KLASTEROVANJE PET TIPOVA ĆELIJA IZ PBMC UZORKA.

A PREPRINT

Aleksandar Vračarević
Matematički fakultet
Beograd
vracarevicaleksandar@gmail.com

February 26, 2020

***K*eywords** Clustering · RNAseq

1 Uvod

U skorije vreme se javio veliki broj radova koji koristi tehniku sekvenciranja RNK. Iz tog razloga, bioinformatičari se bave razvitkom raznih alata za analizu ovih podataka. Jedan od radova koji se bavi ovom vrstom analize je [1]. Autori rada su koristili veštačke neuronske mreže za klasifikaciju ljudskih ćelija na osnovu izraženih gena. Autori pomenutog rada su uspeli da postignu preciznost od oko 90% korišćenjem neuronske mreže. Cilj ovog istraživanja je upoređivanje njihovih rezultata i rezultata dobijenih raznim tehnikama klasterovanja.

2 Opis podataka

Za razliku od metoda koji rade nad gomilama podataka *bulk transcriptome measurements*, transkriptomika pojedinačnih ćelija (*single cell transcriptomics*) omogućava otkrivanje heterogenih populacija ćelija, rekonstrukciju razvojnih putanja ćelija. Naime, meri se nivo informacione RNK u pojedinačnim ćelijama i onda se nad takvim podacima vrši analiza. U ovom radu korišćene su dve grupe podataka: ćelije srca i ćelije aorte kućnog miša (*mus musculus*). Podaci su zapisani u vidu retkih matrica, gde su vrste nazivi gena, a kolone predstavljaju pojedinačne ćelije i vrednosti u ćelijama predstavljaju nivo izmerene informacione RNK sa izraženim određenim genima u pojedinačnim ćelijama. Genom na osnovu koga su obeleženi geni je mm10. Peripheral blood mononuclear cells ili PBMC su ćelije koje imaju jedno sferno jezgro u kome se nalazi genetski materijal. Postoji 5 tipova ovih ćelija: B ćelije (BC), T ćelije (TC) i ćelije ubice *natural killer* (NK) ćelije, monociti (MC) and dendritske ćelije (DC). Procenat zastupljenosti ovih ćelija varira od pojedinca do pojedinca, ali u proseku je tačno da B ćelije čine 5-15%, monociti čine 10-30%, DC čine 1-2%, NK ćelije čine 5-10%, i T ćelije čine 40-70% svih PBMC ćelija kod čoveka.

Svi podaci su organizovani na sledeći način:

- GSM3308814 i GSM3308815 - uzorci iz ćelija tkiva srca kućnog miša
- GSM3316206 i GSM3316207 - uzorci iz ćelija tkiva aorte kućnog miša
- SCT-10x-Metadata_readylist_merged-PBMC-tasks-short-Bgd - metapodaci o raznim uzorcima, nama je potrebna samo informacija o genomu odgovarajućih uzoraka
- common_mouse_list.csv - veza između ćelija i u njima izraženih gena odgovarajućeg genoma

3 Priprema okruženja

Kako bi se skripte korišćene u okviru rada koristile potrebno je najpre podesiti radnu stanicu sa odgovarajućim paketima. Sve skripte su pisane u programskom jeziku python. Eksperimenti su vršeni na Linux sistemu sa 25GB RAM memorije i procesorom sa 4 jezgra arhitekture x86_64.

```
pip install scanpy
pip install python-igraph
pip install louvain
pip install leidenalg
pip install multicoreTSNE
```

Listing 1: Instaliranje neophodnih paketa

4 Preprocesiranje podataka

Vodeći se smernicama iz [2], podaci su preprocesirani na više načina. U obradi su korišćene biblioteke scanpy, pandas, scikit, matplotlib programskog jezika python.

4.1 Priprema podataka

Alati koji se koriste u nastavku očekuju da podaci budu određenog oblika, te je stoga bila neophodna priprema istih. Najpre su iz ulaznih datoteka uklonjeni oni unosi za koje nema podataka u konsultacionoj datoteci *common_mouse_list*. Nakon toga je na osnovu datoteke *SCT-10x-Metadata_readylist_merged-PBMC-tasks-short-Bgd* zaključeno da će se genom *mm10* koristiti u nastavku rada. Indeksi ćelija zamenjeni su njihovim rednim brojevima i tako dobijena matrica je transponovana. Implementacija ovog koraka se nalazi u skripti *prepare_data.py*.

4.2 Kontrola kvaliteta

Kako bi se analiza nastavila, potrebno je ispitati kvalitet podataka. Za ovaj korak se najčešće koriste tri veličine: količina iRNK po ćeliji, broj izraženih gena po ćeliji i procenat mitohondrijske RNK po ćeliji[2]. Generalno je poželjno da ćelije imaju veliku količinu očitane iRNK i broj izraženih gena, a mali procenat mitohondrijske RNK. (Prisustvo mitohondrijske RNK obično signalizira da sa ćelijom nešto nije u redu[3]). Bitno je naglasiti da bi trebalo ove veličine posmatrati zajedno i da bi pojedinačno tumačenje moglo dovesti do pogrešnih zaključaka. Na primer, ćelije koje su uključene u respiratorne procese mogu imati veću koncentraciju mitohondrijske RNK, takođe je moguće da su ćelije sa veoma velikim brojem molekula iRNK veće od ostalih. Ispitivanjem ovih vrednosti i postavljanjem granica odsecanja, mogu se eliminisati umiruće ćelije (veliki procenat mitohondrijkse RNK, mali broj molekula iRNK i mali broj izraženih gena) ili duplirani uzorci (neobično veliki broj molekula iRNK). Kod uz pomoć koga su dobijeni sledeći rezultati je u skripti *qc.py*.

Violinski grafici predstavljaju kombinaciju kutijastih dijagrama i aproksimacije gustine raspodele neke promenljive. Sa obe strane središnje linije nalaze se procene gustine za vrednosti promenljive. U najdebljem delu grafika je najveća verovatnoća da će podatak uzeti vrednost sa y ose, a u najtanjem je verovatnoća najmanja.

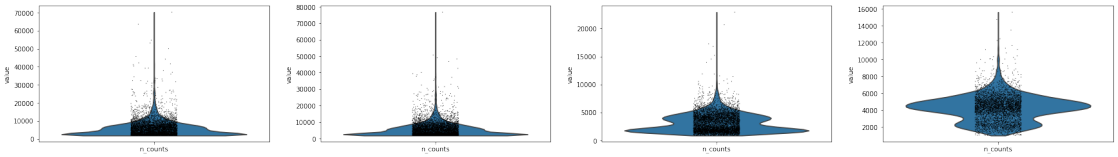

Figure 1: Violinski grafici količine izmerenih molekula iRNK za (a) GSM3308814 (b) GSM3308815 (c) GSM3316206 (d) GSM3316207. Na y osi na prvoj slici predstavljen je broj očitanih molekula iRNK dok x osa služi za vizuelizaciju gustine raspodele broja molekula iRNK. Crne tačke predstavljaju prave podatke i vidi se da im je najveća koncentracija tamo gde je grafik najdeblji odnosno gde je verovatnoća najveća.

