Univerza v Ljubljani
Fakulteta za računalništvo in informatiko

Aneja Furlan

Računski postopki za odkrivanje celičnih tipov v vizualizacijah podatkov scRNA

diplomsko delo

interdisciplinarni univerzitetni
študijski program prve stopnje
računalništvo in matematika

Mentor: prof. dr. Blaž Zupan

Ljubljana, 2018

Besedilo je oblikovano z urejevalnikom besedil LaTeX.

Fakulteta za računalništvo in informatiko izdaja naslednjo nalogo:

Tematika naloge:

Podobnosti izraznih profilov posameznih celic lahko predstavimo z njihovo umestitvijo v dvorazsežne razsevne diagrame. Predlagajte računsko tehniko, ki v teh diagramih poišče in označi skupine celic istega tipa. Vhod v metodo je matrika izraznih profilov celic in skupine genov, značilnih za izbran celični tip. Metodo ovrednotite na podatkih, ki jih pridobite iz literature.

Kazalo

Povzetek

V genomiki se v zadnjem času, zahvaljujoč napredku tehnologije, veliko raziskuje na področju transkriptomike posameznih celic, ki ponuja nov, drugačen vpogled v funkcionalno različnost celic. Pridobivanje takih podatkov se začne z izolacijo RNA iz posameznih celic, zato jih lahko krajše označimo kot podatke scRNA (ang. *single-cell RNA*). V diplomskem delu predlagamo metodo za odkrivanje celičnih tipov na vizualizacijah podatkov scRNA. Metoda prejme podatke o izražanju genov v celicah in seznam genov, značilen za izbran tip celice. Visokorazsežne izrazne profile celic vloži v dvorazsežen prostor, primeren za vizualizacijo v obliki razsevnega diagrama, in nato na podlagi ocen obogatenosti soseščin celic glede na podan seznam genov odkrije regije celic izbranega celičnega tipa. Uspešnost predlagane metode smo preverili na treh različnih naborih podatkov nedavno opravljenih študij sekvenciranja na nivoju posameznih celic. Metoda se je izkazala za robustno glede na začetno vložitev v 2D prostor in glede na izbrano velikost obravnavanih soseščin. Dobili smo spodbudne rezultate, vendar je tu še veliko prostora za dopolnitve in izboljšave. Predvsem bi bilo smiselno metodo preveriti še na večjem naboru podatkov.

Ključne besede: bioinformatika, enocelična genomika, vizualizacije podatkov, analiza obogatenosti, scRNA.

Abstract

The popularity of single-cell analysis has risen due to recent advancements in single-cell transcriptomics, especially in sequencing technologies. Single-cell analysis provides us with a new perspective of cellular data and helps us study cellular heterogeneity. Expression profiles of individual cells can be derived with single-cell RNA sequencing (scRNA data) and corresponding gene expressions. In the Thesis we propose an approach for cell type discovery in such data. Input to the proposed approach is a gene expression matrix, along with a set of marker genes, typical for a specific cell type. Our method starts with visualizing data in 2D and then finds regions of chosen cell type by measuring functional enrichments across local neighborhoods and estimating their significances. We applied the proposed approach on three datasets from recent single-cell sequencing studies. We got encouraging results and showed the approach to be robust to dimensionality reduction and neighborhood size. There is still room for improvement and in order to further illustrate full functionality of the approach, more testing on larger datasets would be required.

Keywords: bioinformatics, single-cell genomics, data visualizations, enrichment analysis, scRNA.

Poglavje 1

Uvod

Razsevni diagram je priljubljena oblika predstavitve dvorazsežnih podatkov v statistiki. Z njim prikažemo relacije med dvema atributoma, pri čemer je vsak izmed njiju predstavljen na svoji osi. Omogoča nam hitro ugotavljanje korelacije med atributi, poenostavlja odkrivanje skupin in osamljenih primerov [1]. Razsevni diagrami so priročni tudi za predstavitev podatkov enocelične genomike scRNA, v zadnjem času z razlogom zelo popularnih podatkov v genomiki, s katerimi se ukvarjamo tudi v pričujočem delu. Z umestitvijo celic v dvorazsežne razsevne diagrame lahko predstavimo podobnosti njihovih izraznih profilov. Velikokrat nas zanima, kje se na vizualizaciji nahajajo podobne celice, denimo celice nekega izbranega celičnega tipa, in ali v naših podatkih obstajajo kake zaključene regije celic, ki imajo podobno nalogo.

V pričujoči nalogi se posvetimo odkrivanju regij v razsevnem diagramu, kjer prevladujejo celice določenega tipa. Naša rešitev na vizualizaciji podatkov scRNA poišče in označi območja izbranega celičnega tipa glede na podano skupino genov, značilnih za ta tip celice. Metoda za vsako celico določi njeno soseščino in zanimiva območja razsevnega diagrama poišče na podlagi izračuna obogatenosti skupine genov v lokalnih soseščinah celic.

Pričnemo s pregledom sorodnih del, na osnovi katerih predlagamo metodo za iskanje z genskimi skupinami obogatehnih regij na vizualizacijah

podatkov scRNA. Metodo preizkusimo na treh različnih naborih podatkov iz nedavno objavljenih študij sekvenciranja na nivoju posameznih celic in dobimo vzpodbudne rezultate. Zaključimo s povzetkom rezultatov in diskusijo o pomanjkljivostih in morebitnih izboljšavah predlagane metode.

Poglavje 2

Pregled sorodnih del

Pred pričetkom razvoja metode smo pregledali dela s področja bioinformatike, funkcijske genomike in genomike posameznih celic, da bi lahko učinkovito začeli z reševanjem zastavljenega problema. Seznaniti se je bilo potrebno z novo vrsto podatkov s področja genomike, najti primerno metodo za vizualizacijo podatkov, preveriti, kako je z analizo izražanj genov v celičnih podatkih in poiskati podobne pristope reševanja zastavljenega problema.

2.1 Podatki s področja genomike

Genomika je znanstvena panoga, ki se ukvarja s preučevanjem genoma. Genom je dedna informacija celotnega organizma, ki je zapisana v DNA (ang. *deoxyribonucleic acid*). Velik mejnik v raziskovanju genoma predstavlja projekt Človeški genom (ang. *Human Genome Project*), ki je bil ustanovljen leta 1990 na pobudo ameriških Nacionalnih inštitutov za zdravje (ang. *National Institutes of Health*) in ameriškega ministrstva za eneregtiko (ang. *U. S. Department of Energy*). Namen projekta je bilo sekvenciranje celotnega človeškega genoma. Leta 2003 so v okviru omenjenega projekta prvič objavili osnutke raziskav odkritja tri miljard dolgega zaporedja nukleotidnih parov človeškega genoma [2, 3].

V večceličnih organizmih skoraj vsaka celica vsebuje enak genom in s tem

enake gene. Gen je del DNA, ki nosi zapis za funkcionalne produkte celic. Informacija o proteinu se prepiše v molekulo RNA (ang. *ribonucleic acid*), ki jo ribosom uporabi pri tvorbi proteina z danim zaporedjem aminokislin. Pravimo, da se gen izraža, procesu izražanja pa s tujko pravimo transkripcija. Vsi geni niso transkripcijsko aktivni v vseh celicah - različne celice kažejo različne trende v izražanju genov (ang. *gene expression*). Del dednega materiala, ki se prepiše v RNA molekule, imenujemo transkriptom. Veda, ki se ukvarja s preiskovanjem transkriptoma je transkriptomika [4].

Do nedavnega se je raziskovalo predvsem na področju transkriptomike populacij celic, bodisi z analizo mikromrež ali uporabo sekvenciranja naslednje generacije (ang. *next-generation sequencing*, NGS) [5]. Pri analizi populacije celic predstavlja izražanje nekega gena povprečje izražanj le-tega skozi celotno populacijo celic. Tak pristop v nekaterih primerih zadošča, npr. pri primerjanju enakega tkiva različnih vrst organizmov, ne pa vedno, npr. pri analizi celic v zgodnjem razvoju, ko je le malo celic in ima vsaka svojo funkcijo in nalogo ali pri analizi kompleksnih tkiv kot je možgansko, ki ga je težko eksperimentalno secirati [5, 6]. Ključnega pomena v nadaljnjem raziskovanju celičnih procesov in celične sestave v vseh vrstah tkiv je analiza izražanj genov na nivoju posameznih celic, ki ponuja vpogled v doslej nepoznano funkcionalno različnost med populacijami celic [7]. Ta je priljubljenost pridobila v zadnjih nekaj letih, zahvaljujoč napredku tehnologije in razvoju novih, cenejših metod za sekvenciranje posameznih celic [8].

