. Razširitve Orange[6] bodo razvite kot dodatni paket za namestitev in bodo shranjene v ločenem repozitoriju na GitHubu.

T1 Pridobivanje in organizacija podatkov. Projekt bo uporabil številne različne nabore podatkov, ki prihajajo iz objavljenih študij in zbirk podatkov kot sta NCBI-jev Gene Expression Omnibus[7] in TCGA, The Cancer Genome Atlas database[8]. Poleg tega bomo ustvarili tudi nabor sintetičnih podatkovnih nizov različnih velikosti in zahtevnosti. Zbrani nabori podatkov bodo shranjeni v lastnem repozitoriju, ustvarjenem v nalogi T1.

T2 Razvoj orodij podatkovnega rudarjenja in bioinformatike za odkrivanje prognostičnih biomarkerjev. Še posebej bomo razvijali in izvajali tehnike za:

- **T2.1 Razvrstitev in izbiro genov na podlagi funkcije preživetja**, kjer bomo uvedli vse klasične metode s področja, vključno z log rank testom in Coxovim modelom sorazmernega tveganja. Vključili bomo tudi novejše in naprednejše pristope modeliranja, ki temeljijo na naključnih gozdovih (ang. *random forest*) in globokem učenju, ter z njimi rangirali gene in preučevali ubčutljivost različnih pristopov modeliranja.
- **T2.2 Priprava novih spremenljivk**, kjer bomo za agregiranje genske ekspresije uporabili napovedne modele na manjši podskupini genov in s tem povečali robustnost biomarkerja. Uporabili bomo $\ell 1$ regularizacijo v kombinaciji s Coxovim in izpeljanimi modeli, ter pristope z omrežji, pri katerih je biomarker sestavljen iz majhnega števila genov istega regulatornega omrežja ali iste metabolne poti.

[5] `https://github.com/biolab`
[6] `https://orangedatamining.com`
[7] `https://www.ncbi.nlm.nih.gov/geo`
[8] `https://portal.gdc.cancer.gov`

T2.3 Analiza genskih interakcij, kjer pričakujemo, da lahko skupina genov deluje nelinearno, da oblikuje močnejši in informativnejši biomarker. Prilagodili bomo pristope k iskanju lastnosti interakcije in pristope za vizualizacijo teh interakcij.

T2.4 Omejitve iskanja s podatkovnimi bazami, kjer bomo omejili iskalni prostor za genske interakcije na skupine genov s skupno funkcijo ali z drugimi skupnimi lastnostmi, kot so opisane v javnih podatkovnih zbirkah, glede na uveljavljene sisteme merjenja ekspresije genov (npr. nanoString paneli) ali glede na markerjih, znanih iz objavljene znanstvene literature.

T2.5 Globoko in preneseno učenje (ang. *deep learning and transfer learning*), kjer je naš cilj poiskati genske vložitve (ang. *embedding*) za njihovo profiliranje v prostoru z manj dimenzijami. Avtokoderje (specifične za tkivo in/ali bolezen) bomo trenirali s pomožnimi podatkovnimi seti [9] in jih s prenesenim učenjem [11] implementirali na problemu napovedovanja preživetja. Uporabili bomo profile genov v latentnem prostoru za vizualizacijo genskih zemljevidov in hevristično omejitev prostora za iskanje.

T2.6 Zemljevidi interakcije genov, kjer bi radi predstavili gene – potencialne biomarkerje – v genskem zemljevidu, kjer bližina genov na zemljevidu kaže na povečan skupni učinek na funkcijo preživetja. Ta orodnja bomo gradili na osnovi podobnih metod v analizi podobnega izražanja genov (ang. *co-expression*). Interakcijski zemljevidi bodo zelo vizualno predstavili prostor rešitev pri iskanju določenega biomarkerja.

T2.7 Samodejno označevanje točkovnih vizualizacij, kjer so točke geni, na primer v zemljevidih genskih interakcij. Oblikovali bomo algoritme, ki najdejo soseske z obogateno gensko funkcijo ali presnovnimi potmi, in s tem znanjem avtomatsko anotirali vizualizacije. Te vizualizacije bodo naslednja razvojna stopnja našega predhodnega dela na področju analize izražanja genov v posameznih celicah (glej sliko 4).

T3 Oblikovanje grafičnih vmesnikov za raziskovalno analizo podatkov o preživetju in odkrivanje biomarkerjev. V tesnem sodelovanju z Genialisom bomo izvedli temeljito analizo potreb in snovanje končnega produkta, da bi ta zagotovo zasnovan za interaktivno odkrivanje biomarkerjev z integracijo javnih podatkovnih zbirk. Zasnova bo vključevala načrtovanje nabora računskih komponent za obravnavo vseh vidikov analize preživetja in odkrivanjem biomarkerjev. Oblikovali bomo grafični vmesnik, interaktivne vizualizacije in možne cevovode za analizo podatkov. Izvedba bo vsebovala skice grafičnega uporabniškega vmesnika (v Balsamiq Mockups). Poudarek pri zasnovai bo kakovost uporabniške izkušnje, dostop do naprednih računskih tehnik, in zmožnost kombiniranja komponent na način Lego kock pri tvorbi potencialno zelo zapletenih in zmogljivih analitskih postopkov.