Primećujemo da se najveći broj ćelija u prvoj grupi (GSM3308814 i GSM3308815) nalazi u opsegu od 0 do 10000 očitanih jedinica iRNK ali da očitane vrednosti idu čak i do 70000. Ne čudi što su rezultati ove dve datoteke gotovo identični jer je GSM3308815 samo kopiran i koristi se za verifikaciju samog procesa uzorkovanja (vidi kolonu 'DESCRIPTION' u datoteci sa metapodacima). U drugoj grupi je situacija drugačija: raspodela vrednosti ima malo nepravilniji oblik i manji opseg a uzorci su međusobno manje slični.

Kada se posmatraju violinski grafici procenta mitohondrijske RNK, primećuju se veliki procenti u prvoj grupi (raspodela je dosta široka u predelu od oko 40%, što je mnogo više od preporučenih 20%-25%[4]) ali još uvek neće biti izvedeni zaključci o kvalitetu uzorka zbog potencijalnih problema koji se mogu javiti i spomenuti su ranije u tekstu. Razlika

između uzoraka možda potiče od različitih stupnjeva u razvoju ćelija između dva merenja. Takođe je moguće da se razlika javlja usled samog procesa merenja.

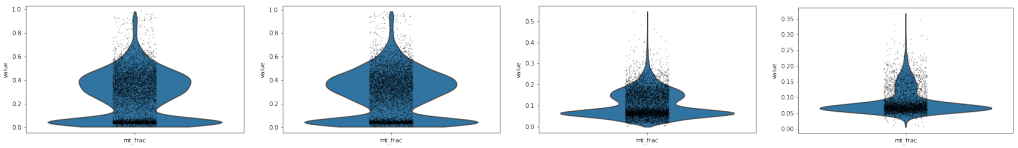

Figure 2: Violinski grafici broja izraženih gena za (a) GSM3308814 (b) GSM3308815 (c) GSM3316206 (d) GSM3316207

Druga grupa uzoraka ima znatno slabije izraženu mitohondrijsku RNK koja se kreće u preporučenim okvirima (najveći deo ćelija ima manje od 20% mitohondrijske RNK). Iako bi se samo na osnovu ovih zapažanja u vezi mtRNK moglo zaključiti da prva grupa uzoraka nema dovoljno dobar kvalitet, ovi podaci će biti korišćeni i u ostatku rada. Na slici 3 je podrobnije ispitan udeo mitohondrijske RNK u prvoj grupi kako bi se podaci dodatno modifikovali.

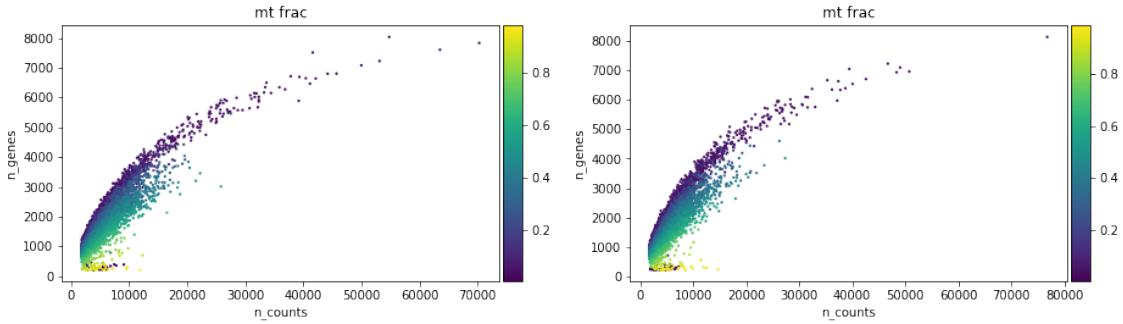

Figure 3: Odnos između broja izraženih gena i broja očitanih molekula iRNK. Procenat mitohondrijske RNK predstavljen je uz pomoć boje (a) GSM3308814 (b) GSM3308815

Jasno se uočava oblast tačaka sa malim brojem izraženih gena i malom količinom iRNK, a velikom koncentracijom mtRNK u donjem levom uglu na oba grafika (prikazano žutom bojom). Dobra stvar je što su nepoželjne tačke u nekoj meri linearno odvojive od glavnice podataka. Da bi se donela odluka o granicama odsecanja ovih tačaka, vršena je dodatna analiza raspodele ovih tačaka (slika 4):

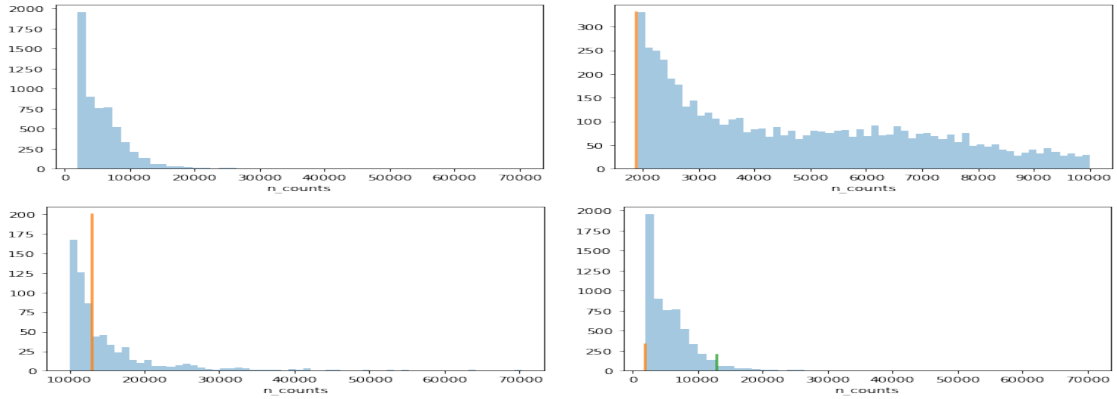

Figure 4: Raspodele broja izmerenih molekula iRNK za GSM3308814: x osa predstavlja broj izmerenih molekula iRNK, a y osa predstavlja broj ćelija (a) vidi se da najveći broj ćelija ima mali broj očitavanja (b) Uveličan deo grafika u opsegu od 0 do 10000 ; odabrana je donja granica za n_counts 1900 (c) Uveličan deo grafika u opsegu od 10000 do 70000 ; odabrana je gornja granica za n_counts 13000 (d) n_counts sa gornjom i donjom granicom. Vidi se sličnost između (a) na slici 2 i (a) na ovoj slici: prva slika je dobijena tako što se histogram sa ove slike zarotirao za 90 stepeni, duplirao kao u ogledalu i "uglačao".

