Sistem za obradu, skladištenje i pretraživanje jednoćelijskih RNK sekvenciranih podataka (eng. *Single Cell RNASeq*)

Ksenija Ćesarević

Sadržaj — Rad predstavlja razvoj integrisane platforme za upravljanje jednoćelijskim RNK sekvencijskim podacima. Platforma kombinuje automatizovane Nextflow radne tokove, relacionu bazu podataka za metapodatke i RAG arhitekturu za kontekstualno pretraživanje. Implementiran Python API i veb interfejs omogućavaju efikasno dodavanje studija i praćenje obrada. Evaluacija potvrđuje značajno poboljšanje dostupnosti jednoćelijskih genomskih podataka.

Ključne reči — automatizacija bioinformatike, genomska bioinformatika, jednoćelijska sekvencija RNK, Nextflow, Python API, RAG sistem pretrage, scRNA-seq, upravljanje genomskim podacima, veštačka inteligencija u biologiji

I. Uvod

Savremena medicina sve više se oslanja na razumevanje bioloških procesa na ćelijskom i molekularnom nivou [1]. Eksperimentalni biolozi i kliničari nastoje da istraže i modulišu ponašanje ćelija primenom ciljanih molekularnih strategija. U cilju postizanja dubljeg molekularnog uvida u ćelijsku funkciju, sprovode se različite vrste analiza koje obuhvataju ispitivanje genomske DNK (eng. *Deoxyribonucleic Acid*), strukture hromatina, sekvenci glasničke RNK (eng. *Messenger RNA*, mRNA), nekodirajućih RNK, ekspresije i posttranslacionih modifikacija, kao i prisustva specifičnih metabolita.

Ksenija Ćesarević, Autor, Računarski fakultet, Srbija (telefon: 381-60-0242714; email: kcesarevic @raf.rs

Tradicionalno su se ove analize vršile na heterogenim populacijama koje obuhvataju hiljade do milijarde ćelija, usled veoma male količine molekula u pojedinačnoj živoj ćeliji. Ćelijska heterogenost predstavlja fundamentalnu karakteristiku bioloških sistema, pri čemu se postavlja pitanje da li uočene razlike između ćelija zaista imaju biološki značaj [2]. Takav pristup omogućio je značajna saznanja u molekularnoj biologiji, posebno kroz studije asocijacije na nivou genoma kojima su identifikovani pojedinačni nukleotidni polimorfizmi (eng. *Single Nucleotide Polymorphisms*, SNP) povezani sa određenim biološkim karakteristikama i sklonostima ka bolestima kod ljudi.

A. Analiza ekspresije na ćelijskom nivou i transkriptomika

Za razumevanje ćelijskih odgovora neophodno je vršiti analizu ekspresije gena ili proteina. U studijama ekspresije proteina, primena višebojne protočne citometrije (eng. *multicolor flow cytometry*) u kombinaciji sa fluorescentno obeleženim monoklonskim antitelima omogućila je istovremeno ispitivanje manjeg broja proteina na velikom broju pojedinačnih ćelija, što je postalo uobičajena praksa u eksperimentalnim i kliničkim istraživanjima.

U novije vreme, razvoj masene citometrije (eng. *mass cytometry*), koja podrazumeva bojenje ćelija antitelima obeleženim jonima teških metala i kvantitativno merenje uz pomoć detektora vremena leta (eng. *time-of-flight detector*), omogućio je višestruko (pet do deset puta) povećanje broja proteina koji se mogu istovremeno analizirati [3]. Ova tehnologija je doprinela otkrivanju do tada nedovoljno uočene heterogenosti i složenosti unutar naizgled homogenih populacija ćelija, poput imunih ćelija. Ipak, i dalje predstavlja značajan izazov istovremeno ispitivanje kompletnog skupa proteina (takozvani proteom) koje genom kodira i koji se eksprimiraju u jednoj jedinstvenoj ćeliji. Kao posredna metoda za proučavanje proteoma, mnogi istraživači se oslanjaju na analizu molekula glasničke RNK koji kodiraju proteine, a koji kolektivno čine transkriptom. Ekspresija transkriptoma često dobro korelira sa osobinama ćelije i promenama u njenom fiziološkom stanju.

U početku, transkriptomika se sprovodila na skupovima od više miliona ćelija, prvo korišćenjem mikronizova zasnovanih na hibridizaciji (eng. *microarrays*), a kasnije primenom tehnika sekvenciranja nove generacije (eng. *next-generation sequencing*, NGS), poznatih kao RNK sekvenciranje [4]. Analize RNK sekvenciranja na spojenim (eng. *bulk*) ćelijama omogućila je dobijanje obimne količine podataka, koje i danas predstavljaju osnovu za brojna otkrića i inovacije u oblasti biomedicine.

Od objavljivanja prve studije sekvenciranja glasničke RNK na nivou pojedinačne ćelije (eng. *single-cell RNA sequencing*, scRNA-seq) 2009. godine [5], interesovanje za ovu tehnologiju kontinuirano raste. Jedan od najubedljivijih razloga za primenu ove metode leži u mogućnosti da se ekspresija RNK molekula analizira sa visokom rezolucijom i na nivou pojedinačnih ćelija, u obimu celog genoma [1]. Iako su studije u početku sprovodile uglavnom visoko specijalizovane istraživačke grupe, postaje sve očiglednije da i istraživači u biomedicinskim naukama, kao i kliničari, mogu ostvariti značajna otkrića korišćenjem ove moćne metode, naročito jer su tehnologije i alati potrebni za sprovođenje takvih analiza postali sve dostupniji.

Analize razlika u transkripciji između pojedinačnih ćelija omogućile su identifikaciju retkih ćelijskih populacija koje bi inače ostale neprimećene u analizama spojenih ćelija — na primer, maligne tumorske ćelije unutar mase tumora ili hiperreaktivne imune ćelije u okviru naizgled homogene populacije. Dodatno, ova tehnologija se sve češće koristi za praćenje ćelijskih loza i razvojnih odnosa u različitim biološkim kontekstima, uključujući embrionalni razvoj, kancer, diferencijaciju mioblasta i epitela pluća, kao i sudbinu limfocita [1].