T4 Izvedba in integracija. Razvite računske tehnike bodo implementirane v okviru odprtokodnega okolja za rudarjenje podatkov Orange[9]. Pri izvedbi bo uporabljena knjižnica metod iz naloge T2 in grafičnih uporabniških načrtov iz T3. Izdelek bo objavljen kot ločen dodatek Orangu in bo sledile uveljavljenim izvedbenim smernicam glede dokumentacije in testiranja vse kode.

T5 Eksperimentalna validacija produkta. Razvita funkcionalnost bo temeljito preizkušena v sodelovanju z Genialisom, našim projektnim partnerjem. Uporabljeni bodo sintetični in resnični podatki (pripravljeni v okviru naloge 1) ter rezultati primerjani s tistimi iz literature. Validacija bo potrdila veljavnost in pravilnost razvitih postopkov in bo služila za zbiranje študij primerov, ki bodo objavljene na spletnem mestu GitHub, spletni strani Orange ter v načrtovanih publikacijah in patentih.

[9] `http://orangedatamining.com`

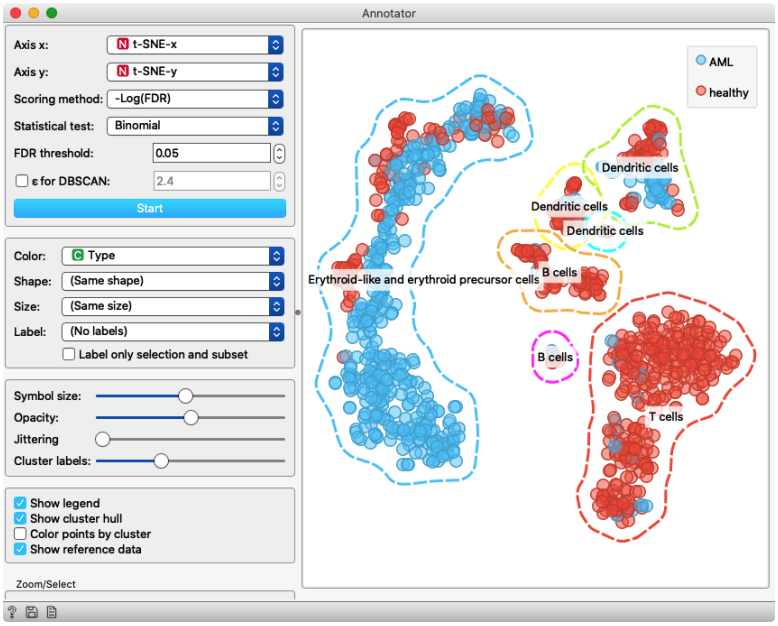

Fig. 4. Prototip modula v okolju Orange s samodejno anotacijo točkaste vizualizacije. Na vhodu ta komponenta dobi profil genske ekspresije posameznih celic in njihove vdelave (npr. koordinate t-SNE) ter seznam markerskih genov za vsak celični tip. V okviru predlaganega projekta bomo uporabil podoben pristop za oblikovanje modula, ki bo pojasnil prostor potencialnih biomarkerjev.

T6 Razširjanje rezultatov. Razvite metode bomo objavili pod splošno licenco (GPL). Napisali in spletno objavili bomo ustrezno dokumentacijo z delovnimi primeri, objavili videe s primeri uporabe na Orange YouTube kanalu[10] in širili spoznanja s predstavitvami na ustreznih konferencah in v znanstvenih revijah. Ciljali bomo na bioinformatske revije, kot so Bioinformatics, Nature Methods, Artificial Intelligence in Medicine ter konference, vključno z najboljše ocenjenima AIM in ISMB. Načrtujemo tudi prijavo skupnega patenta z Genialisom na področju odkrivanja prognostičnih biomarkerjev.

27.3.2 Research Design and Methods

Pregled. Pregled dela v predlaganem projektu je predstavljen na sliki 5. Slika prikazuje, kako bomo povezali podatke o genskega izražanja in klinične metapodatke z dodatnimi podatkovnimi bazami. Ugotovili bomo

1. kateri geni so sami po sebi ključni molekularni markerji preživetja,
2. katere kombinacije genov (genski sklopi) so povezane s preživetjem,
3. katere značilnosti markerskih genov nam lahko pomagajo pri interpretaciji rezultatov.

Čeprav bi zgoraj našteta vprašanja lahko individualno in ročno naslovil molekularni biolog, pa jih je v praktičnem smislu mogoče rešiti le s pomočjo računske analize in vključevanjem velikega števila podatkov iz različnih javnih zbirk. Glavni izziv projekta je torej, kako te različne vire združiti in uporabiti sodobne pristope rudarjenja podatkov za predlaganje novih hipotez. V spodnjem opisu najprej navedemo vire podatkov, na katerih bomo uporabili postopek odkrivanja biomarkerjev, nato pa nabor računskih pristopov, ki jih bomo razvili in uporabili, ter komentiramo izvedljivost njihove implementacije znotraj obstoječe platforme za vizualno programiranje Orange.

Podatki. V projektu bodo zbrane, organizirane in uporabljeni štirje viri podatkov o genskih ekspresijah in preživetju:

DS1 Podatki iz baze podatkov The Cancer Genome Atlas[11] [31] in Gene Expression Omnibus[12] [2].