2.2 Vizualizacijski postopki

V pričujočem delu se ukvarjamo z visokorazsežnimi podatki. Za namen vizualizacije je potrebno te podatke vložiti v predstavljiv prostor nižjih dimenzij, tj. v dvo- ali trorazsežen prostor. V našem primeru se osredtočimo na vizualizacije v dvorazsežnem (2D) prostoru.

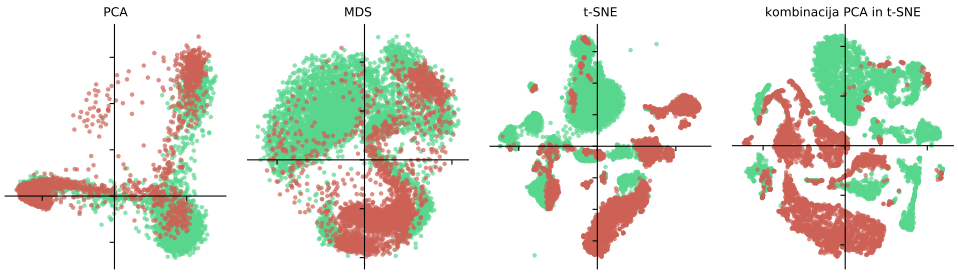
Klasični metodi za preslikavo podatkov v nižje prostore sta metoda glavnih komponent (ang. *principal component analysis*, PCA) in večrazsežnostno

lestvičenje (ang. *multidimensional scaling*, MDS). Metoda glavnih komponent, PCA, je linearna transformacija, pri kateri so nove spremenljivke, ki so linearne kombinacije originalnih spremenljivk, nekorelirane [9]. Da lahko kar najbolje predstavi razpršenost originalnih primerov, PCA te linearne kombinacije poišče tako, da le-te čimbolj korelirajo z originalnimi spremenljivkami. Nov prostor razpenjajo ortogonalni vektorji, ki so lastni vektorji kovariančne matrike originalnih spremenljivk. Linearnost PCA ima svoje omejitve, ker ne more predstaviti nelinearne relacije med podatki [10]. Večrazsežnostno lestvičenje, MDS, skuša pri projiciranju podatkov v nižjerazsežen prostor čimbolj ohraniti razdalje med posameznimi primeri v originalnem prostoru. Pri tem je izbira razdalje prepuščena uporabniku in se izbere glede na domeno vhodnih podatkov. Problem MDS je, da skuša ohranjati razdalje med primeri ne glede na to, kako blizu oz. daleč so si primeri v originalnem prostoru. Pri vizualiziranju podatkov nas pogosto zanima lokalna podobnost primerov; na vizualizaciji želimo imeti podobne podatke skupaj, medtem ko nam ohranjanje razdalje med podatki, ki so si različni, ni pomembno [11, 12].

Pri obeh opisanih metodah je en primer predstavljen izključno z eno lokacijo v novem nižjedimenzionalnem prostoru. Novejša metoda stohastičnega vstavljanja sosedov (ang. *stochastic neighbor embedding*, SNE) deluje na principu verjetnih sosedov in primere vstavlja v nov prostor tako, da kar najbolje ohrani njihovo lokalno soseščino. Metoda temelji na verjetnostni oceni podobnosti parov primerov glede na njihovo medsebojno lego v prostoru [13]. Danes je bolj uporabljena metoda t-SNE (*t-distributed stochastic neighbor embedding*), izboljšana različica metode SNE, ki se od slednje razlikuje v preprostejši cenilni funkciji in z uporabo Studentove (namesto Gaussove) porazdelitve rešuje njen problem prenatrpanosti (ang. *crowding*). Avtorji *t*-SNE predlagajo v primeru visokodimenzijskih podatkov predhodno zmanjšanje dimenzije le-teh s kako drugo metodo (npr. PCA), s tem je postopek hitrejši in delno je odstranjen šum podatkov brez prevelikih razlik v medsebojnih razdaljah primerov [14].

Slika 2.1 prikazuje razsevne diagrame projekcij podatkov o celicah kost-

nega mozga zdravega človeka in bolnika z akutno mieloblastno levkemijo (AML) [15], ki smo jih pridobili z opisanimi metodami za preslikanje podatkov v nizkorazsežne prostore PCA, MDS, t-SNE in kombinacijo PCA in t-SNE, ki jo kasneje uporabljamo v metodi za odkrivanje celičnih tipov, ki jo predlagamo.



Slika 2.1: Projekcije podatkov o celicah kostnega mozga bolnika z AML in zdravega človeka [15] v dvorazsežen prostor; z zeleno so označene celice zdravega človeka, z rdečo pa celice bolnika z AML.

2.3 Analiza obogatenosti

Rezultati visokozmogljivostnih metod na področju genoma so običajno podatki o izražanju več sto ali celo več tisoč genov. Čeprav je natančna biološka interpretacija takih podatkov še vedno izziv, je bilo v zadnjih nekaj desetletjih razvitih kar precej metod za sistematično analizo takšnih podatkov. Te metode skušajo za dano skupino genov poiskati najbolj obogatene (ang. *enriched*) genske pripise oziroma za skupino celic najbolj značilne oznake genov [16]. Z analizo obogatenosti lahko denimo ugotovimo, kateri celični proces izstopa v dani študiji, saj so celični procesi regulirani z uravnavanjem ekspresije več genov.

Genske skupine, ki so značilne za določeno študijo ali podatke, lahko pridobimo z opazovanjem diferencialne izraženosti genov v celicah. Izraženost genov lahko analiziramo na dva načina, z analizo posameznih genov (ang.

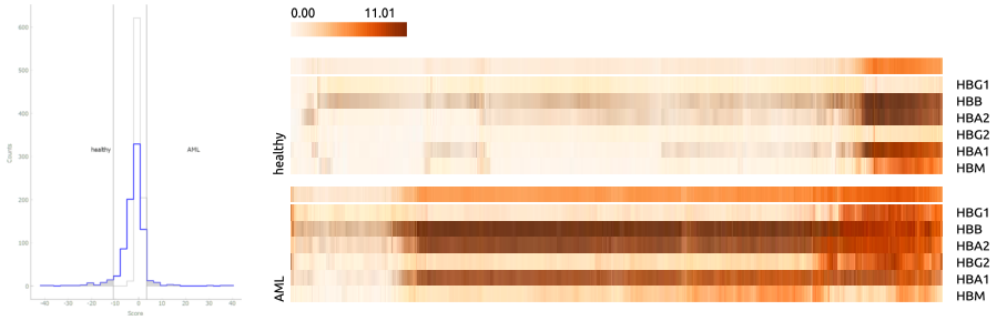
individual gene analysis), kjer ocenimo statistične pomembnosti posameznih genov in na podlagi le-teh določimo pomembnejše skupine genov, ali z analizo skupin genov (ang. *gene set analysis*), kjer neposredno ocenimo vnaprej določene skupine genov glede na korelacijo med izraženostjo genov v skupini in fenotipom [17]. Pri analizi posameznih genov dobimo kot rezultat množico genov, katerih statistična značilnost je nad določenim pragom. Slabosti teh metod sta močan vpliv izbranega praga, izbira različnih pragov vodi do različnih zaključkov o bioloških lastnostih obravnavanih podatkov [18], in napačna predpostavka o neodvisnosti genov (oz. genskih skupin), kar vodi do povečanja število lažno pozitivnih napovedi [17].

V zadnjem času so bolj uporabljane metode analize izraženosti skupin genov. Prednost le-teh je, da neposredno ocenjujejo skupine genov, dobro ocenijo tiste skupine, ki imajo zmerne, vendar usklajene spremembe v izraženosti [17]. Gene povežejo v skupine glede na znane anotacije, torej brez ozira na obravnavane podatke. Anotacije so običajno povzete po znanih knjižnicah kot sta genska ontologija (ang. *Gene Ontology*, GO) [19] in KEGG (ang. *Kyoto Encyclopedia of Genes and Genomes*) [20]. Metode analize obogatenosti skupin genov večinoma temeljijo na statističnih testih s testiranjem hipotez. Lahko jih razdelimo v dve skupini glede na ničelno hipotezo, ki jo testirajo: tekmovalne (ang. *competitive*) in samostojne (ang. *self-contained*) [21]. V splošnem tekmovalne metode testirajo relativno obogatenost diferencialno izraženih genov glede na ostale gene in iščejo skupine genov z usklajenimi spremembami v izraženosti. Samostojne metode za razliko od tekmovalnih metod uporabijo samo informacijo o izraženosti genov znotraj obravnavane skupine in lahko odkrijejo več obogatenih skupin genov kot slednje, saj lahko že en sam diferencialno izražen gen močno vpliva na obogatenost skupine [17, 21, 22].

