Radni okvir za procjenu starosti koristeći knjižnicu epigenetičkih satova

Brečić, Luka

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

DIPLOMSKI RAD br. 384

RADNI OKVIR ZA PROCJENU STAROSTI KORISTEĆI KNJIŽNICU EPIGENETIČKIH SATOVA

Luka Brečić

Zagreb, lipanj 2024.

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

DIPLOMSKI RAD br. 384

RADNI OKVIR ZA PROCJENU STAROSTI KORISTEĆI KNJIŽNICU EPIGENETIČKIH SATOVA

Luka Brečić

Zagreb, lipanj 2024.

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

Zagreb, 4. ožujka 2024.

DIPLOMSKI ZADATAK br. 384

Pristupnik:	**Luka Brečić (0036515775)**
Studij:	Računarstvo
Profil:	Računarska znanost
Mentor:	izv. prof. dr. sc. Klemo Vladimir

Zadatak: **Radni okvir za procjenu starosti koristeći knjižnicu epigenetičkih satova**

Opis zadatka:

Epigenetički satovi su obećavajući alati u području bioinformatike za procjenjivanje starosti i otkrivanje rizičnih faktora koji ubrzavaju starenje. U okviru diplomskog rada potrebno je proučiti i opisati osnovnu bioinformatičku teorijsku podlogu epigenetičkih satova, posebno DNA metilaciju i CpG otoke te najčešće korištene zapise datoteka za dijeljenje podataka o DNA metilaciji. Dodatno, potrebno je proučiti i detaljno prezentirati povijesni kontekst nastanka epigenetičkih satova i temeljne algoritme i tehnike korištene za njihovo ostvarenje s naglaskom na tzv. Horvathov sat. Praktični dio rada obuhvaća implementaciju radnog okvira za računanje procjene kronološke ili biološke starosti koristeći skup od više različitih satova u obliku knjižnice po uzoru na programsku knjižnicu methylCIPHER koja je ostvarena u programskom jeziku R. Radni okvir ostvariti koristeći programski jezik Python i knjižnicu numpy za učinkovito računanje. Sustav mora podržati elegantno dodavanje i postavljanje novih satova u knjižnicu. Također, svi satovi koji koriste istovjetni pozadinski model moraju biti ostvareni koristeći jedinstveni programski modul. Provjeriti ispravnost rada sustava usporedbom s rezultatima dostupnog sustava methylCIPHER. U radni okvir ugraditi potporu za automatsko ispitivanje. Opisati izgrađeni radni okvir, opisati njegovu programsku arhitekturu, objasniti postupak dodavanja novog sata u knjižnicu satova te prikazati načine korištenja.

Rok za predaju rada: 28. lipnja 2024.

Mentoru, profesoru Klemi Vladimiru, za nesebično izdvojeno vrijeme, beskrajno razumijevanje i strpljenje na ovom mom akademskom putu.

Obitelji, djevojci i prijateljima na podršci koju ste mi pružali sve ove godine.

U spomen dragoj profesorici Lidiji Puljan.

SADRŽAJ

1. Uvod

Ljudska vrsta, kao i sva živa bića, podliježu procesu starenja koji se tradicionalno mjeri kronološki u godinama proteklim od rođenja do, naposljetku, smrti. Trenutak u kojem za jedinku nastupa smrt teško je objasniti samo godinama, svjesni da na to konačno stanje utječe jako puno faktora poput životnih navika, psiho-fizičkih stanja i još mnogih drugih što daje naslutiti da bi se u tom tradicionalnom okviru mogla razmatrati i kakva biološka dob, a ne samo kronološka. Takav koncept biološke dobi označava stvarno stanje organizma uključujući sve faktore koji na njega utječu uz sam proces starenja.

S takvom pretpostavkom nikla je grana genetike s ciljem proučavanja gena i njihovog ponašanja koje nisu direktna posljedica nužno uzrokovana promjenama u sekvencama DNK (proces starenja) imena epigenetika. Epigenetika uz proces starenja proučava i stavlja naglasak na druge faktore koji pridonose cjelokupnom procesu starenja, a posebno DNK metilaciju koja se pokazala jednim od ključnih mehanizama u tom procesu. DNK metilacija proces je dodavanja metilnih skupina na molekule DNK koji reguliraju aktivnost gena bez promjene sekvence DNK. Predvidljiva narav mijenjanja obrazaca DNK metilacije omogućilo je epigenetici podlogu za razvoj takozvanih epigenetičkih satova.

Epigenetički satovi su skup tehnologija koji pružaju mogućnost procjene i mjerenja biološke starosti jedinke temeljenih na epigenetičkim markerima od kojih su najvažniji markeri metilacije DNK na specifičnim lokacijama. Iako su epigenetički satovi relativno nova tehnologija, otvorili su vrata novim istraživanjima i promišljanjima u području biomedicine te nude dublje razumijevanje starenja posebno na molekularnoj razini. Mjerenje biološke dobi pokazat će se daleko praktičnijim i preciznijim od tradicionalnog mjerenja kronološke dobi.