Nakon filtriranja podataka korišćenjem pomenutih granica, broj ćelija sa velikim procentom mtRNK je smanjen (slika 5).

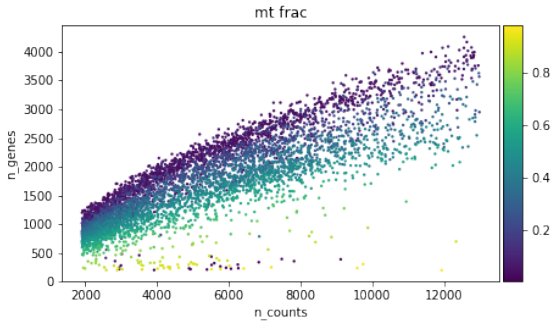

Figure 5: Značajno smanjen broj ćelija sa visokim udelom mtRNK

Na slici 6 su prikazani historgami broja izraženih gena u ćelijama:

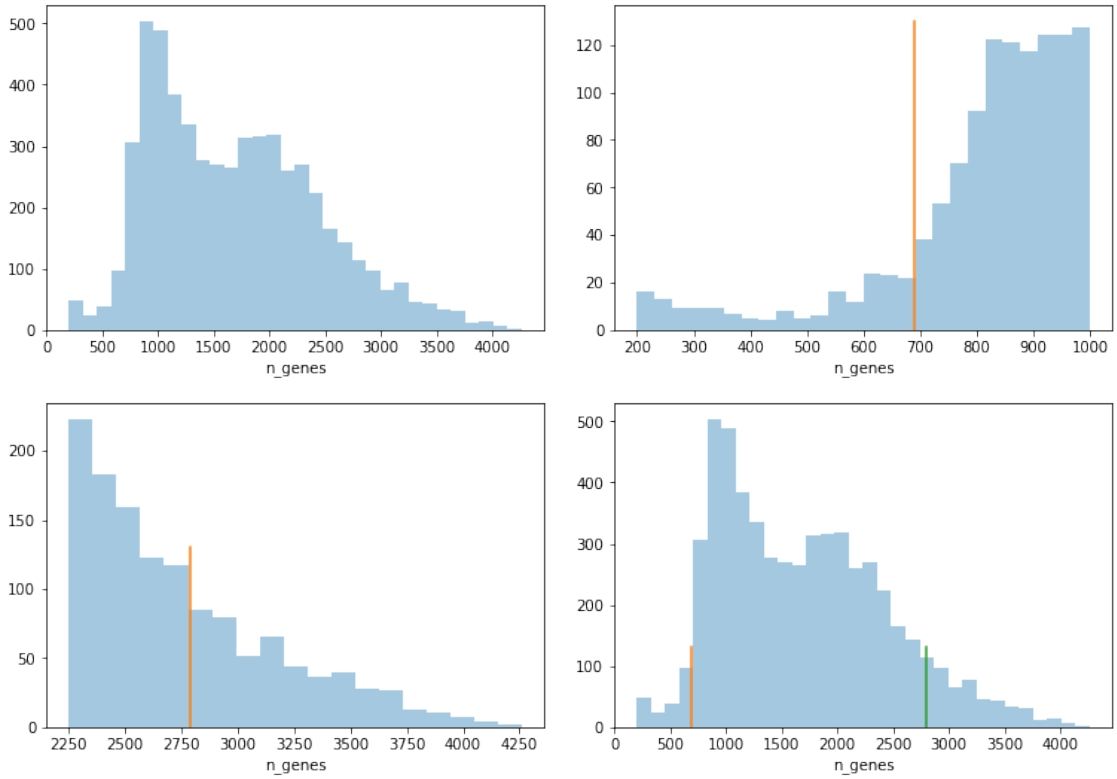

Figure 6: Raspodele broja izraženih gena za GSM3308814: na x osi je broj različitih gena izraženih po ćeliji, a na y osi je broj ćelija (a) vide se dve oblasti sa većom koncentracijom izraženih gena, oko 1000 i 2000 (b) Uveličan deo grafika u opsegu od 0 do 1000 ; odabrana je donja granica za n_genes 690 (c) Uveličan deo grafika u opsegu od 2250 do 4250 ; odabrana je gornja granica za n_genes 2790 (d) n_genes sa gornjom i donjom granicom

Posle filtriranja dobijeni su sledeći rezultati (slika 7):

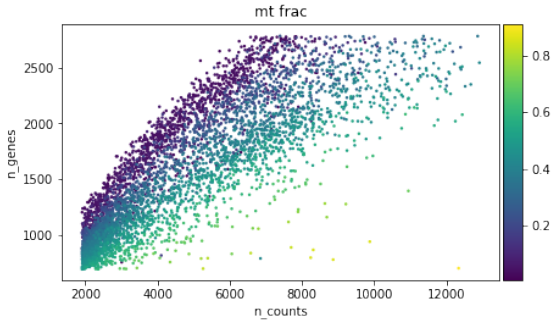

Figure 7: Broj ćelija sa visokim procentom mtRNK je dodatno smanjen

Podaci iz uzorka GSM3308815 analizirani su i obrađeni analogno.

Na slici (8) za uzorke druge grupe se vidi da ne postoji neka ravan odsecanja koja bi mogla da razdvoji podatke sa velikim i malim procentima mtRNK. Takođe, najveće vrednosti za ovu vrstu RNK su oko 50% za GSM3316206 i 35% za GSM3316207 (vidi sliku 8). Iz ovih razloga, izbačene su ćelije sa više od 20% nakon čega se broj ćelija smanjio sa 6053 na 5513 u GSM3316206 i sa 4222 na 4120 u GSM3316207 (slika 9).