Najbolj znana analiza obogatenosti skupin genov je metoda GSEA (*Gene Set Enrichement Analysis*) [23]. GSEA ne moremo razvrstiti v nobeno izmed prej opisanih skupin. Je analitična metoda, ki testira ničelno hipotezo, da nobena od vnaprej določenih skupin genov ni povezana s fenotipom.

Za razliko od tekmovalnih in samostojnih metod, ki testirata pomembnost posameznih skupin genov, GSEA obravnava celoten nabor podatkov in testira, če le-ta vsebuje katerokoli skupino genov, ki bi bila povezana s fenotipom. GSEA je tekmovalna glede na posamezne skupine genov in samostojna glede na celoten nabor podatkov [17].

Na sliki 2.2 je analiza obogatenosti v tisoč celicah kostnega mozga [15] prek analize posameznih genov s pomočjo programskega paketa Orange [24]. S pomočjo t-testa najprej ocenimo statistične pomembnosti posameznih genov glede na fenotip (zdrav človek, bolnik z AML) in izmed tisoč genov izberemo 72 statistično pomembnejših (slika 2.2 levo). S pomočjo iskalnika poizvedb v genski ontologiji (GO)[1] poiščemo zanimiv izraz GO (filtriranje po p-vrednosti in deležu napačnih napovedi (ang. *false discovery rate*, FDR)), denimo prenos kisika (ang. *oxygen transport*). Izraženosti genov, pripadajočih temu izrazu GO, lahko nazorno prikažemo s toplotnim grafom (ang. *heat map*, slika 2.2 desno). Opazimo, da je prenos kiska aktivenjši v celicah bolnika z AML, kar lahko razlagamo s povečanim številnom reaktivnih kisikov spojin in posledično oksidativnega stresa pri bolnikih z AML [25].



Slika 2.2: Analiza obogatenosti v celicah kostnega mozga [15]: distribucija t statistik izbranih 1000 celic glede na fenotip na levi, izrazni profili celic in povprečje leteh za skupino genov, značilnih za prenos kisika, za vsakega izmed dveh fenotipov, zdravega človeka in bolnika z AML, na desni.

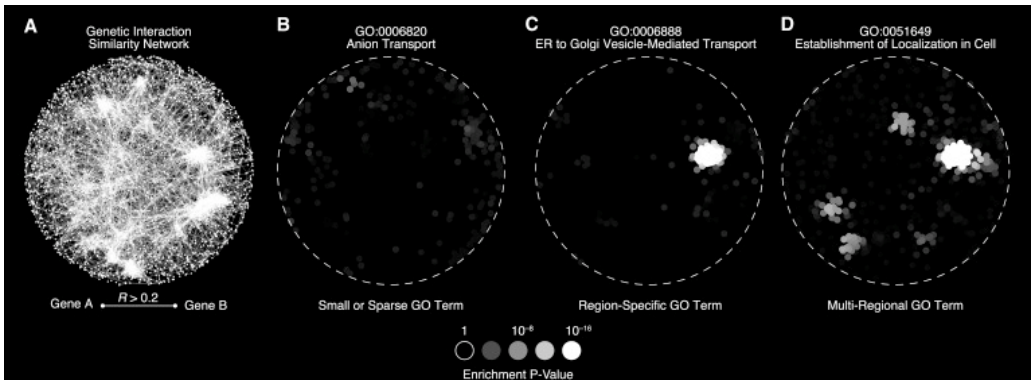
[1] GO browser: `http://orange-bioinformatics.readthedocs.io/en/latest/widgets/gobrowser.html` [dostopano 29. junij 2018]

2.4 Prostorska analiza funkcionalne obogatenosti SAFE

Razvoj naše metode se je v veliki meri zgledoval po novejši metodi prostorske analize obogatenosti SAFE (ang. *spatial analysis of functional enrichment*). SAFE je metoda za funkcijsko anotacijo bioloških omrežij, ki omrežje najprej predstavi v dvorazsežnem prostoru in nato na podlagi porazdelitve obogatenosti v lokalnih soseščinah vozlišč omrežja definira območja obogatenosti posameznih atributov [26].

Metoda SAFE biološko omrežje najprej predstavi z grafom v dveh dimenzijah, za kar uporabi algoritem za risanje grafov na osnovi ravnovesij sil (ang. *force-directed network layout algorithm*). Za vsako vozlišče grafa X definira njegovo lokalno soseščino glede na prag maksimalne razdalje. Privzeta razdalja med vozlišči je najkrajša pot v grafu (ne glede na originalen prostor). Glede na lokalne soseščine vsakemu vozlišču X dodeli oceno, enako seštevku funkcionalnih atributov njegovih sosedov. Atributi so lahko binarni ali numerični. Sledi izračun statistične pomembnosti, p-vrednosti, ki jo oceni s pomočjo hipergeometrijske porazdelitve (binarni atributi) ali empirične ocene (numerični atributi) glede na naključne permutacije grafa. Središču soseščine X dodeli logaritemsko oceno obogatenosti $-log_{10}p$. Izračunane ocene obogatenosti določajo območje atributove obogatenosti, ki ima svojo velikost, obliko in relief. Preko slednjih lahko razberemo porazdelitev izbranega funkcionalnega atributa v preučevanem biološkem omrežju [26].

Avtorica članka je metodo med drugim testirala na biološkem omrežju genetskih interakcij kavasovk *Saccharomyces cerevisiae*, na katerem je označevala področja več izrazov genske ontologije (GO). Vsak izraz GO je predstavljen s seznamom genov, označena področja pa predstavljajo območja, obogatena z geni tega seznama. Na sliki 2.3 lahko vidimo izbrano biološko omrežje in poskus označitve treh izrazov GO, vsak izmed njih ima drugačno regijo obogatenosti [26].



Slika 2.3: Rezultat metode SAFE na biološkem omrežju genetskih interakcij kavasovk: (A) Graf izbranega biološkega omrežja (B) Označitev skupine genov vnosa GO:0006820, ki nima značilnega območja (C) Označitev skupine genov vnosa GO:0006888, ki ima eno značilno območje (D) Označitev skupine genov vnosa GO:0051649, ki ima več značilnih območij. Sliko smo pridobili iz [26].

Poglavje 3

Metoda

Predlagamo metodo, ki na vizualizaciji podatkov scRNA označi območja izbranega tipa celice glede na podane markerske gene za ta tip celice. Metoda za vsako celico določi njeno soseščino in poišče zanimiva območja na podlagi izračuna obogatenosti skupine markerskih genov v vsaki soseščini. Sledi oris metode in podrobnejši opis implementacije posameznih korakov.

3.1 Oris metode

Glavna vhodna podatka metode sta genska ekspresijska matrika celic $C \subset \mathbb{R}^{N \times K}$, kjer je N število obravnavnih celic in K število genov, s katerimi je vsaka celica opisana, in seznam markerskih genov G celičnega tipa, ki ga želimo na vizualizaciji označiti.

Ker skuša metoda odkriti celične tipe na vizualizaciji v dvorazsežnem prostoru, najprej podane podatke vloži v tak prostor. Nato vsako celico oceni s povprečno vrednostjo izražanj markerskih genov in določi tiste celice, v katerih je to izražanje znatno. Za vsako celico določi lokalno soseščino glede k v vizualizaciji najbližjih točk – celic. Vsako soseščino oceni s številom celic, ki kažejo opazno izražanje markerskih genov in to oceno obogatenosti ES dodeli središču soseščine. Sledi izračun statistične značilnosti te ocene, na podlagi katere določi regijo obogatenosti skupine markerskih genov, ki naj bi

sovpadala z regijo celic izbranega celičnega tipa na vizualizaciji. Na sliki 3.1 je grafično prikazan opisan postopek.

3.2 Implementacija

Vhodni podatki

Vhodni podatki metode so:

genska ekspresijska matrika (ang. *gene expression matrix*), ki je matrika izražanj genov v vsaki celici $\in \mathbb{R}^{N \times K}$, kjer je N število celic in K število genov,

seznam markerskih genov G, t.j. seznam genov, značilnih za tip celice, ki ga želimo odkriti,

mejna relativna vrednost izražanja skupine genov v opazovani celici glede na ostale celice FC, privzeta vrednost $FC = 1.75$,

velikost želene opazovane soseščine vsake celice k, privzeta vrednost: $k = 50$,

stopnja tveganja p_{thresh}, privzeta vrednost: $p = 1 \times 10^{-5}$.