II. GENETIČKA OSNOVA ĆELIJSKE FUNKCIJE I KOMPLEKSNOST ORGANIZMA

Svi biološki procesi u organizmima – od razvoja tkiva do odgovora na spoljašnje nadražaje – potiču iz informacija sačuvanih u molekulu DNK (eng. *deoxyribonucleic acid*). DNK se sastoji od četiri azotne baze (adenin - A, timin - T, citozin - C, guanin - G) čiji raspored čini genetički kod. Ovaj kod sadrži uputstva za sintezu RNK (eng. *ribonucleic acid*), a zatim i proteina, koji predstavljaju izvršioce gotovo svih funkcija unutar ćelije [6]. Geni su osnovne funkcionalne jedinice DNK i nose informacije o strukturi i ekspresiji proteina specifičnih za različite tipove ćelija.

A. Centralna dogma molekularne biologije i diferencijalna genska ekspresija

Proces prevođenja genetičke informacije u funkcionalne proteine odvija se kroz centralnu dogmu molekularne biologije: **transkripciju** (sintezu RNK) I **translaciju** (sintezu proteina).

Tokom transkripcije, enzim RNK polimeraza II (eng. *RNA polymerase II*) prepoznaje promotorske sekvence i sintetiše glasničku RNK (eng. *messenger RNA*, mRNA) iz DNK, pri čemu kod eukariotskih organizama novosintetisana

mRNA prolazi kroz sazrevanje koje uključuje dodavanje 5'-kape (eng. *5'-cap*) i 3'-poli(A) repa (eng. *3'-poly(A) tail*), kao i uklanjanje introna i spajanje egzona. Alternativno splajsovanje (eng. *alternative splicing*) omogućava stvaranje različitih izoformi proteina iz jednog gena [6].

Iako sve ćelije organizma poseduju identičan genetički materijal, dramatično se razlikuju po morfologiji, funkciji i ponašanju zbog diferencijalne genske ekspresije - različiti tipovi ćelija aktiviraju različite setove gena u različitim vremenskim trenutcima i obimu. Regulacija genske ekspresije je složen proces koji se odvija na različitim nivoima: od kontrole transkripcije putem transkripcijskih faktora i epigenetičkih modifikacija, preko post-transkripcijske regulacije pomoću nekodirajućih RNK molekula, do post-translacijskih modifikacija proteina [2]. Ova precizna kontrola omogućava ćelijama prilagođavanje promenljivim uslovima i održavanje specifične funkcije. Diferencijalna genska ekspresija predstavlja osnovu ćelijske diferencijacije tokom razvoja, tkivne specijalizacije, odgovora na signale i metaboličkih adaptacija.

Zato je analiza genske ekspresije na nivou pojedinačne ćelije postala nezamenljiv alat za dublje razumevanje bioloških procesa i razvoj novih terapijskih pristupa. Razvoj tehnologija kao što su sekvenciranje RNK (eng. *RNA sequencing*, RNA-Seq) [4] i sekvenciranje RNK na nivou pojedinačne ćelije (eng. *single-cell RNA sequencing*, scRNA-seq) [1,5] omogućio je preciznu analizu transkripcijskih profila na nivou pojedinačnih ćelija i otkrio do tada skrivene obrasce genske regulacije i ćelijske složenosti [2]..

III. Proces sekvenciranja RNK na nivou pojedinačne ćelije

Uspešno sprovođenje eksperimenata sekvenciranja RNK pojedinačne ćelije počinje pažljivim planiranjem i pripremom uzoraka, pri čemu kvalitet finalnih podataka direktno zavisi od kvaliteta početnog materijala i održavanja integriteta RNK molekula [7,8]. Precizna koordinacija između eksperimentalnih i bioinformatičkih pristupa predstavlja ključni element uspešne analize, jer greške u eksperimentalnoj fazi mogu biti ograničeno korigovane tokom kasnije bioinformatičke obrade.

Metodologija se oslanja na tri primarna tehnološka pristupa koji se razlikuju po principu izolacije ćelija, dubini sekvenciranja i protoku uzoraka [8]. Metoda zasnovana na kapljicama (eng. *Droplet-based method*), predstavljena platformama kao što je 10x Genomics Chromium, omogućava masovnu

analizu hiljada ćelija istovremeno kroz enkapsulaciju pojedinačnih ćelija u nanolitarske kapljice sa jedinstvenim molekularnim barkodovima, što rezultuje značajno nižom cenom po ćeliji ali ograničava analizu na 3' krajeve transkripata [9,10].

Priprema ćelijskog uzorka zahteva pažljivo planiranje pri čemu ćelijska suspenzija mora biti pripremljena tako da minimizuje stres ćelija i artifakte u genskoj ekspresiji nastale tokom manipulacije [8,10]. Temperatura, osmolarnost i pH vrednost moraju biti precizno kontrolisani, standardno se koristi temperatura od 4°C do 37°C, osmolarnost se održava na fiziološkim nivoima (280-320 mOsm/kg), dok se pH kontroliše puferskim sistemima. Disocijacija predstavlja proces dezintegracije tkivne arhitekture kroz narušavanje intercelularnih veza, što može biti mehaničko (fizičke sile pogodne za meka tkiva) ili enzimsko (specifične proteaze za ciljano razgradanje ekstracelularne matrice) [7]. Nakon disocijacije, ćelijska suspenzija se podvrgava rigoroznoj kontroli kvaliteta kroz filtriranje i procenu viabilnosti, držeći se na niskim temperaturama i koristi u najkraćem mogućem vremenskom periodu.

Metodologija zasnovana na kapljicama predstavlja naprednu mikrofluidnu tehnologiju koja omogućava masovnu enkapsulaciju i obradu pojedinačnih ćelija kroz formiranje diskretnih reakcijskih komora u vidu kapljica emulzije ulje-u-vodi [9-10]. Ova tehnologija revolucionisala je polje genomike pojedinačnih ćelija omogućavajući istovremenu analizu hiljada do desetina hiljada ćelija sa dramatično smanjenim troškovima po ćeliji, pri čemu svaka kapljica nanolitarske zapremine funkcioniše kao nezavisan bioreaktor koji enkapsulira jednu ćeliju zajedno sa svim neophodnim reagensima za lizu, reverznu transkripciju i molekularno barkodovanje [9].