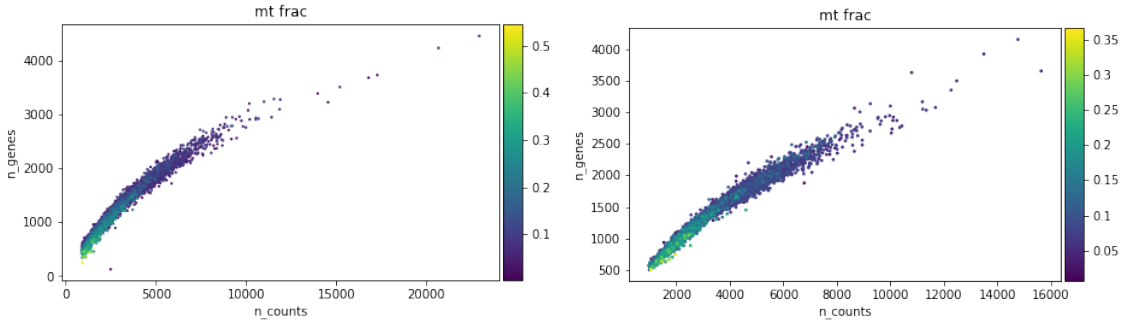

Figure 8: (a) GSM3316206 (b) GSM3316207

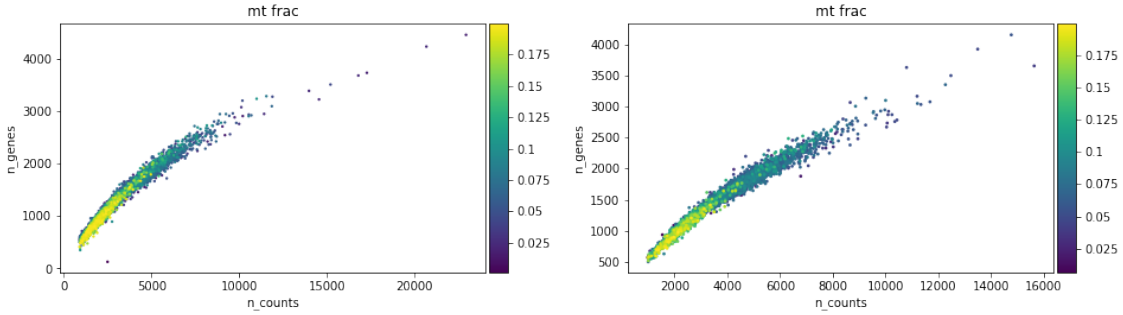

Figure 9: (a) GSM3316206 (b) GSM3316207

4.2.1 Normalizacija

Pošto sigurno nisu svi molekuli iRNK pokupljeni prilikom uzorkovanja, može se desiti da broj molekula varira između ćelija što može biti posledica samog uzorkovanja. Pošto se ne može pretpostaviti da je broj molekula iRNK u svim ćelijama isti, cilj je normalizacijom modifikovati vrednosti tako da dovoljno dobro procenjuju stvaran broj molekula u ćeliji. Ne procenjuje se sam broj molekula, već faktori u okviru svake ćelije koji bi trebalo da budu proporcionalni broju molekula ćelija - **faktori veličine**. Normalizacija se vrši tako što se izmerene vrednosti dele odgovarajućim faktorima.

U okviru paketa scanpy, vrši se CPM (counts per million) normalizacija: $CPM = \frac{\frac{r_i}{R}}{10^6}$, gde je R broj registrovanih

molekula u celom uzorku, a r_i su brojevi molekula za svaki atribut, rezultati se skaliraju sa milion zarad bolje čitljivosti. Nakon toga se podaci logaritmuju. Normalizacija je bitan korak u obradi podataka[2, 5] i zato je primenjena u ovom radu.

4.2.2 Vizualizacija

Podaci sa kojima se rade imaju ogromne dimenzije i nije moguće vizuelno ih predstaviti. Sa druge strane, veliki broj tih dimenzija nema nikakvu ulogu u istraživanju latentnih informacija. Stoga je neophodno smanjenje broja dimenzija podataka.

Tehnike za smanjenje dimenzionalnosti podataka korišćene u ovom radu su **PCA**, **tSNE** i **UMAP**.

PCA je tehnika smanjenja dimenzija podataka koja akcenat stavlja na varijabilnost, odnosno nastoji da predstavi samo one dimenzije koje pokrivaju najveću varijabilnost u okviru podataka. Početan visokodimenzionalan skup se linearnim transformacijama dovodi u skup sa manjim brojem dimenzija, odnosno glavnih komponenti koje su međusobno linearno nezavisne.

tSNE Za razliku od PCA, tSNE ne koristi linearnu transformaciju već se oslanja na lokalne odnose između podataka i stoga je sposobna da prepozna i nelinearne strukture. Uz pomoć normalne raspodele, pravi se raspodela u visokodimenzionom prostoru i nakon toga se uz pomoć studentove t-raspodele rekonstruišu ti odnosi u nižim dimenzijama. Više detalja u [6, 7]

UMAP predstavlja poboljšanje tSNE algoritma, odnosno brže se izvršava, bolje čuva globalnu strukturu podataka i takođe ima svrhu i van same vizualizacije pa se koristi i kao generalna tehnika smanjenja dimenzionalnosti podataka. U suštini, sastoji se iz dva koraka: 1) konstrukcija grafa povezanosti u visokodimenzionom prostoru i 2) optimizacija tog grafa koja traži najsličniju reprezentaciju u nižim dimenzijama. Više detalja može se naći u [8, 9]

Bitno je napomenuti nekoliko stvari u vezi tSNE i UMAP projekcija podataka. Naime, veličina klastera (površina koju klaster zauzima na slici, a ne broj samih tačaka u tim klasterima) na ovim projekcijama ne mora da znači gotovo ništa. Ovo je posledica toga što i tSNE i UMAP koriste lokalne udaljenosti za projektovanje u nižu dimenziju. tSNE algoritam će zato "širiti" guste klastere, a "sabijati" rasprsnute i na taj način će ih izjednačiti. Stoga, nije moguće uzimati u obzir relativne veličine klastera i iz njih izvlačiti ikakve zaključke. Još jedna stvar oko koje treba biti oprezan prilikom tumačenja ovih projekcija je da distance između dobro razdvojenih klastera često nemaju značenje što je još jedna posledica korišćenja lokalnih udaljenosti u algoritmu. Pošto su oba algoritma stohastička, odnosno mogu dati različite rezultate svaki put kada se izvrše, nekada je potrebno više puta ih pokrenuti zbog realnije slike o podacima.

Iako su ove dve tehnike dobre za prikaz visokodimenzionog prostora u manje dimenzija i dobro oslikavaju lokalnu povezanost tačaka u okviru istih klastera, treba biti svestan i navedenih zamki prilikom tumačenja grafika dobijenih ovim tehnikama.

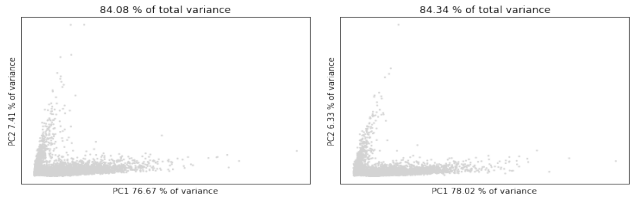

Figure 10: Dimenziono redukovani podaci na dve glavne komponente sa najvećom varijansom korišćenjem PCA dekompozicije: ose na ovom grafiku predstavljaju dva pravca u kojima je varijansa najveća. Ose su zapravo linearne kombinacije visokodimenzionog prostora. Ispod svake slike je oznacen procenat varijanse koji pokriva odredjena glavna komponenta, i zato su razlike u pravcu x ose bitnije od razlika u pravcu y ose. (a) GSM3308814: jasno se uočava da su podaci rasporedjeni u dva pravca (prve dve glavne komponente pokrivaju oko 84% varijanse u obe datoteke) (b) GSM3308815: podaci deluju gotovo identicno prvoj datoteci. Ova zapazanja nam govore da ce u daljoj analizi biti moguce smanjiti dimenzionalnost podataka bez velikih gubitaka informacija.