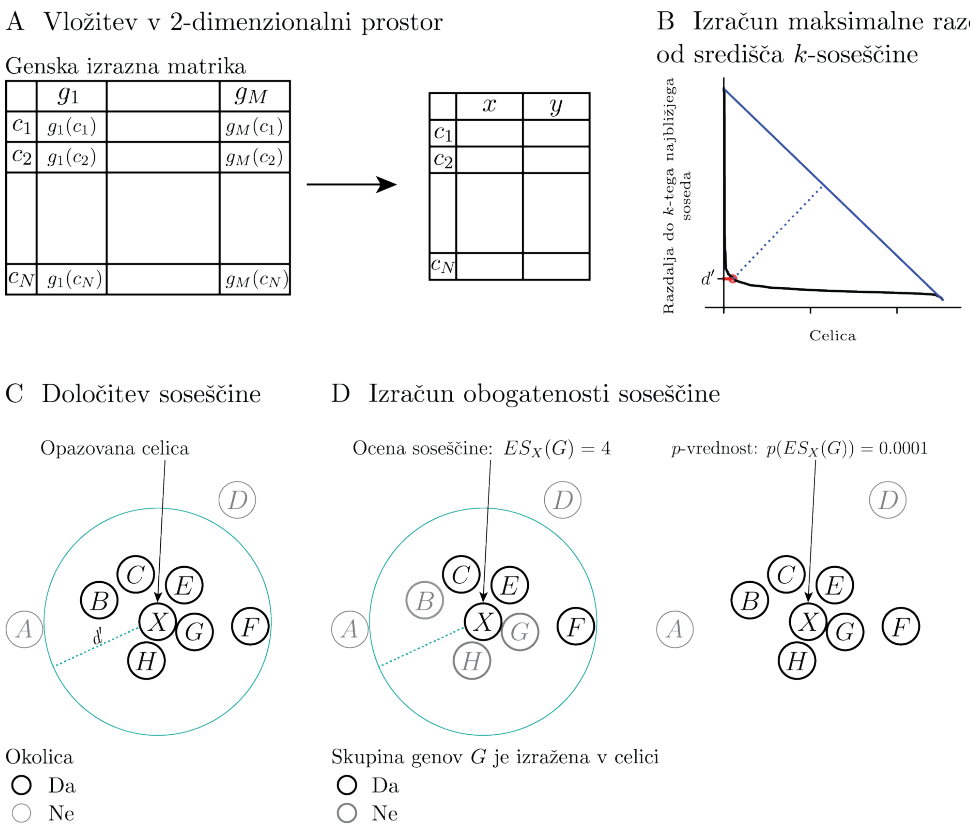
Vložitev v dvodimenzionalni prostor

Vsaka izmed N celic je predstavljena z vektorjem izražanja K genov:

$$\begin{aligned} x &\in C \subset \mathbb{R}^{N \times K} \\ x &= (x_1, x_2, ..., x_K) \end{aligned} \tag{3.1}$$

Metoda množico celic C vloži v dvorazsežen prostor s kombinacijo metode glavnih komponent (PCA) in t-SNE (t-distributed stochastic neighbor embedding). Pri PCA vzame prvih 20 glavnih komponent in nato nad novim prostorom naredi t-SNE (3.2).

$$C \subset \mathbb{R}^{N \times K} \xrightarrow{PCA_{20}} C' \subset \mathbb{R}^{N \times 20} \xrightarrow{t-SNE} C'' \subset \mathbb{R}^{N \times 2} \tag{3.2}$$



Slika 3.1: Skica predlagane metode
(A) Gensko izrazno matriko po potrebi vložimo v dvodimenzionalni prostor s kombinacijo preslikav PCA in t-SNE.
(B) S pomočjo urejenega grafa razdalj do k-tega najbližjega soseda vsake celice izračunamo maksimalno razdaljo od središča za definicijo k-soseščine.
(C) Za vsako celico X določimo njeno soseščino na podlagi izračunane razdalje.
(D) Za vsako soseščino izračunamo njeno oceno, ki je enaka številu celic, v katerih je izbrana skupina genov G izražena. Glede na globalno izraženost izbrane skupine genov izračunamo obogatenost soseščine in rezultat dodelimo celici X.
(E) Določitev regije obogatenosti genske skupine G glede na podan prag za p-vrednost.

Metoda dovoljuje tudi uporabo vnaprej določene lastne vizualizacije in s tem preskok tega koraka.

Vrednotenje izražanja skupine genov v celicah

Vsako celico metoda oceni s povprečno vrednostjo markerskih genov za opazovan tip celice (3.3):

$$w_x(G) = \frac{1}{|G|} \sum_{g \in G} g(x), \tag{3.3}$$

kjer je $g(x)$ izraz gena g v celici x. Nato določi celice, v katerih je opazovana skupina genov opazno izražena. Skupina genov je v celici opazno izražena, če je razmerje izražanja skupine genov med opazovano celico in povprečjem izražanja v vseh celicah večja od mejne relativne vrednosti izražanja FC (ang. *fold-change*) (3.4).

$$\begin{aligned} w'_x(G) &= \begin{cases} 1 & \frac{w_x(G)}{\bar{w}(G)} > FC, \\ 0 & sicer. \end{cases}, \\ \text{kjer je } \bar{w}(G) &= \frac{1}{|C|} \sum_{c \in C} w_c(G) \end{aligned} \tag{3.4}$$

Določitev lokalnih soseščin celic

Za vsako celico X metoda definira njeno lokalno soseščino C_X, tj. skupino celic, ki so od izbrane celice oddaljene največ za neko razdaljo d' (3.5).

$$C_X = \{Y \in C \mid d(X,Y) \leq d'\} \tag{3.5}$$

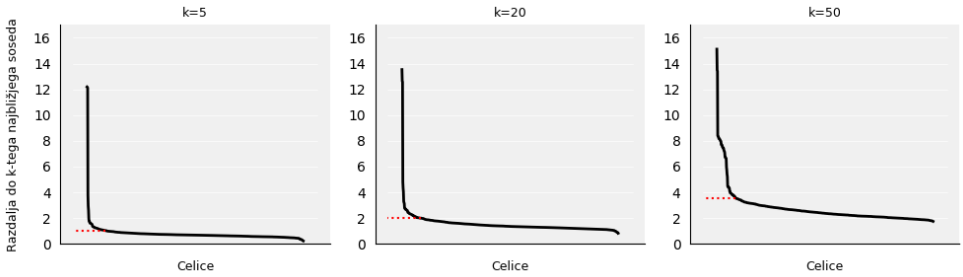
Pri tem za mero uporabi evklidsko razdaljo glede na položaj celic v projekcijskem prostoru. Razdalja med celicama X in Y, kjer je $X(x_1, y_1)$ in $Y(x_2, y_2)$, je definirana kot (3.6):

$$d(X,Y) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}. \tag{3.6}$$

Ker je za uporabnika določitev maksimalne razdalje za definicijo lokalne soseščine zaradi nepoznavanja prostora podatkov težka, metoda omogoča oceno primerne razdalje d' na podlagi navedbe želene velikosti lokalne soseščine

k. Na podlagi števila sosedov k metoda izračuna maksimalno razdaljo d' za definicijo lokalne soseščine. To stori s pomočjo grafa urejenih razdalj vsake celice do njenega k-tega najbližjega soseda po vzoru predlagane določitve parametra *eps* pri tehniki gručenja DBSCAN [27].

Na sliki 3.2 vidimo grafe urejenih razdalj celic do njihovih k-tih sosedov, grafe urejenih k-razdalj, na primeru kasneje obravnavanih bioloških podatkov za tri različne vrednosti k. V grafu urejenih k-razdalj iščemo točko največje ukrivljenosti, takoimenovano koleno – kandidat za koleno je na sliki 3.2 označen s črtkasto rdečo črto. Določitev kolena v grafu ni enostavna naloga – priporoča se ročno izbiro glede na izrisan graf urejenih k-razdalj. Zaradi hitrega izračuna kot privzeto nastavitev naše metode uporabimo poenostavljeno verzijo algoritma Kneedle po [28], ki se je izkazala za najboljšo od preverjenih hitrih metod. Za koleno izberemo tisto točko, ki je najdlje oddaljena od premice, ki teče skozi točki, ki predstavljata maksimalno in minimalno k-razdaljo, torej skozi prvo in zadnjo točko na grafu urejenih k-razdalj. Kasneje smo pokazali, da izbira razdalje ne vpliva pretirano na točnost predlagane metode, dokler razdalja ni prevelika.



Slika 3.2: Grafi urejenih k-razdalj za tri različne vrednosti k (podatki po [29]), črtkana rdeča črta označuje kandidata za koleno grafa.