Finalni skup podataka, nakon kompletne kontrole kvaliteta i filtriranja, predstavlja osnovu za naknadne bioinformatičke analize koje mogu obuhvatiti identifikaciju tipova ćelija, analizu razvojnih putanja, uporedne studije između različitih uslova i rekonstrukciju genskih regulatornih mreža [7-8]. Kvalitet ovog finalnog skupa podataka direktno određuje moć i pouzdanost svih naknadnih bioloških zaključaka.

IV. MAPIRANJE SEKVENCI NA REFERENTNI GENOM I KVANTIFIKACIJA EKSPRESIJE GENA

Tokom procesa visokoprotočnog sekvenciranja, biološke informacije kodovane u DNK sekvencama prolaze kroz složenu transformaciju od fizičkih

signala do digitalno dostupnih podataka pogodnih za bioinformatičku analizu [13-14]. Ovaj proces predstavlja kritičnu poveznicu između eksperimentalne metodologije i računarske obrade podataka. Rezultat primarne obrade optičkih signala su BCL fajlovi (eng. Base Call files) koji predstavljaju binarni format specifičan za Illumina platforme [13].

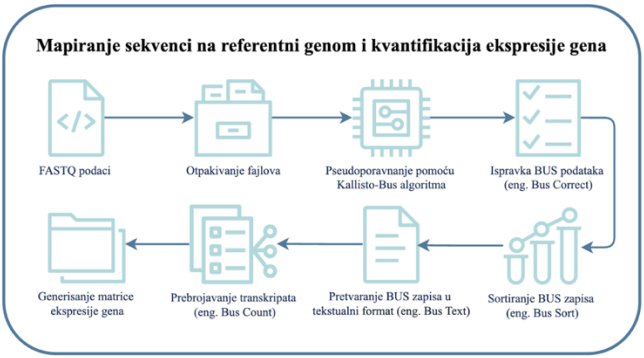
Transformacija BCL fajlova u FASTQ format predstavlja ključni korak koji omogućava interoperabilnost između sekvenciranja platformi i bioinformatičkih alata [13-14]. Ova konverzija se izvršava koristeći specijalizovane softvere kao što je bcl2fastq (Illumina) ili alternativi kao što su bcl-convert ili picard-tools. FASTQ format predstavlja tekstualni standard za skladištenje nukleotidnih sekvenci zajedno sa odgovarajućim ocenama kvaliteta, što ga čini univerzalno prihvaćenim formatom u bioinformatici [14].

Kod jednoćelijski sekvenciranih RNK eksperimenata, konverzija u FASTQ format mora da uzme u obzir složenu strukturu očitavanja koji sadrže brojne funkcionalne elemente [1,17]. Tipična konfiguracija za metode zasnovane na kapljicama generiše tri odvojena FASTQ fajla:

- Očitavanje 1 (eng. ***Read 1***) koji sadrži ćelijske barkodove i UMI sekvence
- Očitavanje 2 (eng. ***Read 2***) koji sadrži genske sekvence
- Indeksno očitavanje (eng. ***Index read***) koji sadrži uzorčeve barkodove za demultipleksovanje

Ova organizacija podataka omogućava efikasnu obradu velikih količina informacija uz očuvanje svih funkcionalnih elemenata potrebnih za naknadnu analizu (eng. downstream analysis) [17]. Specijalizovani bioinformatički paketi kao što su Cell Ranger, STARsolo, alevin-fry ili kallisto-bustools direktno procesiraju FASTQ fajlove kako bi generisali kvantifikacije genskih ekspresija na nivou pojedinačnih ćelija [15-17].

Završni rezultat ove transformacije je matrica ekspresije gena gde očitavanja predstavljaju gene, kolone predstavljaju ćelije, a vrednosti predstavljaju kvantifikovane nivoe ekspresije [1,8]. Ova matrica, često izvožena u formatima kao što su CSV, HDF5, ili specijalizovani objekti kao što su Seurat (R, Python) ili AnnData (Python), predstavlja osnovu za sve naknadne bioinformatičke analize uključujući identifikaciju tipova ćelija, analizu razvojnih putanja i uporedne genomske studije [17].



DIJAGRAM 1. PROCES MAPIRANJA SEKVENCI NA REFERENTNI GENOM I GENERIANJE MATRICE EKSPRESIJE GENA

V. BIOINFORMATIČKA ANALIZA JEDNOĆELIJSKIH RNK PODATAKA

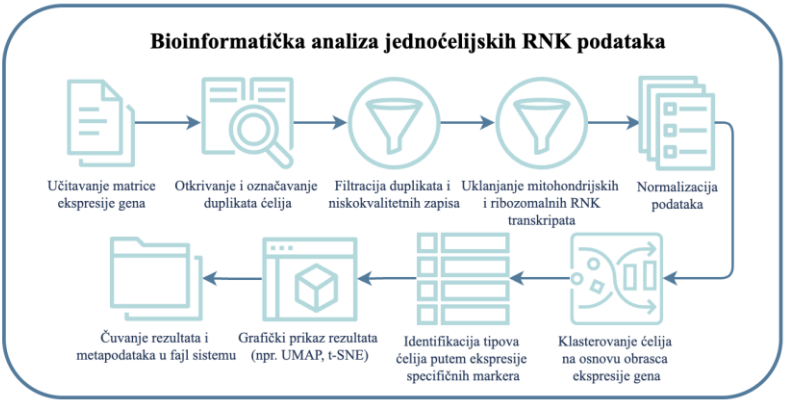
Bioinformatička analiza jednoćelijskih RNK podataka predstavlja složen, višestepeni proces koji transformiše sirove sekvencijalne podatke u biološki značajne uvide o ćelijskoj heterogenosti, funkciji i razvojnim putanjama [1,2]. Ova analiza zahteva primenu specijalizovanih algoritma i metodologija koje su prilagođene jedinstvenim karakteristikama jednoćelijskih podataka, uključujući visoku dimenzionalnost, izraženu retko-popunjenost matrice (eng. sparsity) i značajnu tehničku varijabilnost [5,8,17].