[10] `http://youtube.com/orangedatamining`
[11] `https://www.cancer.gov/tcga`
[12] `https://www.ncbi.nlm.nih.gov/geo/`

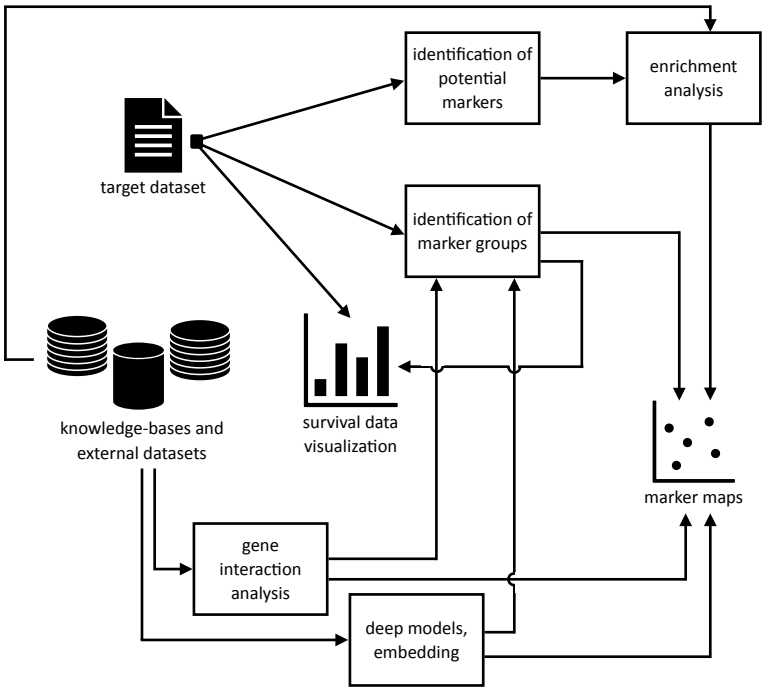

Fig. 5. Od podatkov do odkritja biomarkerja. Naš cilj je združiti podatkovni sete s podatki genskem izražanju in kliničnem preživetju z drugimi razpoložljivimi nabori podatkov. Tako bomo pridobili na hitrosti, natančnosti in razlagi rezultatov. Znaten del našega projekta se ukvarja z razvojem grafičnih vmesnikov, ki združujejo vizualizacijo podatkov, biomarkerjev in modelov. Računski in vizualizacijski pristopi bodo postavili rezultate v kontekst celotnega iskalnega prostora (npr. interakcijski zemljevidi) in nam bodo v pomoč pri polavtomatskem procesu odkrivanja novih biomarkerjev. Metode bodo tako podale hipoteze, ki jih bo ocenil domenski strokovnjak, v pomoč pa mu bodo agregirani podatki iz različnih podatkovnih zbirk.

DS2 Podatki o analizi preživetja iz kliničnih raziskav iz baze podatkov The Cancer Genome Atlas database.

DS3 Zasebni podatki o transkripciji in kliničnih preskušanjih, ki jih upravlja Genialis.

DS4 Simulirani nabori podatkov.

Podatki iz DS1 in DS2 so razdrobljeni; nabore podatkov je treba integrirati, tako da so podatki iz kliničnih raziskav poravnani z ustreznimi meta-podatki. Metodološka poročila o razvoju računskih tehnik, vključno s tistimi, ki smo jih navedli v pregledu literature, redko objavijo organizirane podatke, pripravljene za uporabo v prodajnih programskih paketih. Namen projekta je prekiniti s to prakso in sestaviti repozitorij podatkov z usklajenim kliničnim in transkripcijskim podatkovnim nizom, pripravljenem za primerjalno analizo in tehnike odkrivanja biomarkerjev.

Podjetje Genialis d.o.o. že ima zbirko tovrstnih podatkov, ki izhajajo iz njihovih obstoječih partnerstev z nekaterimi večjimi farmacevtskimi podjetji. Podatki so zasebni, vendar jih bodo v okviru projekta delili z nami za namene razvoja in testiranja.

Izdelali bomo tudi nabor simuliranih podatkovnih nizov. Podatki bodo služili za testiranje in primerjalno analizo predlaganih metod.

Projekt bo dodatno uporabil tudi druge vire informacij. Ti vključujejo, vendar niso omejeni na:

DS6 Genske funkcije iz zbirke genskih ontologij konzorcija Gene Ontology (GO).[13]

DS7 Presnovne poti iz KEGG, Kjotske enciklopedije genomov in genov.[14]

DS8 Podatkovna baza poti NDEx.[15]

DS9 Različne podatkovne baze markerskih genov, vključno s CellMarker[16] [37] in PanglaoDB [17] [10].

[13] `https://www.geneontology.org`
[14] `https://www.genome.jp/kegg`
[15] `http://www.ndexbio.org/`
[16] `http://biocc.hrbmu.edu.cn/CellMarker`
[17] `https://panglaodb.se`

Računski pristopi, pridobivanje podatkov in bioinformatika Računski pristopi in razvoj načinov podatkovnega rudarjenja bo vključeval:

Organizacija podatkov (naloga T1). Projekt bo razvil računalniško platformo z dostopom do skupnih podatkovnih baz s strežniške zbirke, ki bodo hranile podatke o genskem izražanju in preživetju. Za to arhitekturo bomo uporabili običajne prakse programskega inženirstva (podatkovni strežnik z varnim dostopom HTTP, poizvedbe, ki temeljijo na HTTP, komponente na odjemalcih podatkov). Lokalno bomo shranili druge vire informacij, kot so genske ontologije in poti, da bodo te informacije hitro na voljo. Za njihovo ponovno uporabo bomo uporabili obstoječo arhitekturo Orange [8, 7, 11]. Na splošno je cilj skriti kompleksnost dostopanja do podatkov pred uporabnikom. Uporabniki bi morali imeti možnost dostopanja do teh funkcij z enim samim klikom in se osredotočiti na analizo in interpretacijo podatkov.