PCA nam očigledno ne pruža dovoljno informacija za zaključivanje o pripadnosti nekom klasteru, ali nam može koristiti kao smernica za izbor broja glavnih komponenti. Na osnovu grafika pokrivenosti varijanse u odnosu na broj glavnih komponenti, biće izabran broj komponenti za konstrukciju grafa suseda (slika 10).

Za UMAP i tSNE vizualizacije potrebna je konstrukcija grafa susedstva za date podatke. Grafovi su konstruisani metodom `sc.pp.neighbors` paketa `scanpy`. Za svaki skup podataka korisceno je po dve razlicite metrike: euklidsko i

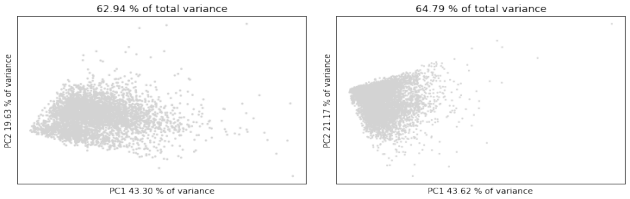

Figure 11: (a) GSM3316206: situacija je malo drugacija sto je i ocekivano s obzirom da prve dve komponente pokrivaju manji procenat varijanse. (b) GSM3316207: sve tačke su veoma slične. U ovom slucaju, potrebno nam je vise glavnih komponenti u odnosu na prvu grupu za istu pokrivenost varijanse.

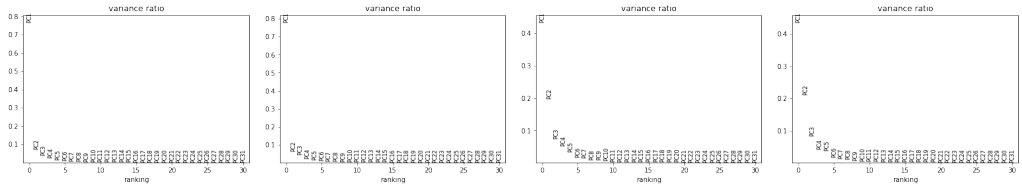

Figure 12: Prikaz odnosa broja komponenti i procenta pokrivenosti varijanse. Često se ovaj prikaz koristi za odabir broja glavnih komponenti, ali ovde će biti korišćen drugačiji pristup. Naime, biće izabran dovoljan broj glavnih komponenti da se pokrije određen procenat varijanse.

kosinusno. Broj glavnih komponenti odabran je tako da je pokriveno barem 90% varijanse u svakom skupu (vidi sliku 12).

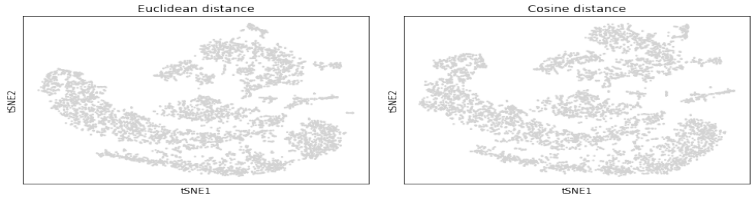

Figure 13: GSM3308814 (a) tSNE vizualizacija koriscenjem euklidskih rastojanja: sa slike se moze zakljuciti da postoji preko 5 klastera, ali treba uzeti u obzir i navedene mere opreza, odnosno suzdrzati se od tumacenja razdvojenosti i oblika klastera. Ova tehnika nam pruza preliminarni uvid u podatke i imace vise smisla kada bude primenjeno klasterovanje. (b) kosinusna rastojanja: I ovde se vidi preko pet klastera malo drugacijeg oblika. Ono sto se moze videti sa slike je reprezentacija topologije podataka u dvodimenzionom prostoru.

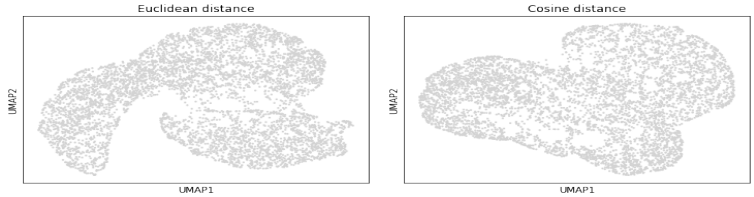

Figure 14: GSM3308814 (a) UMAP vizualizacija koriscenjem euklidskih rastojanja: jasno je razdvojeno dva klastera, ali treba uzeti u obzir i navedene mere opreza, odnosno suzdrzati se od tumacenja razdvojenosti grafova. Ova tehnika nam pruza preliminarni uvid u podatke i imace vise smisla kada bude primenjeno klasterovanje. (b) kosinusna rastojanja: podaci deluju malo slicnije i sa ove slike se ne moze puno zakljuciti, bice lakse tumaciti rezultate nakon klasterovanja. Ono sto se moze videti sa slike je reprezentacija topologije podataka u dvodimenzionom prostoru. Takodje se moze videti da su podaci manje razdvojeni u odnosu na tSNE vizualizaciju, ali i tu treba biti oprezan jer je za UMAP poznato da bolje cuva globalnu topologiju.

Primećuje se da tSNE grafici uglavnom daju veliki broj klastera sa delimično nejasnim granicama, dok se UMAP vizualizacije ponašaju relativno bolje. UMAP algoritam se brže izvršava i bolje oslikava topologiju podataka. Vizualizacije

neklasterovanih podataka su u dodatku**??**. U nastavku ce biti koriscena UMAP vizualizacija, a tSNE vizualizacije mogu se dobiti pokretanjem metoda iz skripta clustering.

5 Klasterovanje

Za klasterovanje će biti korišćeni algoritmi louvain i leiden implementirani u okviru scanpy paketa. Takođe će biti izvršeni algoritmi za klasterovanje prisutni u scikit-learn paketu. Kod koji je korišćen za dobijanje navedenih rezultata nalazi se u skripti cluster. Rezultati klasterovanja sa ocenama kvaliteta i ćelijama koje pripadaju svakom klasteru dati su u okviru foldera clustering_results.