Analitički tok (eng. workflow) podataka obuhvata više kritičnih uzastopnih koraka, od kojih svaki doprinosi kvalitetu i pouzdanosti finalnih rezultata [8,17]:

- **Kontrola kvaliteta ćelija (eng. Quality Control, QC)** - Identifikacija i uklanjanje ćelija sa niskim kvalitetom podataka, mrtvih ćelija ili duplikata koji mogu kompromitovati naknadne analize [7,8]. Ovaj korak uključuje evaluaciju metrika kao što su ukupan broj detektovanih gena po ćeliji, ukupan broj UMI, i procenat mitohondrijalnih gena [9,10].
- **Kvantifikacija transkripata** - Precizno određivanje nivoa ekspresije svakog gena u pojedinačnim ćelijama na osnovu mapiranih sekvencijskih očitavanja, uz korekciju za tehnička ograničenja kao što su PCR duplikacija i sekvencijska dubina [15,16]. Ovaj proces se

oslanja na napredne probabilističke algoritme za tačno kvantifikovanje transkripta u uslovima izražene retko-popunjenosti podataka [15].

- **Normalizacija podataka** - Standardizacija ekspresijskih vrednosti radi eliminisanja sistematskih varijacija koje potiču od tehničkih faktora, omogućavajući komparaciju između ćelija i uzoraka [1,8,17]. Ovo uključuje korekciju za razlike u ukupnom broju detektovanih molekula po ćeliji i kompenzaciju za različite efikasnosti sekvenciranja [4].
- **Uklanjanje faktora koji dovode do zabuna** - Identifikacija i eliminacija dubleta (ćelija koje enkapsuliraju više od jedne ćelije) i drugih artefakata koji mogu dovesti do pogrešnih bioloških zaključaka [8,17]. Ovaj korak je kritičan za održavanje integriteta podataka, posebno u protokolima visokog protoka kao što su droplet-based metode [9,10].
- **Redukcija dimenzionalnosti i selekcija karakteristika** - Identifikacija najinformativnijih gena (eng. *highly variable genes*) i primena matematičkih tehnika kao što su PCA (eng. *Principal Component Analysis*) i UMAP (eng. *Uniform Manifold Approximation and Projection*) za vizualizaciju i analizu podataka u nižedimenzionalnom prostoru [1,8,17]. Ove tehnike omogućavaju efikasno rukovanje velikim brojem dimenzija karakterističnih za jednoćelijske podatke.
- **Klaster analiza** - Grupisanje ćelija sa sličnim transkriptomskim profilima u diskretne populacije koje verovatno predstavljaju različite ćelijske tipove ili funkcionalna stanja [2,8,17]. Moderna pristupa koriste grafne algoritme i mašinsko učenje za identifikaciju biološki značajnih ćelijskih subpopulacija.
- **Naknadne analize (eng. *Downstream analyses*)** - Složene analize koje uključuju identifikaciju marker gena za različite ćelijske populacije, rekonstrukciju razvojnih putanja (eng. trajectory analysis), funkcionalne analize genskih setova, i komparativne studije između različitih eksperimentalnih uslova [8,17,18]. Ove analize omogućavaju dublje razumevanje ćelijske funkcije i regulatornih mreža na jednoćelijskom nivou.



DIJAGRAM 2. TOK BIOINFORMATIČKE ANALIZE JEDNOĆELIJSKOG RNK SEKVENCIRANJA

Svaki od ovih koraka zahteva pažljivo razmatranje parametara analize, validaciju rezultata i integraciju sa biološkim znanjem radi interpretacije dobijenih podataka [17,18]. Uspešna primena ove metodologije omogućava dublje razumevanje ćelijske biologije na nivou koji je bio nedostupan tradicionalnim pristupima analize genskih ekspresija [1,6].

VI. ARHITEKTURA SOFTVERSKOG REŠENJA ZA ANALIZU I SKLADIŠTENJE PODATAKA DOBIJENIH IZ EKSPERIMENATA JEDNOĆELIJSKE RNK

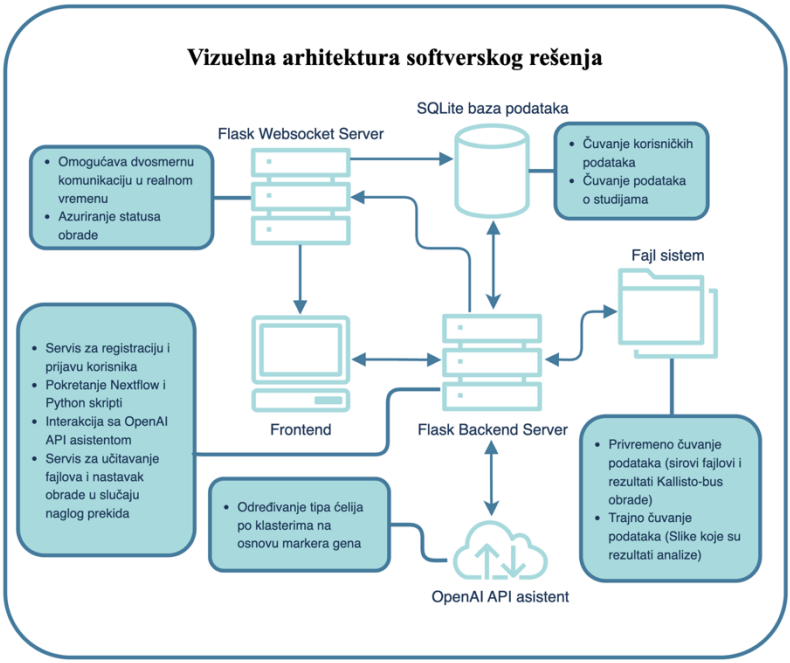
Nakon što je detaljno prikazana biološka osnova istraživanja i metodologija analize podataka dobijenih jednoćelijskim RNK sekvenciranjem [1,5,8], naredni segment rada usmeren je ka prikazu arhitekture razvijenog softverskog sistema. Cilj ovog sistema predstavlja sveobuhvatna automatizacija, standardizacija i integracija svih kritičnih faza analitičkog procesa – počevši od obrade sirovih sekvencijskih podataka [12-14], preko sistematičnog upravljanja metapodacima i eksperimentalnim parametrima, pa sve do izvođenja kompleksnih analitičkih procedura [15-17] i interaktivne vizualizacije dobijenih rezultata [18-19].