Razvrstitev genov (T2.1). Za primerjavo dveh ali več krivulj preživetja bomo uporabili standardni log-rank test in tako ocenili informacijsko vrednost izbranih genovin prag njihovih ekspresije. Napredni modeli bodo vključevali naključne gozdove [28] in globoko učenje [14, 6]. Posredne vrednosti prispevkov posameznih genov bomo določevali z metodami teorije iger, npr. SHAP[18] [18].

Analiza obogatitve genskega sklopa (del T2.1). Za interpretacijo rezultatov razvrstitve biomarkerjev bomo uporabili analizo obogatitve genskega nabora [36] in pregledali skupne točke najbolje uvrščenih genov glede na funkcije in poti.

Konstrukcija novih spremenljivk in identifikacija naborov genov (naloga T2.2). Uporabili bomo neposreden in posreden izbor. Kakovost nabora gena bo ocenjena s pomočjo kvalitete napovedi preživetja. Kot alternativni pristop bomo uporabili izbiro lastnosti z modelom, kot je $\ell 1$ regularizacija Coxovih modelov in modelov omrežja.

Analiza interakcije genov (T2.3) bo preučila možne kombinacije genov na njihov skupni učinek na funkcijo preživetja. Nameravamo uporabiti modelno vrednotenje genskih parov (T2.1) in predstaviti rezultate v interakcijskih zemljevidih (T2.6) in omrežjih.

Odkritje biomarkerja s pomočjo drugih podatkov (T2.4) bo omejilo iskanje uporabnih kombinacij v nalogi T2.2 na gene s skupno biološko funkcijo, po možnosti na tiste, ki so že povezani s preučevano patologijo v literaturi. Uspeh tega pristopa se bo meril s povečanjem hitrosti algoritma in zmanjšanjem prostora za iskanje, ki ga je treba upoštevati, s tem da bomo zahtevali podobno kakovost rezultatov kot tistih iz izčrpnega iskanja.

Globoko učenje (T2.5) bo za vdelavo uporabljalo avtoenkoderje, ki bodo predstavljali kandidate za biomarkerje z vektorji v latentnem prostoru, kjer lahko enostavno pregledamo njihovo povezanost in strukturo prostora biomarkerjev. Iz zbirke transkripcijskih podatkov in naborov kliničnih podatkov (naloga T1) bomo izdelali avtokoderje in nato uporabili prenosno učenje [11] za prilagoditev modela določenemu ciljnemu naboru podatkov. Take predstavitve potencialnih markerjev bodo osnova za oblikovanje interpretativnih vizualizacij (npr. naloga T2.6).

Zemljevidi o interakcijah genov (T2.6) bodo podajali informacije iz nalog T2.3 in T2.7 z namenom vizualne interpretacije rezultatov. Za upodabljanje genskih zemljevidov bomo uporabili lastno različico t-SNE [29], imenovano openTSNE[19] [23], ki lahko ohrani globalno strukturo.

Razlaga točkovnih vizualizacij bo opremila genske zemljevide (naloga T2.6) in druge vizualizacije s funkcijskimi oznakami in oznakami presnovnih poti. Ta pristop bomo uporabili za pomoč pri interpre-

[18] `https://github.com/slundberg/shap`
[19] `https://github.com/pavlin-policar/openTSNE`

taciji iskalnega prostora. Da bi razvili to tehniko, bomo razširili naš pristop, ki smo ga prej razvili za analizo izražanja genov v posameznih celicah (glej sliko 4).

Uporaba programske opreme. V zadnjih dveh desetletjih smo razvijali obsežno analizo podatkov z imenom Orange[20] [8, 7, 11]. Orange uporablja tisoče uporabnikov po vsem svetu in je izbrano orodje pri izobraževanju podatkovnih znanosti na stotinah univerz[21]. Orange ima skriptni del in vizualno programsko okolje. Vizualno programiranje ponuja intuitivno sredstvo za kombiniranje orodij analiz in vizualizacij v zmogljive aplikacije. Orange vsebuje nabor vizualnih komponent za funkcijsko genomiko (slika 3), ki uporabnikom, ki niso programerji, omogoča analizo transkripcijskih podatkov in prilagoditev njihove analize s kombiniranjem običajnih orodij za analizo podatkov, ki ustrezajo njihovim potrebam [11]. Orodje Orange bomo uporabili za razvoj metod, ki jih predlagamo v projektu, in jih bomo delili z znanstveno skupnostjo znotraj odprtokodnega modela.

V okviru projekta bomo razvili nabor komponent, ki bodo specifične za naš problem, hkrati pa dovolj splošne za druge naloge iz analize preživetja. Razviti želimo komponente za modeliranje preživetja, oceno natančnosti, razvrstitev biomarkerjev, identifikacijo skupin markerjev, vgradnjo (ang. *embedding*) genov in izdelavo zemljevidov genskih interakcij na osnovi preživetja. Orange že vsebuje nekatere gradnike, ki so ključnega pomena za predlagani projekt. Te vključujejo dostop do knjižnic genskih pripisov, obogatitve genskega nabora, brskalnika GO in orodij za vizualizacijo, vključno z vizualizacijama t-SNE in krivulj Kaplan-Meier.