5.0.1 Louvain i Leiden

Louvain[10, 11] je algoritam hijerarhijskog klasterovanja koji otkriva klastere u mrežama tako što maksimizuje nivo modularnosti svakog klastera. To znači da se procenjuje koliko su jako povezani čvorovi nekog klastera u odnosu na to koliko bi bili povezani u nasumičnoj mreži. Rekurzivno se spajaju klasteri i primenjuje se maksimizacija modularnosti na spojenim klasterima. Matematička formulacija modularnosti:

$$H = \frac{1}{2m}\sum_{c}(e_c - \gamma\frac{K_c^2}{2m})$$

Gde je K_c suma stepeni čvorova u klasteru c, m je ukupan broj grana u mreži, e_c je broj grana u klasteru c i γ je rezolucija. Veća rezolucija dovodi do većeg broja klastera a manja do manjeg.

Leiden[11] algoritam predstavlja poboljšanje louvain algoritma u kome se može desiti da neki čvor koji je imao ulogu spone između klastera uđe u sastav nekog novog klastera i na taj način napravi nepovezane klastere. Ovo se rešava uvođenjem dodatnog međukoraka prečišćavanja klastera. U tom koraku se čvorovi ne pridružuju nužno onom klasteru za koji je modularnost najveća već se nasumično bira neki od klastera za koji se povećava ciljna funkcija.

U nastavku ce biti prikazani najbolji (po koeficijentu senke) rezultati za ova dva algoritma klasterovanja po datotekama. Algoritmi su pokrenuti za interval rezolucija od 1.0 do 0.15. Ovaj parametar odredjuje koliko blizu tacke trebaju biti medjusobno kako bi se smestile u isti klaster. Vece vrednosti daju vise klastera, odnosno lokalna slicnost je naglasenija, dok manje vrednosti ovog parametra daju manji broj klastera sa vise tacaka.

U sledecoj tabeli dati su najbolji rezultati klasterovanja u odnosu na koeficijent senke i Davies-Bouldin indeks.

	GSM3308814		**GSM3308815**		**GSM3316206**		**GSM3316206**	
	Distance used		Distance used		Distance used		Distance used	
	Euclidean	Cosine	Euclidean	Cosine	Euclidean	Cosine	Euclidean	Cosine
Silhouette	0.3446	0.3933	0.3723	0.4737	0.3853	0.3109	0.4015	0.3695
Calinski-Harabasz	7202.4210	7892.7844	7390.1634	9046.2576	3368.4585	3186.1564	5616.0195	2916.7540
Davies-Bouldin	0.9388	0.7874	0.6779	0.7431	0.8893	0.8316	0.7750	1.0421
Resolution	0.3	0.15	0.15	0.15	0.15	0.5	0.3	0.15
Clusters found	8	5	6	4	3	7	5	3

Table 1: Rezultati klasterovanja koriscenjem louvan algoritma: iz tabele se vidi da je najbolji rezultat (2 od 3 najbolje ocene) klasterovanje datoteke GSM3308815 za rezoluciju 0.15. Ovaj rezultat nije toliko ocigledan na slici15 sto je i ocekivano s obzirom da su brojcane vrednosti ocena veoma bliske. Generalno se vidi da se za prvu grupu (GSM3308814 i GSM3308815) kosinusna rastojanja bolje ponasaju od euklidskih dok je za druge dve datoteke situacija suprotna. Iako postoje razlike izmedju izbora metrika rastojanja, ne deluje da su toliko znacajne.

Rezolucije klasterovanja je izabrane su eksperimentalno. Zbog nepoznatih stvarnih klasa uzoraka, kvalitet pronađenih klastera je procenjen merama koje ne zahtevaju oznake stvarnih klasa. Korišćeni su: Calinski-Harabasz[12] indeks, Davies-Bouldin[13] indeks i koeficijent senke.

5.0.2 KMeans klasterovanje

Ideja algoritma je da pokuša da rasporedi podatke u k klastera tako da rastojanje među podacima u okviru jednog klastera bude minimalno. Od korisnika očekuje da unese željeni broj klastera. Rad algoritma je testiran na intervalu od 2 do 10 klastera. Dobro se skalira i na više dimenzije. Funkcija koja se minimizuje definisana je kao

$$\sum_{i=0}^{n}\min_{\mu_j \in C}(||x_i - \mu_j||^2)$$

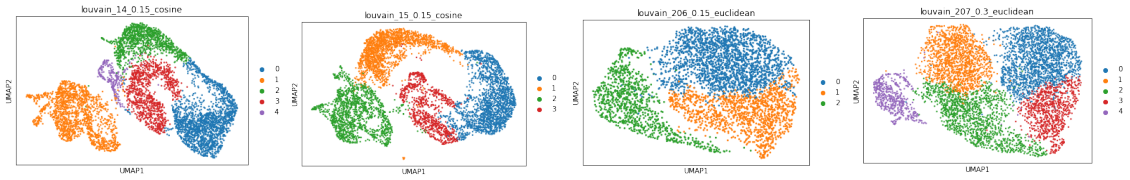

Figure 15: UMAP reprezentacija rezultata najboljih klasterovanja za algoritam Louvain po datotekama. U prvoj grupi (a) i (b) deluje da je algoritam odradio dosta dobar posao i da je na logican nacin razdvojio kalstere. Kao sto je vec spomenuto, velicina ovih klastera ne mora da ima bilo kakvo znacenje, bitno je da su tacke u okviru klastera medjusobno slicne. Takodje udaljenost klastera nema neku posebnu vaznost. Sa slika se moze videti topoloska organizacija podataka projektovana na 2 dimenzije. Na slici pod (a) vidimo da su klasteri "prosarani" i tackama koje mozda logicno ne bi trebalo smestiti u njih, sto je potvrdjeno i koeficijentom senke koji je relativno mali. Na slici (b) se takodje vide neke tacke koje odstupaju. Slicna situacija je i sa slikama (c) i (d) s tim sto deluje da je algoritam pronasao vise klastera nego sto bi se golim okom procenilo, odnosno na slici (c) bi se klasteri 0 i 1 mogli spojiti u jedan veci na osnovu slike, a na slici (d) bi se klasteri 1 i 2 mogli spojiti u jedan, a klasteri 0 i 3 spojiti u drugi klaster na taj nacin formirajuci samo tri klastera. Povecan prostor pretrage za parametar rezolucije bi mozda proizveo bolje rezultate, ali u eksperimentima nije doslo do poboljsanja i ovo su najbolji rezultati nakon vise pokusaja. U nastavku ce takodje biti dati najbolji dobijeni rezultati.