Izračun obogatenosti lokalnih soseščin

Za vsako soseščino C_X metoda določi njeno oceno ES_X (ang. *enrichment score*), ki je enaka številu celic v soseščini, v katerih je skupina genov opazno

izražena (3.7).

$$ES_X(G) = \sum_{x \in C_X} w'_x(G) \tag{3.7}$$

Sledi izračun statistične značilnosti, p-vrednosti, ocenjene za dano soseščino ES_X. Ocena soseščine ES ima, tako kot smo jo definirali, hipergeometrijsko porazdelitev, zato je izračun statistične značilnosti enostaven in hiter. p-vrednost je v našem primeru verjetnost, da v skupini N celic, med katerimi je v n celicah izražena iskana genska skupina, izberemo skupino M celic ($M = |C_X|$), med katerimi je v več kot $ES_X(G)$ celicah izražena opazovana skupina genov G (3.8):

$$\begin{aligned} p(ES_X(G)) &= P(Y \geq ES_X(G)) \\ &= 1 - P(Y < ES_X(G)) \\ &= 1 - F(ES_X(G)) \\ &= 1 - \sum_{k=0}^{s} \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}} \end{aligned} \tag{3.8}$$

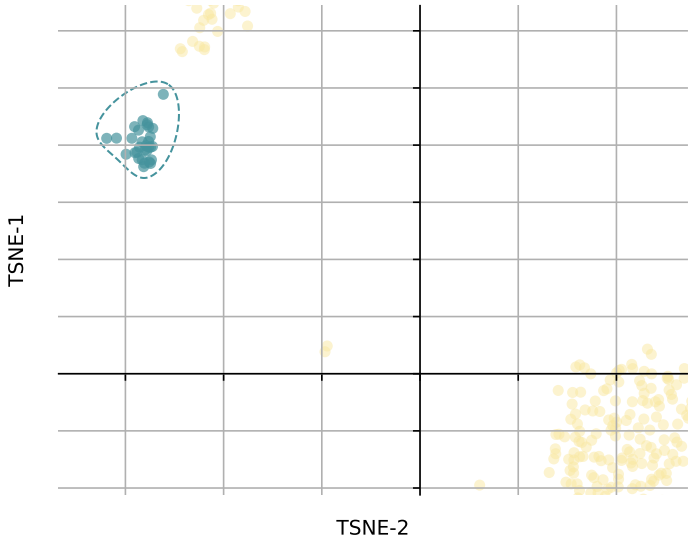
Določitev regije obogatenosti

Zadnji korak metode predstavlja določitev regije obogatenosti ER (ang. *enrichment region*). Regijo obogatenosti določimo glede na podano stopnjo tveganja, pragu za p-vrednost, p_{thresh} (3.9).

$$ER(G) = \{x \in C \mid p(ES_x(G)) < p_{thresh}\} \tag{3.9}$$

Metoda ponuja tudi možnost obrisa definirane regije s konveksno ovojnico (3.10). Primer takega obrisa regije je viden na sliki 3.3.

$$Conv(ER(G)) = \left\{ \sum_{i=1}^{|ER(G)|} \alpha_i x_i \mid (\forall i : \alpha_i > 0) \wedge \sum_{i=1}^{|ER(G)|} \alpha_i = 1 \right\} \tag{3.10}$$



Slika 3.3: Obris odkrite skupine točk, ki so značilne za izbrani tip celic, s konveksno ovojnico.

Poglavje 4

Poskusi in rezultati

Predlagano metodo smo uporabili pri anlizi nekaterih nedavno objavljenih podatkov s področja študij izražanja genov v posameznih celicah.

4.1 Podatki

Poskuse smo izvedli na treh različnih skupinah scRNA podatkov.

Dva nabora podatkov smo pridobili iz novejših člankov, ki preučujejo specifične skupine celic in jih preko metod sekvenciranja na nivoju posameznih celic scRNA-seq želijo podrobneje klasificirati. Podatki smo izbrali zato, ker avtorji obeh člankov podajo vizualizacije in klasifikacije vseh proučevanih celic. Poleg tega opredelijo tudi nabor markerskih genov za vsakega izmed razredov [29, 30]. To nam je služilo kot izhodišče za preverjanje uspešnosti naše metode odkrivanja teh skupin celic.

Tretji sklop podatkov smo pridobili iz članka [15], ki opisuje novo metodo sekvenciranja na nivoju posametnih celic, v okviru katere so sekvencirali več različnih sklopov celic. Odkrivali smo skupine prisotne v podatkih glede na najbolj tipične markerske gene izbranih tipov celic.

Človeške krvne celice z visoko gensko tipizacijo HLA-DR

Prve podatke smo povzeli po članku, ki je z analizo zdrave človeške krvi s pomočjo metod scRNA-seq odkril nove tipe dendritičnih celic v človeški krvi, monocitov in progenitornih celic [29]. Iz vzorca zdrave človeške krvi so izolirali približno 2400 celic, ki so kazali visoko tipizacijo tkivnega antigena HLA-DR in jih sekvencirali na nivoju posameznih celic. To sekvenciranje jih je, skupaj s kasnejšim porfiliranjem in funkcijsko karakterizacijo, privedlo do posodobitve originalne klasifikacije krvnih celic z vključitvijo šestih razredov dendritičnih celic (ang. *dendritic cells*, DCs), štirih podtipov monocitov (ang. *monocytes*) in ene DC progenitorne celice (ang. *DC progenitor*) [29].

Vzeli smo podatke o 742 dendritičnih celicah, za vsako izmed njih je bilo podano izražanje 26593 genov. Avtorji so celice klasificirali v 6 razredov, za katere so podali med 31 in 390 markerskih genov. Pregled markerskih genov je razviden v tabeli 4.1. Redkost podatkov (ang. *data sparsity*), tj. delež ničelnih vnosov v genski izrazni matriki, izpeljani iz podatkov, je 0.8.

Tabela 4.1: Tabela skupin dendritičnih celic s seznami za skupine značilnih markerskih genov po [29].

Oznaka skupine	Opis	Nekaj najbolj značilnih markerskih genov	Št. vseh markerjev
DC1	CD141/CLEC9A	CLEC9A, C1ORF54, HLA-DPA1	113
DC2	CD1C_A	CD1C, FCER1A, CLEC10A, ADAM8	31
DC3	CD1C_B	S100A9, S100A8, VCAN, LYZ, ANXA1	59
DC4	$CD1C^-$, $CD141^-$	FCGR3A, FTL, SERPINA1, LST1	342
DC5	nova populacija: AXL^+, $SIGLEC6^+$	AXL, PPP1R14A, SIGLEC6, CD22, DAB2	85
DC6	plazmacitoidne dendritične celice	GZMB, IGJ, AK128525, SERPINF1, ITM2C	390

Bipolarne mrežnične živčne celice miši

Drugi sklop podatkov smo črpali iz članka, ki je uporabnost transkriptomike posameznih celic pokazal na živčnih celicah mrežnice miši (ang. *mouse retinal bipolar cells*). Bipolarne celice predstavljajo približno 7% vseh mrežničnih celic v miših. Te so avtorji članka pridobili iz transgenih[1] miši z visokim izražanjem GFP (zelenega fluoscirajočega proteina) in jih preko sortiranja fluorscenčno aktivnih celic (FACS) izolirali. Sledilo je sekvenciranje na nivoju posameznih celic scRNA-seq z metodo Drop-seq (več o tem v [32]). S tem so pridobili gensko ekspresijsko matriko, s pomočjo katere so, prek analize z metodami bioinformatike, preslikanjem v nižji prostor in gručenjem (ang. *clustering*), pridobili molekularno klasifikacijo. Ta med celicami identificira več razredov, med katerimi sta dva popolnoma nova. Markerske gene za odkrite razrede so identificirali s kombinacijo metod analize obogatenosti genov in metod analize morfologije celic [30].

Vzeli smo del podatkov v velikosti 6383 celic, za vsako celico so bile podane vrednosti izražanja 13166 genov. Redkost podatkov je 0.93. Na podatkih so avtorji odkrili 26 razredov, od tega so jih za informativne označili 18 in le-te povezali z znanimi skupinami celic, paličnicami, čepnicami, amakrinimi in Müllerjevimi celicami, bipolarnimi paličnicami in bipolarnimi čepnicami tipov BC, ostale celice so označili kot dvojnike oz. kontaminante (ang. *doublets/contaminants*, D/C). Paličnice in čepnice zaznavajo in prenašajo svetlobne dražljaje v kemične signale na bipolarne paličnice (RBC) in bipolarne čepnice (BC), tip BC je odvisen od globine sinapse bipolarne celice na ganglijske celice mrežnice v notranjem mrežastem skladu mrežnice (ang. *inner plexiform layer*, IPL). Ilustracijo in opis za lažjo predstavo najdemo v [33]. Za vsakega izmed razredov je bilo podanih med 63 in 374 markerskih genov. Markerski geni so podani v tabeli 4.2 [30].