Razvijena aplikacija implementira napredni modularni pristup koji omogućava međusobnu povezanost ključnih komponenti sistema kroz jasno definisane interfejse i protokole komunikacije. Arhitektura obuhvata četiri fundamentalna sloja: **sloj bioinformatičke obrade podataka** odgovoran za

implementaciju algoritma analize jednoćelijski RNK sekvenciranih podataka [1,8,17], **sloj skladištenja i upravljanja bazama podataka** koji osigurava perzistentnost i integritet eksperimentalnih podataka [17], **sloj izvođenja analitičkih algoritama** koji omogućava skalabilno izvršavanje računski zahtevnih procedura [18-21], kao i **sloj prikaza rezultata** koji kroz sofisticiran korisnički interfejs omogućava intuitivnu interpretaciju i eksploraciju rezultata [18-19].

Ovakav pristup arhitekture obezbeđuje nekoliko ključnih prednosti u odnosu na tradicionalne sisteme za bioinformatičku analizu:

- **Efikasnost i performanse** - Modularni dizajn omogućava optimizaciju svakog sloja nezavisno, paralelizaciju obrade podataka i efikasno upravljanje resursima sistema [16-17].
- **Reproduktibilnost rezultata** - Standardizovani protokoli obrade i verzioniranje analitičkih procedura garantuju konzistentnost rezultata kroz različite eksperimentalne sesije i omogućavaju validaciju naučnih nalaza [8,17,18].
- **Skalabilnost sistema** - Arhitektura podržava obradu podataka različitih veličina, od malih pilot studija do velikoskalnih populacijskih istraživanja [9-10], uz mogućnost horizontalnog skaliranja na oblačnoj platformi (eng. *cloud platforms*) [16,19].
- **Proširivost funkcionalnosti** - Jasno definisani interfejsi (eng. interfaces) omogućavaju integraciju novih analitičkih metoda, algoritma i vizualizacionih komponenti bez modifikacije postojećeg koda [17-19].
- **Interoperabilnost** - Sistem podržava standardne bioinformatičke formate podataka i protokole [12,17], omogućavajući integraciju sa postojećim softverskim rešenjima i bazama podataka [18-19].



DIJAGRAM 3. ARHITEKTURA SOFTVERSKOG REŠENJA ZA ANALIZU I SKLADIŠTENJE PODATAKA DOBIJENIH IZ EKSPERIMENATA JEDNOĆELIJSKE RNK

A. Korisnički interfejs aplikacije (eng. Frontend)

Korisnički interfejs predstavlja primarnu komponentu analitičke platforme za obradu podataka jednoćelijskog RNK sekvenciranja, implementiran kao nezavisan frontend modul prema principima savremenog korisničkog iskustva [18,19]. Sistem omogućava upravljanje korisničkim nalozima kroz kompletan životni ciklus korisničkih sesija, uključujući registraciju, autentifikaciju sa sigurnim protokolima i enkripcijom podataka [21,22]. Bezbednosni aspekti obuhvataju višeslojnu zaštitu sa validacijom korisničkih unosa, CSRF zaštitom za kritične operacije i ograničavanjem pristupa resursima na osnovu korisničkih privilegija [20,22].

Jedna od najkompleksnijih funkcionalnosti je sistem za učitavanje velikih bioinformatičkih FASTQ fajlova koji mogu dosegnuti veličine od nekoliko gigabajta [1,17]. Sistem implementira mehanizam nastavka prekidanja (eng. *resumable upload*) koji deli velike fajlove na manje segmente koji se nezavisno šalju na server. U slučaju prekida konekcije, sistem automatski identifikuje koje segmente treba ponovno poslati, omogućavajući nastavak procesa bez ponovnog slanja celokupnog fajla. Tokom slanja, sistem prati brzinu transfera, procenjuje preostalo vreme i validira integritet podataka kroz algoritme sumiranja [22].

Praćenje statusa dugotrajnih analitičkih procesa realizuje se kroz WebSocket komunikacijski protokol koji omogućava bidirektionalnu, asinhronu komunikaciju između korisničkog interfejsa i pozadinskog dela aplikacije [17]. Korisnici mogu da prate napredovanje analiza kroz vizuelne indikatore koji prikazuju trenutnu fazu obrade, procenjeno vreme završetka i detaljne log poruke. WebSocket infrastruktura omogućava trenutno obaveštavanje o kritičnim događajima, uključujući završetak analitičkih faza ili potrebu za dodatnim korisničkim inputom, bez potrebe za ručnim osveženjem stranice.

Frontend implementira sveobuhvatan sistem za vizualizaciju bioinformatičkih rezultata koji omogućava intuitivnu eksploraciju kompleksnih setova podataka [18,19]. Modul uključuje dinamičke dijagrame rasipanja za UMAP/t-SNE vizualizacije, toplotne mape za prikaz ekspresijskih profila gena i grafike oblika violine za komparativnu analizu ćelijskih populacija [19]. Svi grafički elementi podržavaju uvećavanje, pomeranje i funkcionalnosti selekcije za detaljno istraživanje podataka. Rezultati diferencijalne analize prikazuju se kroz sortirajuće tabele sa mogućnostima filtriranja, a sistem omogućava izvoz rezultata u standardnim formatima prilagođenim naučnim publikacijama [17,18].

B. Sistem za upravljanje bazom podataka

Rešenje za perzistenciju podataka zasnovano je na SQLite sistemu koji predstavlja optimalno rešenje uzimajući u obzir zahteve za lakoću implementacije, prenosivost i robusnost za predviđeno opterećenje sistema [22]. Struktura baze podataka organizovana je kroz ključne entitete koji odražavaju logičke odnose u domenu bioinformatičke analize [17]. Entitet korisnika čuva informacije o registrovanim korisnicima uključujući jedinstvene identifikatore, korisničke adrese elektronske pošte i kriptografski zaštićene lozinke sa implementiranim mehanizmima indeksiranja za brzu validaciju kredencijala [20,22].