Eksperimentalna potrditev. Podatke DS1 do DS4 bomo uporabili za eksperimentalno potrditev računskih tehnik in vizualnih vmesnikov. Ocenjevali bomo tri vidike:

- **napovedna uspešnost**, kjer bomo primerjali napovedi biomarkerjev z znanimi biomarkerji in primerjali rezultate naših tehnik odkrivanja markerjev z objavljenimi;
- **razumljivost**, ki jo bodo ocenili domenski eksperti, ki sodelujejo z Genialisom. Ocena razumljivosti bo kvantitativna glede na to, kako dobro lahko identificirane biomarkerje povezujemo z znanimi funkcionalnimi genskimi potmi, in subjektivna v smislu potrditve napovedanih genskih poti s strani domenskih ekspertov;
- **uporabnost**, kjer bomo z domenskimi eksperti na praktičnih delavnicah ocenili, ali lahko po triurnem treningu orodje uporabljajo samostojno.

27.4 Razpoložljiva raziskovalna oprema nad 5.000 €

Genialis bo prispeval spletno orodje Genialis Expressions (glej sliki 6 in 7) za agregacijo, pregledovanje in analizo podatkov genskih zaporedij ter kliničnih in eksperimentalnih podatkov. Orodje Genialis Expressions je nameščeno na spletno platformo Amazon AWS Kubernetes in je v lasti podjetja Genialis.

Projekt bo uporabljal računalniško infrastrukturo Laboratorija za bioinformatiko Univerze v Ljubljani, ki vključuje računalniško gručo z računskimi procesorji CPU (približno 500 procesorjev), pomnilnik NFS (približno 500 TB) in računalniško gručo z grafičnimi procesorji GPU (približno 20 grafičnih procesorjev) v skupni vrednosti približno 200.000 EUR. Za predlagani projekt ne bomo potrebovali posebne računske opreme poleg obstoječe.

27.5 Vodenje projekta

Projekt bo združil dve komplementarni skupini raziskovalcev iz univerze in gospodarstva. Laboratorij za bioinformatiko Univerze v Ljubljani prinaša široka strokovna znanja na področjih rudarjenja podatkov, strojnega

[20] `https://orangedatamining.com`
[21] `https://orangedatamining.com/blog/2021/2021-01-11-orange-in-classroom/`

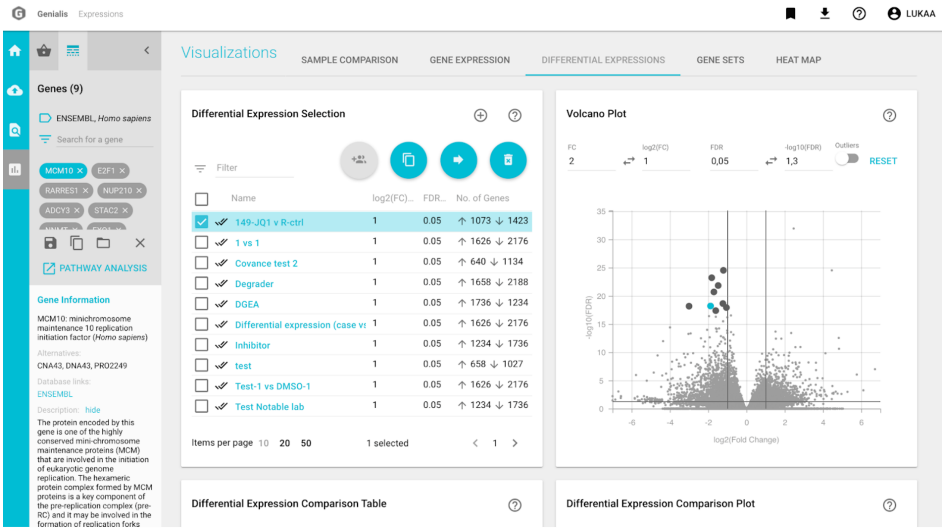

Fig. 6. Genialis Expressions je spletno stičišče biomedicinskih podatkov, ki zagotavlja principe dostopnosti, interoperabilnosti in ponovne uporabe podatkov (angl. FAIR principles – Findable, Accessible, Interoperable and Re-usable). Podpira avtomatsko primarno analizo podatkov geniskih sekvenc, hrani informacijo o izvoru podatkov in omogoča ponovljivost računskih analiz. Genialis Expressions omogoča sodelovanje uporabnikov, ki lahko podatke izmenjujejo, in podpira integracijo s poljubnimi programi kot je npr. Orange.