[1] **transgen**: *ki ima dodan ali spremenjen gen:* transgeni organizem; proizvodnja transgene hrane; transgena podgana; transgena rastlina [31]

Tabela 4.2: Tabela skupin bipolarnih celic mrežnice miši s seznami za skupine značilnih markerskih genov po [30].

Oznaka skupine	Opis	Nekaj najbolj značilnih markerskih genov	Št. vseh markerjev
BC1A		Wfdc2, Mylk, Igsf3, Fat4, Pcdh17	72
BC1B		Fezf1, RP23-124H2.4, Nxph1, Wls	66
BC2		Ebf1, Lamp5, Kcna1, Syt2, Neto1	125
BC3A		Angpt1, Erbb4, Kcnh7, Irx6, Irx5	84
BC3B		Gng2, A930009A15Rik, Syt13, Ift20	80
BC4		Col11a1, Calml4, Nxph4, Postn	116
BC5A	bipolarne celice	BC046251, Ptprt, Sox6, Neurod2, Hpca	105
BC5B		Chrm2, Adarb2, C1ql3, Ddit4l, Nwd2	97
BC5C		Areg, Slitrk5, Tmem200a, Gabrr3	87
BC5D		Irx3, Kirrel3, Lrrtm1, Cxcl14, Ddit4l	128
BC6		Cck, Spock3, Lect1, Lhx3, Cdh8	63
BC7		Gnat3, A930006I01Rik, Rgs3, Kcnab1	104
BC8/9		Cpne9, Tshz2, Nsg2, Adarb2, Plcb1	66
AC	amakrine celice	C1ql2, Slc32a1, Gad1, Tfap2a, Kcnip1	351
MG	Müllerjeve celice	Slc14a1, Gm3764, Cd44, Bmp3	374
RBC	bipolarne paličnice	Lrrtm4, Kcne2, Il1rap, Casp7, Adrb1	100
RP	paličnice	A930036K2Rik, Samd7, Cabp4, Impg1	200
CP	čepnice	Arr3, Pde6h, Opn1mw, Opn1sw, Egflam	249

Celice kostnega mozga bolnika z AML pred presaditvijo in zdravega človeka

Zadnji uporabljeni podatki so bili eni izmed rezultatov nove visoko paralelne metode sekveniranja scRNA-seq, predlaganega v [15]. Avtorji članka predlagajo novo metodo na osnovi kapljične mikrofluidike, ki omogoča digitalno štetje mRNA tisočih posameznih celic. Zajame približno polovico vseh celic in lahko procesira do osem vzorcev hkrati. Metodo so med drugim preverili na mononuklearnih celicah kostnega mozga bolnika z akutno mieloično levkemijo (AML) pred presaditvijo krvotvornih matičnih celic in enakih celicah zdravega človeka [15].

Na voljo smo imeli podatke o izražanjih 1004 genov v 8390 celicah, od tega predstavljajo 47% celice kostnega mozga bolnika z AML in 53% celice zdravega človeka. Redkost podatkov je 0.93. Na podatkih smo želeli poiskati znane celične tipe, ki so prisotni v tem tkivu in sicer B celice, T celice in monocite. Markerske gene smo povzeli po članku [15] in so razvidni v tabeli 4.3.

Tabela 4.3: Tabela celičnih tipov, prisotnih v celicah kostnega mozga, s seznami za skupine značilnih markerskih genov po [15].

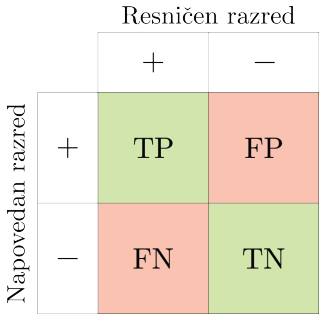
Skupina	Markerski geni
B celice	CD19, CD79A
T celice	CD4, CD3D
Monociti	CD14, S100A9

4.2 Metode vrednotenja

Uspešnost predlagane metode smo lahko ovrednotili na dveh sklopih podatkov, človeških krvnih celicah in bipolarnih mrežničnih celicah miši. Na voljo smo imeli klasifikacijo celic v razrede in sezname markerskih genov za vsakega izmed odkritih omenjenih razredov. Metodo smo pognali za vsakega izmed razredov in rezultat združili v eno napoved.

Uspešnost klasifikacijskih napovedi pogosto prikažemo z matriko klasifikacij (ang. *confusion matrix*), ki je predstavljena na sliki 4.1. Na diagonali najdemo pravilne napovedi, pravilno klasificirani pozitivni primeri (ang. *true positives*, TP) in pravilno klasificirani negativni primeri (ang. *true negatives*, TN), drugod pa napačne, primere, napačno klasificirane kot pozitivne (ang. *false positives*, FP) in primere, napačno klasificirane kot negativne (ang. *false negatives*, FP). S pomočjo le-te lahko izpeljemo več mer uspešnosti [34, 35].

Uspešnost predlagane metode smo merili s tremi statistikami, klasifikacijsko točnostjo in mero F1, ki smo ju izračunali za vsak razred posebej in izračunali povprečje, ter odstotkom pravilnih napovedi po zgledu iz [36].



Slika 4.1: Matrika klasifikacij

Klasifikacijska točnost (ang. *accuracy*) predstavlja odstotek pravilno klasificiranih primerov (4.1).

$$točnost = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.1}$$

Mera F1 je definirana kot harmonično povprečje med natančnostjo (ang. *precision*) (4.2), ki je razmerje med pravilno klasificiranimi pozitivnimi primeri in vsemi napovedanimi pozitivnimi primeri, in priklicem (ang. *recall*) (4.3), ki je razmerje med pravilno klasificiranimi pozitivnimi primeri in vsemi v resnici pozitivnimi primeri (4.4) [34, 35].

$$natančnost = \frac{TP}{TP + FP} \tag{4.2}$$

$$priklic = \frac{TP}{TP + FN} \tag{4.3}$$

$$F1 = 2 \cdot \frac{natančnost \cdot priklic}{natančnost + priklic} \tag{4.4}$$

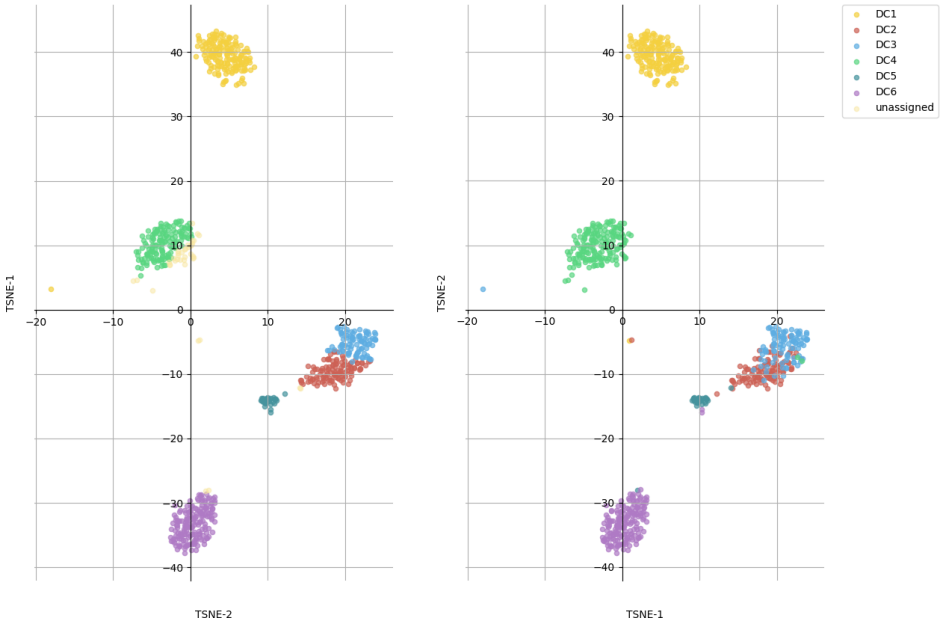
Odstotek pravilnih napovedi upošteva klasifikacijo v vse razrede hkrati in je razmerje med seštevkom pravilno klasificiranih pozitivnih primerov za vsak razred, ki je enako številu primerov, klasificiranih v pravilni razred, in številom vseh primerov (4.5):

$$Q = \frac{\sum_{i \in C} TP_i}{N}, \tag{4.5}$$

kjer je C množica razredov, TP_i število pravilno klasificiranih pozitivnih primerov za razred i in N število vseh primerov [36].