Centralni entitet sesije za učitavanje podataka sadrži metapodatke o sesijama za učitavanje i obradu bioinformatičkih podataka, skladišti informacije o nazivu datoteke, ukupnoj veličini, trenutnom statusu obrade i detaljima o biološkom uzorku poput imena studije, broja uzorka, tipa uzorka i zdravstvenog stanja [1,17]. Rezultati analize čuvaju se kroz komplementarne entitete koji skladište međurezultate i konačne rezultate bioinformatičkih analiza sa referencama ka datotekama sa rezultatima ekspresijskih profila, identifikovanim klasterima ćelija i oznakama ćelijskih tipova [18,19]. Za interakciju sa bazom podataka koristi se SQLAlchemy objektno-relaciono preslikavanje što omogućava održavanje visokog nivoa apstrakcije pri manipulaciji podacima i nezavisnost aplikacione logike od specifičnosti konkretnog sistema za upravljanje bazom podataka [22].

Implementiran je mehanizam praćenja učitavanja po segmentima koji omogućava nastavak prenosa u slučaju prekida mrežne veze, dok parametri analize skladište se u zasebnim entitetima omogućavajući reproduktibilnost i ponovljivost naučnih rezultata [17,22]. Ovakva arhitektura omogućava efikasno skladištenje i pristup kritičnim informacijama neophodnim za funkcionisanje celokupnog sistema, uz istovremeno očuvanje integriteta podataka i optimizaciju performansi pri paralelnom pristupu većeg broja korisnika [20,22].

C. Flask pozadinski deo aplikacije (eng. Backend)

Centralnu logiku i obradu unutar aplikacije realizuje Flask backend server [23], koji predstavlja jezgro sistema zaduženo za upravljanje pravilima, pokretanje analitičkih procesa i koordinaciju između različitih komponenti [17]. Ova komponenta implementirana je korišćenjem Python Flask okvira, poznatog po svojoj fleksibilnosti i jednostavnosti integracije sa različitim servisima i bibliotekama za bioinformatičku analizu [18]-[19]. Flask je minimalistički Python web-okvir (eng. *microframework*) koji omogućava efikasnu izgradnju RESTful API-ja i web-servisa pri minimalnom umetnutom sloju apstrakcije [23]. Za razliku od sveobuhvatnijih monolitnih rešenja, Flask pruža programeru potpunu kontrolu nad izborom i integracijom biblioteka za upravljanje bazom podataka [23], autentifikaciju [22]-[23], validaciju i druge infrastrukturne komponente, dok arhitektura aplikacije i struktura fajl sistema ostaju potpuno prilagodljivi specifičnim zahtevima projekta [23].

Backend server odgovoran je za nekoliko kritičnih aspekata funkcionisanja sistema:

- **Registracija i autentifikacija korisnika** - Implementiran je sveobuhvatan sistem za upravljanje korisničkim nalozima koji obuhvata sigurne mehanizme registracije, autentifikacije i autorizacije [22,23]. Korisnički podaci se čuvaju u kriptografski zaštićenom obliku, dok se sesijske informacije upravljaju kroz sigurne tokene sa ograničenim vremenom važenja [23].
- **Pokretanje i koordinacija analitičkih procesa** - Server orkestrira izvršavanje automatizovanih toka obrade podataka pomoću Nextflow alata [24] za sirove FASTQ fajlove [12] i generisanje matrica ekspresije [15]-[16]. Paralelno sa tim, upravlja Python analizama koje uključuju filtriranje, normalizaciju, klasterovanje i generisanje kompleksnih vizualizacija rezultata [18]-[21].
- **Integracija sa spoljašnjim servisima** - Implementirana je sofisticirana integracija sa OpenAI API asistentom u cilju automatske identifikacije ćelijskih tipova na osnovu izraženih markera u identifikovanim klasterima [1,8]. Ova funkcionalnost predstavlja naprednu implementaciju veštačke inteligencije u domensko-specifičnim bioinformatičkim analizama [17].
- **Upravljanje fajlovima i resursima** - Server implementira napredne mehanizme za upravljanje velikim bioinformatičkim fajlovima koje korisnici učitavaju putem frontend-a [12]. Ovo uključuje validaciju fajlova [12,14], segmentovano učitavanje sa mogućnošću nastavka prekidanja, kao i optimizovano skladištenje i pristup podacima [23].
- **Komunikacija u realnom vremenu** - Flask backend služi kao posrednik između frontend komponente i ostalih servisa, prosleđujući rezultate analize i ažurirajući stanje obrada u realnom vremenu pomoću WebSocket komunikacije [21]. Na ovaj način se postiže visoka modularnost sistema i jasno razdvajanje odgovornosti između slojeva aplikacije [17, 23].

D. Integracija sistema veštačke inteligencije za automatksu identifikaciju ćelijskih tipova

Jedan od najnaprednijih aspekata razvijene aplikacije predstavlja integraciju OpenAI GPT-4o modela za automatsku identifikaciju i kategorizaciju ćelijskih tipova na osnovu ekspresionih profila marker gena [1, 8]. Ova komponenta kombinuje tradicionalne bioinformatičke pristupe [17]-[18] sa savremenim metodama veštačke inteligencije, omogućavajući automatizovanu interpretaciju kompleksnih bioloških podataka koja inače zahteva visok nivo domenske ekspertize.

Arhitektura AI komponente dizajnirana je oko OpenAI Assistant API-ja koji omogućava kreiranje specijalizovanih asistenata sa predefinisanim instrukcijama i domensko-specifičnim znanjem. Za razliku od direktnog pristupa kroz standardni Chat API, Assistant API pruža mogućnost kreiranja persistentnih kontekstualnih sesija sa specijalizovanim ponašanjem [25], što značajno poboljšava konzistentnost i tačnost interpretacije biomedicinskih podataka.

Centralna funkcionalnost ovog sistema zasniva se na sposobnosti modela da interpretira liste marker gena karakteristične za određene ćelijske klastere [2, 6] i da na osnovu biološkog znanja identifikuje najverojatnije ćelijske tipove. Model koristi opširnu bazu znanja o genskim ekspresionim profilima različitih ćelijskih tipova [7], uključujući imune ćelije, epitelijalne ćelije, mezenhimalne ćelije i druge specijalizovane ćelijske populacije karakteristične za single-cell RNA-seq analize [19]-[20].