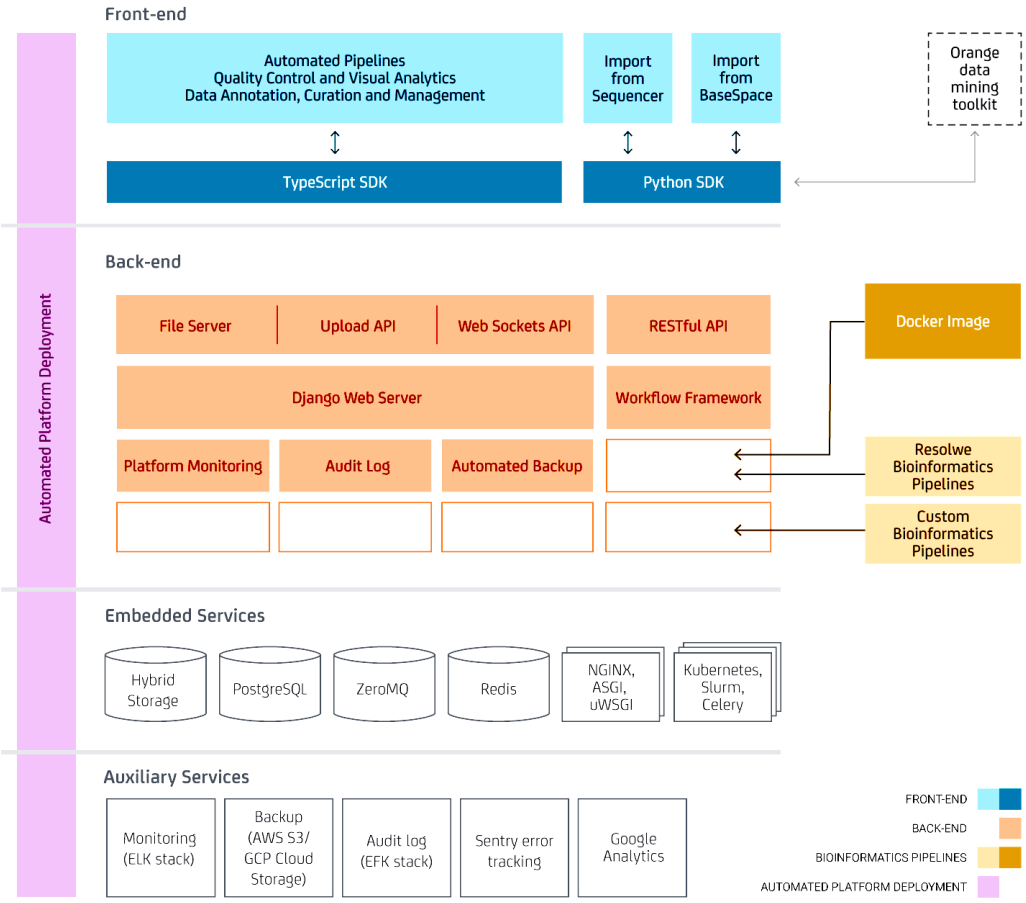

Fig. 7. Genialis Expressions ima modularno arhitekturo programske opreme. Prednja plast (angl. frontend) komunicira z zaledjem (angl. backend) preko vmesnika RESTful API. Na voljo so knjižnice za razvoj povezane programske opreme v programskih jezikih TypeScript in Python, na primer za integracijo s programom Orange kot je prikazano s sivo barvo v zgornjem desnem kotu. Tudi komponente za podatkovno analizo imajo modularno zasnovo (glej rjava polja na sliki desno). Analize tečejo vzporedno v gruči računalnikov Kubernetes, vsaka v svojem izoliranem okolju v vsebnikih Docker.