4.3 Rezultati z analizo robustnosti

Uspešnost metode smo najprej preverjali na podatkih o človeških krvnih celicah [29]. Na sliki 4.2 je grafični prikaz primerjave skupin, odkritih s predlagano metodo ($k = 30$), s skupinami, klasificiranimi v članku [29]. Razvidno je, da je metoda dobro odkrila izolirane skupine DC1, DC4, DC5 in DC6. Dobro je odkrila tudi skupni prostor kombinacije skupin DC2 in DC3, slabše pa mejno območje med omenjenima. Slabše odkritje teh dveh skupin ni presenečenje, saj sta obe skupini novi in obe izhajata iz predhodne skupine z močnim izražanjem gena DC1C, avtorji pa so, poleg markerskih genov, navedli tudi posebne gene, ki razlikujejo izključno med omenjenima skupinama.



Slika 4.2: Poskus odkritja skupin celic, predlaganih v [29]. Levo je rezultat predlagane metode, desno so skupine označene po [29].

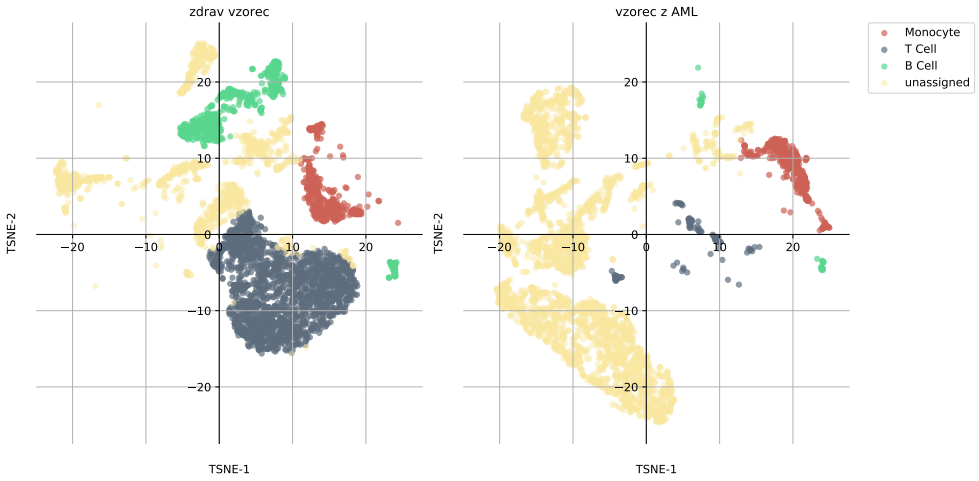
S preverjanjem uspešnosti metode smo nadaljevali na drugi množici podatkov, bipolarnih živčnih celicah mrežnice miši. Na sliki 4.3 je grafični prikaz razvrstitve celic po celičnih tipih, kot jo je predlagala naša metoda

($k = 100$), na sliki 4.4 je razdelitev celic v skupine celičnih tipov, kot so jih predlagali v [30]. Opazimo lahko, da je metoda zelo dobro odkrila večino celičnih tipov. Slabše je odkrila najbolj razpršeno skupino, skupino bipolarnih paličnic (ang. *rod bipolar cells*, RBC), mejo med na novo definiranima podrazredoma bipolarnih celic BC1, BC1A in BC1B, ter mejo med razredoma bipolarnih celic BC3B in BC4 – to ni nenavadno, saj sta si razreda zelo blizu in jih denimo metoda sekvenciranja Smart-seq2 sploh ni mogla razločiti [30].

Slika 4.3: Poskus odkritja celičnih tipov, odkritih v bipolarnih celicah mrežnice miši [30].

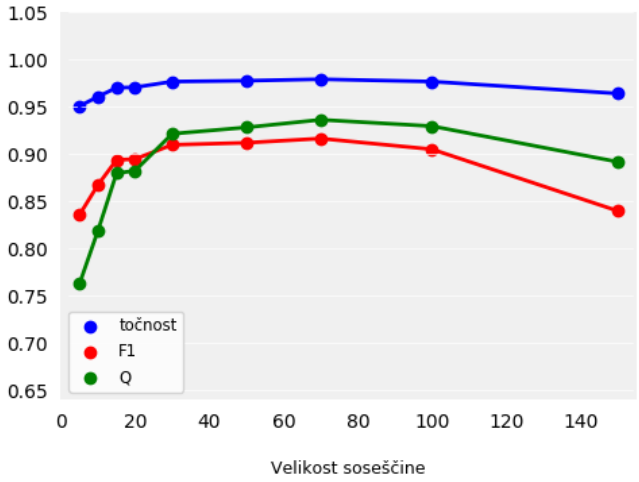
Slika 4.4: Razdelitev celic v skupine po celičnih tipih, predlagana v [30].

Nazadnje smo metodo preizkusili še na podatkih o celicah kostnega mozga bolnika z akutno mieloblastno levkemijo (AML) pred presaditvijo in celicah enakega tkiva zdravega človeka [15]. Na sliki 4.5 je razvidno, kako je metoda odkrila izbrane celične tipe na podatkih iz [15]. V vzorcu z AML je metoda odkrila bistveno manj T celic in B celic kot v zdravem vzorcu, medtem ko je število odkritih monocitov v obeh vzorcih približno enako. Do podobne ugotovitve so prišli tudi avtorji članka po katerem smo povzeli podatke [15].

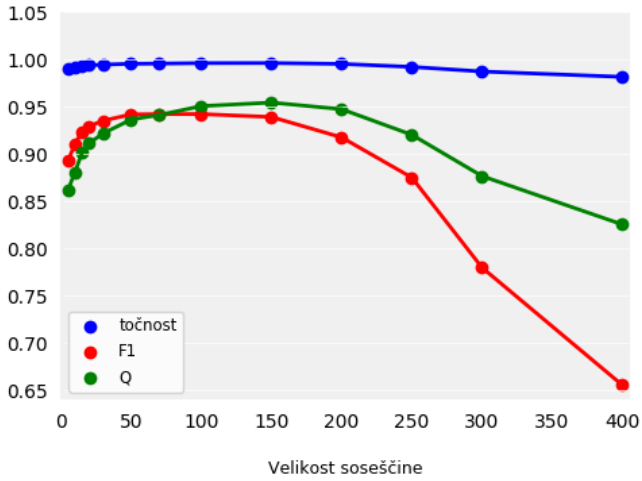


Slika 4.5: Poskus odkritja T-celic, B-celic in monocitov na podatkih o celicah kostnega mozga [15].

Na obeh množicah podatkov smo preverjali vpliv parametra velikosti lokalne soseščine k. Poskuse smo izvedli z 9 različnimi parametri za velikosti lokalne soseščine $k \in [5, 150]$ na prvih in 13 različnimi $k \in [5, 400]$ na drugih podatkih. Na slikah 4.6 in 4.7 je razvidno, da šele zelo velika velikost soseščine glede na velikost vhodnih podatkov bistveno vpliva na uspešnost napovedi naše metode.



Slika 4.6: Vpliv velikosti soseščine na uspešnost metode



Slika 4.7: Vpliv velikosti soseščine na uspešnost metode

Kako na odkrivanje skupin vpliva vložitev v dvorazsežen prostor, ki smo ga izvedli prek PCA s številom glavnih komponent 20 in t-SNE, smo preverjali le na prvi množici podatkov. Vložitev v 2D prostor je ozko grlo metode in porabi največ časa, zato smo si izbrali manjšo izmed obeh množic podatkov. Izvedli smo 100 ponovitev metode. Izkazalo se je, da vložitev v 2D prostor, ne vpliva bistveno na rezultat, kar je razvidno v tabeli 4.4.

Tabela 4.4: Tabela izračunanih mer za preverjanje stabilnosti metode glede na začetno vložitev v 2D prostor.

mera	povprečje	standardni odklon
povprečje klasifikacijskih točnosti	0.9792	0.0023
povprečje f1	0.9020	0.0080
odstotek pravilnih napovedi	0.9295	0.0078

Poglavje 5

Zaključek

Predlagali smo novo metodo za odkrivanje celičnih tipov na vizualizacijah podatkov scRNA v 2D prostoru. Metodo smo opisali in preizkusili na treh različnih naborih podatkov. Glede na spodbudne rezultate opravljenih poskusov bi lahko nadaljevali z razvojem metode, preverjanjem uspešnosti in izboljšanjem posameznih korakov. Izvorna koda je napisana v programskem jeziku Python in je dostopna na portalu Github[1].