VII. AUTOMATIZOVANI TOK OBRADE POMOĆU NEXTFLOW ALATA

Nextflow predstavlja naprednu platformu za dizajn, implementaciju i upravljanje kompleksnim bioinformatičkim tokovima rada koji zahtevaju visoku reproduktivnost, skalabilnost i fleksibilnost izvršavanja [24]. Razvijen specifično za potrebe savremene bioinformatike, Nextflow omogućava kreiranje tokova izrade koji mogu efikasno rukovati velikim količinama biomedicinskih podataka u heterogenim računarskim okruženjima [17].

Kako bi se obezbedila reproduktivnost i skalabilnost obrade, svi koraci procesa – od preuzimanja podataka [12], preko kvantifikacije i sortiranja [15]-[16], do generisanja ekspresione matrice [1, 5] – automatizovani su korišćenjem alata Nextflow [24]. Nextflow je fleksibilan okvir za definisanje i upravljanje bioinformatičkim tokovima rada (eng. *pipelines*), koji omogućava paralelizaciju izvršavanja, jednostavno rukovanje ulazno-izlaznim fajlovima, kao i lako ponavljanje analiza u različitim računarskim okruženjima [26]. Zahvaljujući deklarativnom pristupu i modularnoj strukturi, Nextflow omogućava transparentno praćenje svakog koraka analize, što značajno doprinosi transparentnosti i reproduktivnosti istraživanja [23].

Nextflow pipeline-i su organizovani kroz modularne komponente koje se mogu nezavisno razvijati, testirati i ponovno koristiti kroz različite projekte [24]. Svaki modul enkapsulira specifičnu funkcionalnost i definiše jasne interfejse za komunikaciju sa ostalim komponentama, što omogućava lako komponovanje složenih analitičkih tokova karakterističnih za single-cell RNA-seq analize [17]-[18].

Modularni pristup omogućava kreiranje biblioteka ponovno upotrebljivih komponenti koji mogu biti podeljeni kroz naučnu zajednicu [23]. Ovo značajno poboljšava kolaboraciju između istraživača i omogućava brže usvojavanje najboljih praksi u bioinformatičkim analizama [8, 17].

VIII. SETOVI PODATAKA IZ OBLASTI KANCER ISTRAŽIVANJA

Implementirani softverski sistem za analizu jednoćelijskih RNK sekvenciranih podataka testiran je i validiran korišćenjem reprezentativnih setova podataka iz oblasti kancer istraživanja, što predstavlja jedan od najkompleksnijih domena bioinformatičke analize [2, 8]. Za potrebe validacije, odabran je pristup testiranja na javno dostupnim, standardizovanim setovima podataka koji omogućavaju objektivno poređenje performansi sistema sa postojećim rešenjima u oblasti [1, 17].

Kompanija 10x Genomics predstavlja jedan od vodećih proizvođača tehnologije za jednoćelijsko sekvenciranje i kao takva obezbeđuje javno dostupne referentne setove podataka koji služe kao zlatni standard u naučnoj zajednici [9, 10]. Ovi podaci su generisani korišćenjem analize ekspresije gena na nivou pojedinačnih ćelija pomoću Chromium platforme (eng. *Chromium Single Cell Gene Expression*, 10x Genomics), koja omogućava precizno izdvajanje i analizu RNK iz individualnih ćelija [1, 5].

Kompanija 10x Genomics čini svoje referentne podatke javno dostupnim kroz svoju web platformu, što omogućava naučnoj zajednici da koristi iste podatke za razvoj i validaciju novih metoda [9, 10]. Ova otvorena politika podataka doprinosi transparentnosti i reproducibilnosti istraživanja u oblasti jednoćelijskog sekvenciranja [23], što je kritično za razvoj i validaciju bioinformatičkih alata [17]-[18].

Kao primarni testni dataset odabran je "3k PBMCs from a Healthy Donor" koji sadrži 3000 perifernih mononuklearnih ćelija krvi (PBMC) ekstraktovanih iz zdravog donora [9, 10]. Tehničke karakteristike ovog seta podataka čine ga idealnim za validaciju bioinformatičkih alata [17, 18]:

- **Visoki kvalitet sekvenciranja**: Prosečan broj detektovanih gena po ćeliji preko 1000, što omogućava preciznu karakterizaciju ćelijskih tipova [1, 14]
- **Standardizovana priprema**: Korišćena je 10x Genomics Chromium tehnologija koja obezbeđuje konzistentne rezultate kroz različite laboratorije i eksperimente [9, 10]

- **Dobro okarakterisane populacije**: Ćelijski tipovi su prethodno validovani kroz citometriju protoka i druge standardne metode [3, 7], što omogućava preciznu evaluaciju tačnosti automatske identifikacije [8]
- **Kompletna dokumentacija**: Uz podatke se obezbeđuje detaljna dokumentacija o eksperimentalnim uslovima, protokolima pripreme i parametrima sekvenciranja [9]

IX. ZAKLJUČAK

Ovaj master rad predstavlja sveobuhvatan pristup razvoju i implementaciji naprednog softverskog sistema za analizu podataka dobijenih jednoćelijskim RNK sekvenciranjem, sa posebnim fokusom na primenu u onkološkim istraživanjima. Istraživanje je uspešno integrisalo teoretske osnove molekularne biologije kancera sa praktičnim aspektima savremenog softverskog inženjerstva.

Razvijena je modularna softverska arhitektura koja integriše četiri fundamentalna sloja: bioinformatičku obradu podataka, optimizovano skladištenje, analitičke algoritme i intuitivni korisnički interfejs. Ovakav pristup obezbeđuje skalabilnost, reproduktibilnost i proširivost, što predstavlja značajan napredak u odnosu na konvencionalne bioinformatičke platforme. Implementirane napredne funkcionalnosti, uključujući resumable upload protokol za velike FASTQ datoteke, WebSocket komunikaciju za praćenje dugotrajnih analitičkih procesa u realnom vremenu, kao i sofisticiran sistem multidimenzionalne vizualizacije rezultata, čine platformu pristupačnom širokom spektru korisnika - od bioinformatičara do kliničara sa ograničenim tehničkim iskustvom.

Modularna arhitektura sistema postavlja čvrstu osnovu za buduću integraciju algoritma mašinskog učenja i veštačke inteligencije u analizu jednoćelijskih podataka. Fleksibilnost platforme omogućava inkrementalno dodavanje novih analitičkih komponenti, što je čini adaptivnom za tehnološke inovacije u oblasti scRNA-seq metodologija. Implementacija cloud-native tehnologija i horizontalnog skaliranja pozicionira sistem kao potencijalno rešenje za velika populacijska istraživanja kancera, što može značajno ubrzati translaciju naučnih otkrića u kliničku praksu kroz omogućavanje analize velikog broja pacijenata.