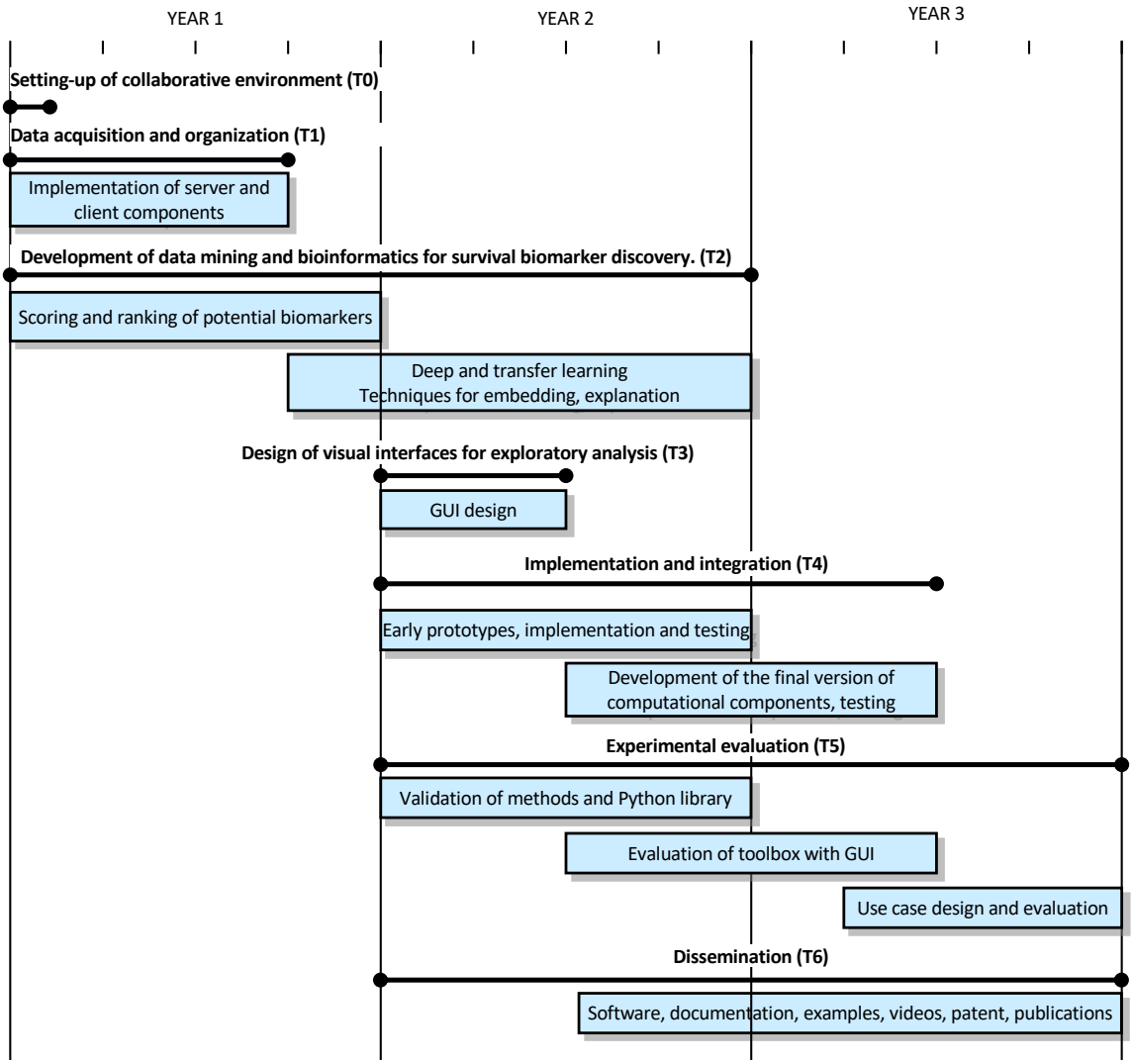

Fig. 8. Časovnica projekta. Začeli bomo z urejanjem podatkov. Nato bomo razvili Python knjižnico za odkrivanje genskih markerjev pri analizi preživetja. Nadaljevali bomo z načrtovanjem in izvedbo grafičnega uporabniškega vmesnika za nove komponente knjižnice podatkovnega rudarjenja Orange. V sodelovanju z Genialisom bomo testirali komponente na realnih problemih. Posebno pozornost bomo namenili diseminaciji rezultatov, vključno z objavo odprtokodne programske opreme, dokumentacije, primerov uporabe, video materialov, znanstvenih objav in vložitvijo patenta.

učenja, bioinformatike in računskega fenotipiziranja. Projekt bo sofinanciralo podjetje Genialis, ki raziskuje in trži nove pristope zdravljenja bolezni, na področjih klinične in translacijske medicine. Genialis sodeluje s farmacevtskimi podjetji pri identifikaciji tarč, iskanju in razvoju biomarkerjev in pozicioniranju novih zdravil. Skupaj so razvili nove pristope za odkrivanje zdravil in zdravljenj, ki bistveno izboljšajo kakovost življenja ljudi z zapletenimi boleznimi.

Projekt bo vodila UL (vodja projekta). Vodenje projekta bo organizirano na rednih sejah upravnega odbora s po enim imenovanim predstavnikom iz vsake institucije in z rednimi sestanki članov projekta. Platforma za sodelovanje bo temeljila na GitHub in bo na razpolago že v zgodnejši fazi projekta.

Projekt bomo zaključili v treh letih. Slika 8 prikazuje podrobno časovnico projekta.

Reference

[1] D. G. Altman. Categorising continuous variables. *British Journal of Cancer*, 64(5):975–975, Nov 1991.

[2] T. Barrett, S. E. Wilhite, P. Ledoux, et al. NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Research*, 41(Database issue):D991–995, Jan 2013.

[3] J. Budczies, F. Klauschen, B. V. Sinn, et al. Cutoff Finder: a comprehensive and straightforward Web application enabling rapid biomarker cutoff optimization. *PLOS One*, 7(12):e51862, 2012.

[4] H. B. Burke. Predicting clinical outcomes using molecular biomarkers. *Biomarkers in Cancer*, 8:BIC–S33380, 2016.

[5] J. Chen, Z. Wang, W. Wang, et al. SYT16 is a prognostic biomarker and correlated with immune infiltrates in glioma: A study based on TCGA data. *International Immunopharmacology*, 84:106490, Jul 2020.

[6] T. Ching, X. Zhu, and L. X. Garmire. Cox-nnet: an artificial neural network method for prognosis prediction of high-throughput omics data. *PLOS Computational Biology*, 14(4):e1006076, 2018.

[7] T. Curk, J. Demsar, Q. Xu, et al. Microarray data mining with visual programming. *Bioinformatics*, 21(3):396–398, Feb 2005.

[8] J. Demšar and B. Zupan. Orange: Data mining fruitful and fun - a historical perspective. *Informatica*, 37:55–60, 2013.

[9] C. Doersch. Tutorial on variational autoencoders, 2021.

[10] O. Franzén, L. M. Gan, and J. L. M. Björkegren. PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data. *Database (Oxford)*, 2019, 01 2019.

[11] P. Godec, M. Pančur, N. Ilenič, et al. Democratized image analytics by visual programming through integration of deep models and small-scale machine learning. *Nature Communications*, 10(1):4551, 10 2019.

[12] J. Hao, Y. Kim, T. Mallavarapu, et al. Interpretable deep neural network for cancer survival analysis by integrating genomic and clinical data. *BMC Medical Genomics*, 12(10):1–13, 2019.