Pokazali smo, da metoda glede na podane markerske gene dokaj dobro poišče območja celic izbranih celičnih tipov. Časovna kompleksnost predlagane metode za iskanje regij obogatenosti ene skupine genov je pogojena z začetno vložitvijo v dvorazsežen prostor in je enaka $\mathcal{O}(max(n,m)^2 \cdot min(n,m))$, kjer je n število celic in m število genov, s katerimi je vsaka celica opisana. Časovna zahtevnost je linearna, $\mathcal{O}(n)$, če podamo začetno vložitev v dvodimenzionalni prostor. Za podatke v velikosti $n \doteq 8000$, $m \doteq 1000$ porabi približno eno minuto.

Da bi preverili vsestranskost predlagane metode, bi uporabnost in uspešnost metode v okviru nadaljnjega dela morali testirali na še več različnih množicah podatkov. Smiselno bi bilo poskusiti označiti izraze genske ontologije, s čimer bi lahko poiskali regije funkcionalno podobnih celic in se ne bi omejevali na celične tipe, za katere so markerski geni že znani.

[1] `https://github.com/anejaf/label-scrna-vis`

Pokazali smo tudi, da je delovanje metode stabilno glede na začetno vložitev v dvodimenzionalni prostor. V predlogu metode smo uporabili kombinacijo dveh znanih metod za zmanjšanje dimenzije podatkov, PCA in t-SNE, in se pri tem nismo ozirali na domeno podatkov. Smiselno bi bilo mogoče preveriti, kako se obnese začetno informativno zmanjševanje dimenzij z izbiro informativnih genov [37].

Eden izmed mogoče nepotrebnih parametrov metode je želena velikost lokalnih soseščin. Kako izbrana velikost soseščine k vpliva na točnost podatkov, smo sicer preverili, vendar bi lahko naredili razširjeno analizo in poskušali parameter k oz. maksimalno razdaljo za definicijo lokalne soseščine določiti glede na podatke same in s tem izboljšali uproabniško izkušnjo.

Metoda, kot smo jo predlagali, deluje na binarni oceni celic. Celice najprej oceni s povprečno vrednostjo markerskih genov za opazovan tip celice in nato določi celice, v katerih so ti geni opazno izraženi. Binarno oceno celic uporabljamo zato, ker lahko potem poenostavimo izračun statistične značilnosti neke skupine celic s pomočjo hipergeometrijske porazdelitve. V začetnem razvoju metode smo hoteli, da bi le-ta omogočala tudi empirični izračun statistične značilnosti, s čimer bi lahko vsako celico ocenili z numerično vrednostjo glede na količino markerskih genov in ne bi bilo potrebe po binarni oceni. To se je izkazalo kot časovno zelo zahtevna operacija, zato bi v okviru nadaljnjega dela lahko zagotovili njeno optimizacijo.

Literatura

[1] M. Friendly and D. Denis. The early origins and development of the scatterplot. *Journal of the History of Behavioral Sciences*, 41(2):103–130, 2005.

[2] International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*, 409:860–921, 2001.

[3] J. C. Venter, M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, et al. The sequence of the human genome. *Science*, 291(5507):1304–1351, 2001.

[4] J. Adams. Transcriptome: Connecting the genome to gene function. *Nature Education*, 1(1):195, 2008.

[5] O. Stegle, S. A. Teichmann, and J. C. Marioni. Computational and analytical challenges in single-cell transcriptomics. *Nature Reviews Genetics*, 16:133–145, 2015.

[6] D. Wang and S. Bodovitz. Single cell analysis: the new frontier in 'omics'. *Trends in Biotechnology*, 28(6):281–290, 2010.

[7] V. Proserpio and T. Lönnberg. Single-cell technologies are revolutionizing the approach to rare cells. *Immunology and Cell Biology*, 94:225–229, 2016.

[8] R. Sandberg. Entering the era of single-cell transcriptomics in biology and medicine. *Nature Methods*, 11(1):22–24, 2014.

[9] H. Hotelling. *Analysis of a complex of statistical variables into principal components*. Warwick and York, 1933.

[10] H. Yin. Nonlinear multidimensional data projection and visualisation. *IDEAL, Lecture Notes in Computer Science*, 2690:377–388, 2003.

[11] J. B. Kruskal. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29, 1964.

[12] L. van der Maaten, E. O. Postma, and H. J. van den Herik. Dimensionality reduction: A comparative review. *Journal of Machine Learning Research*, 10:66–71, 2009.

[13] G. E. Hinton and S. T. Roweis. Stochastic neighbor embedding. *Advances in Neural Information Processing Systems*, pages 857–864, 2003.

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

[15] G. X. Y. Zheng, J. M. Terry, et al. Massively parallel digital transcriptional profiling of single cells. *Nature Communications*, 8, 2017.

[16] D. W. Huang, B. T. Sherman, and R. A. Lempicki. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research*, 37(1):1–13, 2009.

[17] D. Nam and S.-Y. Kim. Gene-set approach for expression pattern analysis. *Briefing in Bioinformatics*, 9(3):189–197, 2008.

[18] K.-H. Pan, C.-J. Lih, and S. N. Cohen. Effects of threshold choice on biological conclusions reached during analysis of gene expression by dna microarrays. *Proceedings of the National Academy of Sciences of the United States of America*, 102(25):8961–8965, 2005.

[19] M. Ashburner, C. A. Ball, et al. The gene ontology consortium. *Nature Genetics*, 25(1):25–29, 2000.

[20] M. Kanehisa and S. Goto. Kegg: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28(1):27–30, 2000.

[21] J. J. Goeman and P. Bühlmann. Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics*, 23(8):980–987, 2007.

[22] M. Mramor, M. Toplak, T. Curk, and B. Zupan. Uporaba skupin genov pri analizi podatkov o izraženosti genov pri raku. *Informatica Medica Slovenica*, 13(2):1–10, 2008.

[23] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *PNAS*, 102(43):15545–15550, 2008.

[24] J. Demšar, T. Curk, A. Erjavec, Č. Gorup, T. Hočevar, M. Milutinovič, M. Možina, M. Polajnar, M. Toplak, A. Starič, M. Štajdohar, L. Umek, L. Žagar, J. Žbontar, M. Žitnik, and B. Zupan. Orange: Data mining toolbox in python. *Journal of Machine Learning Research*, 14:2349–2353, 2013.

[25] F.-L. Zhou, W.-G. Zhang, Y.-C. Wei, S. Meng, G.-G. Bai, B.-Y. Wang, et al. Involvement of oxidative stress in the relapse of acute myeloid leukemia. *The Journal of Biological Chemistry*, 285(20):15010–15015, 2010.

[26] A. Baryshnikova. Systematic functional annotation and visualization of biological networks. *Cell Systems*, 2(6):412–421, 2016.

[27] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters a density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, pages 226–231. AAAI Press, 1996.

[28] V. Satopaa, J. Albrecht, D. Irwin, and B. Raghavan. Finding a "kneedle"in a haystack: Detecting knee points in system behavior. In *2011 31st International Conference on Distributed Computing Systems Workshops*, pages 166–171, 2011.

[29] A.-C. Villani, R. Satija, G. Reynolds, S. Sarkizova, K. Shekhar, J. Fletcher, et al. Single-cell rna-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. *Science*, 356(6635), 2017.

[30] K. Shekhar et al. Comprehensive classification of retinal bipolar neurons by single-cell transcriptomics. *Cell*, 166(5), 2016.

[31] Sodelavci Inštituta za slovenski jezik Frana Ramovša ZRC SAZU. Slovar slovenskega knjižnega jezika. Dosegljivo: https://fran.si/130/sskj-slovar-slovenskega-knjiznega-jezika. [Dostopano: 23. 5. 2018].

[32] E. Z. Macosko, A. Basu, R. Satija, et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, 161(5):1202–1214, 2015.

[33] T. Euler, S. Haverkamp, T. Schubert, and T. Baden. Retinal bipolar cells: elementary building blocks of vision. *Nature Reviews Neuroscience*, 15:507–519, 2014.

[34] D. J. Hand, H. Mannila, and P. Smyth. *Principles of Data Mining*. The MIT Press, 2001.

[35] D. M. W. Powers. Evaluation: From precision, recall and f-factor to roc, informedness, markedness & correlation. *International Journal of Machine Learning Technology*, 2:37–63, 2011.

[36] P. Baldi, S. Brunak, Y. Chauvin, C. A. F. Andersen, and H. Nielsen. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics Review*, 16:412–424, 2000.

[37] T. S. Andrews and M. Hemberg. Identifying cell populations with scrnaseq. *Molecular Aspects of Medicine*, 59:114–122, 2018.