	GSM3308814		**GSM3308815**		**GSM3316206**		**GSM3316206**	
	Distance used		Distance used		Distance used		Distance used	
	Euclidean	Cosine	Euclidean	Cosine	Euclidean	Cosine	Euclidean	Cosine
Silhouette	0.3561	0.4199	0.4045	0.3968	0.3363	0.3421	0.4054	0.3830
Calinski-Harabasz	6646.1356	7531.1001	7754.5354	6914.3054	3271.5613	3092.7325	4770.8381	3512.9175
Davies-Bouldin	0.8271	1.3735	0.7408	0.9940	0.9888	0.8927	0.7530	0.8721
Resolution	0.15	0.15	0.15	0.3	0.5	0.3	0.15	0.15
Clusters found	5	5	6	8	7	5	4	4

Table 2: Rezultati klasterovanja koriscenjem leiden algoritma: iz tabele se vidi da je najbolji rezultat (2 od 3 najbolje ocene) klasterovanje datoteke GSM3308815 za rezoluciju 0.15 i euklidsko rastojanje. S druge strane, klasterovanje za datoteku GSM3308814 i kosinusno rastojanje daje bolju ocenu senke, tako da nije najjasnije koji klaster je bolji. Zanimljivo je da ovde nema pravilnosti vezanih za izbor ocene rastojanja i kvaliteta klasterovanja prvu datoteku je bolje kosinusno, za drugu euklidsko, za trecu kosinusno i za cetvrtu euklidsko. Ni sada rezultati nisu toliko uocljivi na slici **??**

Gde je μ_i srednja vrednost podataka u klasteru i a x_i su sami podaci. Zbog formulacije ove formule, KMeans algoritam ne radi dobro za izdužene ili podatke nepravilnog oblika. Vizuelizacija najboljih rezultata (po koeficijentu senke) na slici 17.

	GSM3308814		GSM3308815		GSM3316207		GSM3316207	
	Euclidean	Cosine	Euclidean	Cosine	Euclidean	Cosine	Euclidean	Cosine
Silhouette	0.4059	0.3583	0.3719	0.3932	0.3020	0.3325	0.3600	0.3293
Calinski-Harabasz	5390.8242	3518.0090	3283.3920	4471.3486	1824.6711	2308.4630	3969.1302	3338.0132
Davies-Bouldin	0.7037	0.7942	0.8642	0.7426	1.1715	1.0531	0.8804	0.9125
k	4	2	2	2	2	2	3	3

Table 3: Rezultati klasterovanja algoritmom k-sredina: najbolje se pokazalo klasterovanje datoteke GSM3308814 i to sa k = 4, i uz koriscenje euklidskog rastojanja. Vrednosti ocena najboljih klastera su slicne onima dobijenim i za louvain i leiden klasterovanja. Rezultati su vizuelno predstavljeni na slici 17

5.0.3 Ward klasterovanje

Ovde će biti prikazani rezultati hijerarhijskog klasterovanja korišćenjem ward-ove veze. Opšta ideja algoritama hijerarhijskog klasterovanja je da svaki podatak u početku predstavlja jedan klaster. Iteraciju po iteraciju, klasteri se na osnovu sličnosti spajaju u veće klastere sve dok svi podaci ne upadnu u isti klaster. U okviru modula scikit-learn programskog jezika python postoji metod za ovu vrstu klasterovanja u kojoj je moguće postaviti traženi broj klastera. Kako bi se lakše poredili različiti algoritmi, i ovde je izabran interval od 2 do 10 klastera. Wardova veza minimizuje sumu kvadriranih razlika između svih klastera i zbog toga što se na taj način minimizuje varijansa, slična je KMeans algoritmu. Vizuelizacija rezultata je na slici **??**.

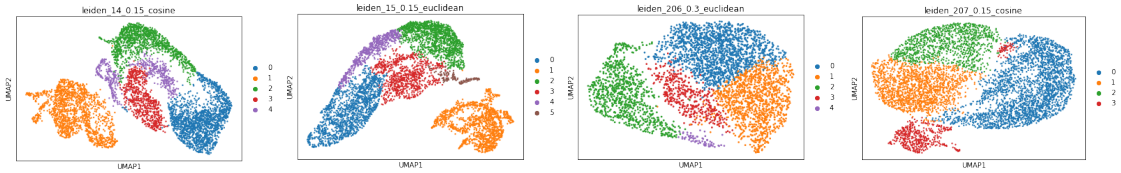

Figure 16: UMAP reprezentacija rezultata najboljih klasterovanja za algoritam leiden po datotekama: Slicna je situacija kao i kod louvain algoritma, na prvi pogled podaci deluju smisleno organizovano u klastere s tim da su na nekim mestima prosarani tackama koje mozda ne pripadaju datoj okolini. Na primer, na slici (a) se vidi da je klaster 4 dosta razvucen i da su tacke koje mu pripadaju dosta udaljene jedne od drugih. Takodje se moze primetiti mala grupa tacaka klastera 0 u donjem levom uglu klastera 3 sto je takodje zanimljivo. Na slici (b) postoji malo manje ovakvih anomalija i klasteri su generalno dobro razdvojeni i moze se reci da bi se i rucno slicno izdvojili klasteri. Na slici (d) izgleda kao da je algoritam pronasao vise klastera nego sto je ih je vizuelno, ali opet treba naglasiti da se kod tumacenja UMAP reprezentacije rastojanje izmedju klastera treba uzeti sa rezervom. Na slici (d) vidimo crvene tacke koje "upadaju" u region koji uglavnom zauzimaju plave tacke. U poredjenju sa louvain algoritmom, (a) ima bolji koeficijent senke, a gore Calinski-Harabasz i Davies-Bouldin indekse. Prilikom poredjenja slika (a) takodje se moze primetiti da se ljubicasti klasteri razlikuju, odnosno da je louvain algoritam pronasao malo logicniju podelu. Slike pod (b) i (d) se ne mogu porediti jer je za vizualizaciju koriscena drugacija metrika u ovim algoritmima, ali mozemo im uporediti ocene i zakljuciti da je leiden algoritam za euklidsko rastojanje u datotedci GSM3308815 nasao malo bolje klastere u odnosu na pokretanje louvain algoritma za iste parametre. S druge strane, koriscenje kosinusnog rastojanja dovelo je do znatno veceg kvaliteta klastera koriscenjem louvain algoritma. Uz to, louvain klasterovanje je pronaslo 4 klastera sto je blize broju od 5 ocekivanih klastera u podacima[1]. Za klasterovanje predstavljeno slikom (d) dobijeni su slicni rezultati za euklidsko rastojanje koriscenjem oba algoritma. Leiden se pokazao malo bolje kada je koristio kosinusna rastojanja. Kada se uiporede slike za ova dva algoritma i datoteku GSM3316207 deluje da su klasteri lepse razdvojeni koriscenjem louvain algoritma i euklidskih rastojanja sto je i potvrdjeno unosima u tabeli.