[13] D. Jia, S. Li, D. Li, et al. Mining TCGA database for genes of prognostic value in glioblastoma microenvironment. *Aging (Albany NY)*, 10(4):592–605, 04 2018.

[14] J. L. Katzman, U. Shaham, A. Cloninger, et al. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Medical Research Methodology*, 18(1):24, 2018.

[15] Y.-W. Kim, D. Koul, S. H. Kim, et al. Identification of prognostic gene signatures of glioblastoma: a study based on TCGA data analysis. *Neuro-Oncology*, 15(7):829–839, 2013.

[16] W. Li, X. Kong, T. Huang, et al. Bioinformatic analysis and in vitro validation of a five-microRNA signature as a prognostic biomarker of hepatocellular carcinoma. *Ann Transl Med*, 8(21):1422, Nov 2020.

[17] X. Liao, G. Zhu, R. Huang, et al. Identification of potential prognostic microRNA biomarkers for predicting survival in patients with hepatocellular carcinoma. *Cancer Management and Research*, 10:787, 2018.

[18] S. M. Lundberg, G. Erion, H. Chen, et al. From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, 2(1):56–67, Jan 2020.

[19] E. Martinez-Ledesma, R. G. Verhaak, and V. Treviño. Identification of a multi-cancer gene expression biomarker for cancer clinical outcomes using a network-based algorithm. *Scientific Reports*, 5:11966, Jul 2015.

[20] M. Mazumdar and J. R. Glassman. Categorizing a prognostic variable: review of methods, code for easy implementation and applications to decision-making about cancer treatments. *Statistics in Medicine*, 19(1):113–132, 2000.

[21] J. Monforte and S. McPhail. Strategy for gene expression-based biomarker discovery. *Biotechniques*, Suppl:25–29, Apr 2005.

[22] R. Pazdur. Endpoints for assessing drug activity in clinical trials. *Oncologist*, 13(2):19, 2008.

[23] P. G. Poličar, M. Stražar, and B. Zupan. openTSNE: a modular python library for t-SNE dimensionality reduction and embedding. *bioRxiv*, 2019.

[24] Y. F. Ramos, S. D. Bos, N. Lakenberg, et al. Genes expressed in blood link osteoarthritis with apoptotic pathways. *Annals of the rheumatic diseases*, 73(10):1844–1853, 2014.

[25] R. T. Relator, A. Terada, and J. Sese. Identifying statistically significant combinatorial markers for survival analysis. *BMC Medical Genomics*, 11(2):45–55, 2018.

[26] R. Singh and K. Mukhopadhyay. Survival analysis in clinical trials: Basics and must know areas. *Perspectives in Clinical Research*, 2(4):145, 2011.

[27] A. R. Sonawane, S. T. Weiss, K. Glass, and A. Sharma. Network medicine in the age of biomedical big data. *Frontiers in Genetics*, 10:294, 2019.

[28] J. M. Taylor. Random Survival Forests. *Journal of Thoracic Oncology*, 6(12):1974–1975, Dec 2011.

[29] L. van der Maaten and G. Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

[30] Z. Wang, G. Chen, Q. Wang, et al. Identification and validation of a prognostic 9-genes expression signature for gastric cancer. *Oncotarget*, 8(43):73826, 2017.

[31] J. N. Weinstein, E. A. Collisson, G. B. Mills, et al. The Cancer Genome Atlas Pan-Cancer analysis project. *Nature Genetics*, 45(10):1113–1120, Oct 2013.

[32] D. M. Witten and R. Tibshirani. Survival analysis with high-dimensional covariates. *Statistical Methods in Medical Research*, 19(1):29–51, 2010.

[33] S. Y. Woo and S. Kim. Determination of cutoff values for biomarkers in clinical studies. *Precision and Future Medicine*, 4(1):2–8, 2020.

[34] G. Yang, Y. Zhang, and J. Yang. A five-microRNA signature as prognostic biomarker in colorectal cancer by bioinformatics analysis. *Frontiers in Oncology*, 9:1207, 2019.

[35] S. Yousefi, F. Amrollahi, M. Amgad, et al. Predicting clinical outcomes from large scale cancer genomic profiles with deep survival models. *Scientific Reports*, 7(1):1–11, 2017.

[36] A. Zacho, J. Nielsen, and C. Cederqvist. Relationship between type of tobacco used and localization of tumour in patients with gastric cancer. *Acta Chirurgica Scandinavica*, 141(7):676–679, 1975.

[37] X. Zhang, Y. Lan, J. Xu, et al. CellMarker: a manually curated resource of cell markers in human and mouse. *Nucleic acids research*, 47(D1):D721–D728, 2019.

[38] Y.-Z. Zhang, L.-H. Zhang, Y. Gao, et al. Discovery and validation of prognostic markers in gastric cancer by genome-wide expression profiling. *World Journal of Gastroenterology: WJG*, 17(13):1710, 2011.

[39] M. Žitnik, E. A. Nam, C. Dinh, et al. Gene Prioritization by Compressive Data Fusion and Chaining. *PLOS Computational Biology*, 11(10):e1004552, Oct 2015.

[40] M. Žitnik, F. Nguyen, B. Wang, et al. Machine learning for integrating data in biology and medicine: Principles, practice, and opportunities. *Information Fusion*, 50:71–91, Oct 2019.