Ovaj rad predstavlja uspešnu sintezu biološkog razumevanja malignoma, savremenih eksperimentalnih metoda jednoćelijskog sekvenciranja i

naprednog softverskog inženjerstva. Rezultujući sistem ne samo da automatizuje analizu kompleksnih genomskih podataka, već aktivno katalitizuje istraživanja u oblasti precizne onkologije kroz omogućavanje novih naučnih pristupa i metodologija. Razvijeni sistem predstavlja značajan korak napred u digitalizaciji onkoloških istraživanja i postavlja temelje za buduće inovacije u oblasti personalizovane medicine zasnovane na jednoćelijskoj genomici.

X. LITERATURA

[1] Kolodziejczyk, A. A., Kim, J. K., Svensson, V., Marioni, J. C., & Teichmann, S. A. (2015). The technology and biology of single-cell RNA sequencing. *Molecular Cell*, 58(4), 610-620.

[2] Altschuler, S. J., & Wu, L. F. (2010). Cellular heterogeneity: do differences make a difference? *Cell*, 141(4), 559-563.

[3] Bandura, D. R., Baranov, V. I., Ornatsky, O. I., Antonov, A., et al. (2009). Mass cytometry: technique for real time single cell multitarget immunoassay based on inductively coupled plasma time-of-flight mass spectrometry. *Analytical Chemistry*, 81(16), 6813-6822.

[4] Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1), 57-63.

[5] Tang, F., Barbacioru, C., Wang, Y., Nordman, E., et al. (2009). mRNA-Seq whole-transcriptome analysis of a single cell. *Nature Methods*, 6(5), 377-382.

[6] Alberts, B., Johnson, A., Lewis, J., Morgan, D., Raff, M., Roberts, K., & Walter, P. (2014). *Molecular Biology of the Cell* (6th ed.). Garland Science.

[7] Proserpio, V. (2019). *Single Cell Methods - Sequencing and Proteomics*. Springer Nature.

[8] Haque, A., Engel, J., Teichmann, S. A., & Lönnberg, T. (2017). A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. *Genome Medicine*, 9(1), 75.

[9] Macosko, E. Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., ... & McCarroll, S. A. (2015). Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, 161(5), 1202-1214.

[10] Klein, A. M., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., ... & Kirschner, M. W. (2015). Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*, 161(5), 1187-1201.

[11] Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., et al. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 456(7218), 53-59.

[12] Cock, P. J., Fields, C. J., Goto, N., Goto, N., Heuer, M. L., & Rice, P. M. (2010). The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Research*, 38(6), 1767-1771.

[13] Illumina Inc. (2017). BCL Convert v3.7.5 Software Guide. Document # 1000000163594. Illumina, Inc.

[14] Schirmer, M., D'Amore, R., Ijaz, U. Z., Hall, N., & Quince, C. (2016). Illumina error profiles: resolving fine-scale variation in metagenomic sequencing data. *BMC Bioinformatics*, 17(1), 125.

[15] Bray, N. L., Pimentel, H., Melsted, P., & Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology*, 34(5), 525-527.

[16] Melsted, P., Booeshaghi, A. S., Liu, L., Gao, F., Lu, L., Min, K. H., ... & Pachter, L. (2021). Modular, efficient and constant-memory single-cell RNA-seq preprocessing. *Nature Biotechnology*, 39(7), 813-818.

[17] Yuan, G. C. (Ed.). (2019). *Computational Methods for Single-Cell Data Analysis*. Humana Press.

[18] Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck III, W. M., ... & Satija, R. (2019). Comprehensive integration of single-cell data. *Nature Biotechnology*, 37(4), 423-43

[19] Wolf, F. A., Angerer, P., & Theis, F. J. (2018). SCANPY: large-scale single-cell gene expression data analysis. *Genome Biology*, 19(1), 15.

[20] OWASP Foundation. (2021). OWASP Top Ten 2021: The Ten Most Critical Web Application Security Risks. Open Web Application Security Project.

[21] Jones, M., Bradley, J., & Sakimura, N. (2015). JSON Web Token (JWT). RFC 7519. Internet Engineering Task Force.

[22] Howard, M., & LeBlanc, D. (2003). Writing Secure Code (2nd ed.). Microsoft Press.

[23] Grinberg, M. (2018). Flask Web Development: Developing Web Applications with Python (2nd ed.). O'Reilly Media.

[24] Di Tommaso, P., Chatzou, M., Floden, E. W., Barja, P. P., Palumbo, E., & Notredame, C. (2017). Nextflow enables reproducible computational workflows. Nature Biotechnology, 35(4), 316-319.

[25] Fette, I., & Melnikov, A. (2011). The WebSocket Protocol. RFC 6455. Internet Engineering Task Force.

[26] Grigorik, I. (2013). High Performance Browser Networking: What every web developer should know about networking and web performance. O'Reilly Media.

ABSTRACT

SYSTEM FOR PROCESSING, STORING AND SEARCHING SINGLE-CELL RNA SEQUENCING DATA

Ksenija Ćesarević

Single-Cell RNA Sequencing (scRNA-seq) technology enables gene expression analysis at individual cell resolution, revealing cellular heterogeneity and developmental pathways, but generates massive, complex datasets that pose significant bioinformatics challenges for processing, storage, and analysis. This master's thesis presents an integrated platform that addresses these challenges by combining automated data processing through Nextflow workflows with advanced search and analysis systems, including centralized metadata storage in optimized relational databases and intelligent data searching using both traditional indexing and artificial intelligence approaches.

The system's key innovation is a RAG (Retrieval-Augmented Generation) architecture that enables natural language-based contextual searching and semantic analysis of scientific studies, integrated with both a Python API and an intuitive web interface for researchers. Designed for scalability, high performance, and data security with support for distributed computing, the platform significantly improves accessibility and usability of single-cell RNA sequencing data, accelerating research in functional genomics, developmental biology, biomedicine, and personalized medicine while providing practical value to the bioinformatics community.