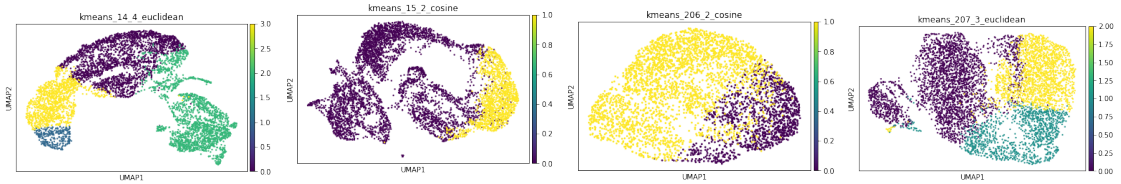

Figure 17: Vizualizacija rezultata dobijenih algoritmom k sredina: najbolje klasterovanje je dobijeno na slici (a) ali se jasno vidi nedostatak ovog algoritma jer se cini da su klasteri podeljeni gotovo linearno i uprkos solidnim ocenama iz 3, ne deluje da je podela toliko logicna. Za preostale datoteke, pogotovo na (b) i (c) se uocava isti problem i tu je razlika u odnosu na prethodne algoritme primetna i po ocenama kvaliteta. Uocljive su i tacke jednog klastera koje zalaze u druge, pogotovo na slici (d).

	GSM3308814		**GSM3308815**		**GSM3316206**		**GSM3316206**	
	Euclidean	Cosine	Euclidean	Cosine	Euclidean	Cosine	Euclidean	Cosine
Silhouette score	0.5193	0.3428	0.3849	0.4543	0.3265	0.2970	0.3673	0.3403
Calinski-Harabasz score	8750.7358	3184.5519	3552.9775	6254.4627	2530.7055	1798.1102	3896.1915	3542.3782
Davies-Bouldin score	0.6254	0.7954	0.8672	0.7270	0.8706	1.2140	0.8751	0.9513
Clusters	3	2	2	2	3	2	3	3

Table 4: Rezultati klasterovanja koriscenjem hijerarhijskog klasterovanja i ward-ove veze:

5.0.4 Birch klasterovanje

Ovaj algoritam klasterovanja od podataka pravi drvo odlika klasterovanja (eng. Clustering Feature Tree) koje zapravo predstavlja vid kompresije podataka sa gubitkom. Čvorovi ovog drveta sadrže podklastere koji mogu imati podklastere itd. Ovi podklasteri u sebi sadrže potrebne podatke za klasterovanje i na taj način eliminišu potrebu za čuvanje celog skupa podataka u memoriji. Parametri ovog algoritma su granica (eng. threshold) i faktor grananja. Granica određuje dovoljnu udaljenost nesvrstanog podatka do postojećeg klastera, a faktor grananja ograničava broj podklastera u drvetu. U metodi iz scikit-learn modula moguće je podesiti i broj klastera. Kao vrednosti u ovom radu uzeta je kombinacija granica 0.3,0.5, faktora grananja 50 i broja klastera 5 i 10. Vizuelizacija rezultata je na slici 18.

	GSM3308814		GSM3308815		GSM3316206		GSM3316206	
	Euclidean	Cosine	Euclidean	Cosine	Euclidean	Cosine	Euclidean	Cosine
Silhouette	0.2337	0.1053	0.1050	0.0350	0.2488	0.1592	0.2979	0.2578
Calinski-Harabasz	4041.6852	2717.8585	4031.4666	1950.8816	2098.3719	1812.8360	3483.6280	3078.4488
Davies-Bouldin	0.9788	2.8007	1.0325	4.5868	0.9741	1.8790	1.0443	1.0730
Threshold	0.3	0.5	0.3	0.3	0.3	0.3	0.3	0.3
Branching factor	50	50	50	50	50	50	50	50
Clusters	10	10	10	10	5	5	5	5

Table 5: Rezultati birch algoritma klasterovanja:

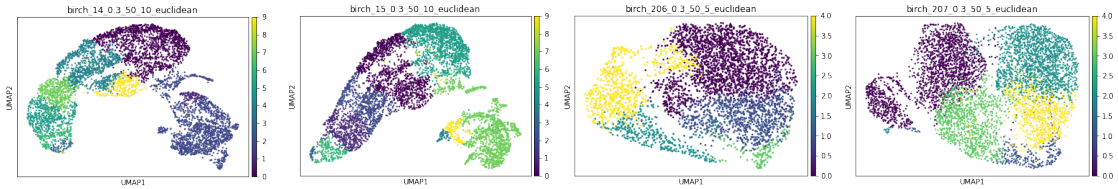

Figure 18: Vizuelizacija rezultata dobijenih birhc algoritmom klasterovanja:

5.1 Zaključak

Iako je korišćeno više metoda klasterovanja, na kraju su dobijeni veoma slični rezultati.

Vizualizacije

References

[1] Razin Abdulrauf Shaikh, Jiahui ZHONG, Minjie LYU, Sen LIN, Derin KESKIN, Guanglan ZHANG, Lou CHITKUSHEV, and Vladimir BRUSIC. Classification of five cell types from pbmc samples using single cell transcriptomics and artificial neural networks. *arXiv preprint arXiv:1804.09028*, 2019.

[2] Malte D Luecken and Fabian J Theis. Current best practices in single-cell rna-seq analysis: a tutorial. *Molecular systems biology*, 15(6), 2019.

[3] Tomislav Ilicic, Jong Kyoung Kim, Aleksandra A Kolodziejczyk, Frederik Otzen Bagger, Davis James McCarthy, John C Marioni, and Sarah A Teichmann. Classification of low quality cells from single-cell rna-seq data. *Genome biology*, 17(1):29, 2016.

[4] Single cell tutorial. `https://github.com/theislab/single-cell-tutorial/blob/master/latest_notebook/Case-study_Mouse-intestinal-epithelium_1906.ipynb`. Accessed: 2020-02-14.

[5] Jiawei Long and Yu Xia. Cluster analysis of high-dimensional scrna sequencing data. *arXiv preprint arXiv:1912.08400*, 2019.

[6] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

[7] Martin Wattenberg, Fernanda Viégas, and Ian Johnson. How to use t-sne effectively. *Distill*, 2016.

[8] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[9] Understanding umap. `https://pair-code.github.io/understanding-umap/`. Accessed: 2020-02-22.

[10] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008, 2008.

[11] Vincent A Traag, Ludo Waltman, and Nees Jan van Eck. From louvain to leiden: guaranteeing well-connected communities. *Scientific reports*, 9(1):1–12, 2019.

[12] Tadeusz Caliński and Harabasz JA. A dendrite method for cluster analysis. *Communications in Statistics - Theory and Methods*, 3:1–27, 01 1974.

[13] David L Davies and Donald W Bouldin. A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence*, (2):224–227, 1979.