Osim same procjene biološke dobi, epigenetički satovi pokazali su se iznimno važnima za područje istraživanja koje se bavi zdravstvenim stanjima poput karcinoma, kardiovaskularnih bolesti i sličnih drugih bolesti povezanih sa starenjem. Potencijal epigenetičkih satova leži u transformaciji pristupa liječenju i prevenciji raznih bolesti te boljem očuvanju zdravlja u starijoj dobi.

2013. godine profesor na sveučilištu u Kaliforniji, Los Angeles, Steve Horvath izumio je jedan od najznačajnijih epigenetičkih satova takozvani Horvathov sat [7]. Horvathov sat revolucionaran je zbog svoje univerzalnosti jer pruža mogućnost procjene biološke starosti za različite vrste i različita tkiva, a informacije korištene za procjenu biološke starosti pohranjenje su u 353 CpG lokacije. Uz Horvathov sat, naravno, postoji još jako puno epigenetičkih satova s raznim primjenama poput epigenetičkog sata Hannum [6].

Ovaj diplomski rad pruža detaljan pregled epigenetičkih satova, njihove bioinformatičke podloge, povijesni razvoj, mehanizme te dostupne implementacije i ograničenja. Zbog relativno oskudne programske podrške za jednostavnu implementaciju i korištenje epigenetičkih satova, uglavnom pisanu u programskom jeziku R, rad pruža pregled implementacije, arhitekture i rezultata radnog okvira za podršku epigenetičkih satova pisanu u programskom jeziku Python imena *Epygenetics*.

Implementirani radni okvir izgrađen je po uzoru na programsku knjižnicu *methylCIPHER* pisanu u programskom jeziku R te okvir pruža predefiniranu podršku za većinu satova podržanih u knjižnici *methylCIPHER* te se rezultati implementacije radnog okvira evaluiraju s rezultatima iz knjižnice.

2. Osnovna bioinformatička podloga

Epigenetički satovi zapravo su napredni bioinformatički modeli koji procjenjuju biološku starost organizma analizom epigenetičkih modifikacija, ponajprije DNK metilacije. Analizom specifičnih obrazaca metilacije na ključnim CpG lokacijama genoma, moguća je precizna procjena starenja i predikcija biološke starosti, koja se može znatno razlikovati od kronološke starosti zbog utjecaja raznih faktora poput životnih navika i genetike.

2.1. CpG otoci i DNK metilacija

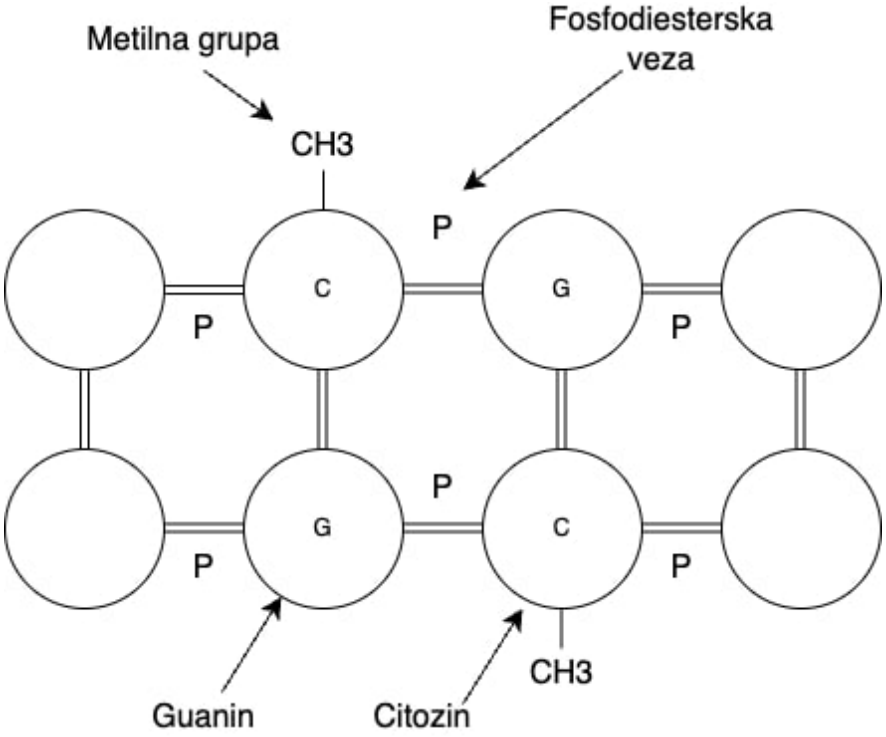
Za potpuno razumijevanje procesa DNK metilacije i njene korelacije s procesom starenja, najvažnije je razumjeti što su to takozvana CpG područja odnosno CpG otoci.

Specifične regije genoma bogate CpG dinukleotidima, to jest regije u kojima se citozin nukleotid ("C") nalazi neposredno prije guanin nukleotida ("G") nazivamo CpG područjem odnosno CpG otokom. Citozin i guanin dinukleotid povezani su fosfodiesterskom vezom što je u samom imenu CpG otoka označeno malim slovom "p".

CpG otoci u genomu su iznimno rijetki, a razlog tome je tendencija citozina da metilira, a potom se procesom deaminacije citozin pretvara u timin. Međutim, specifične regije bogate CpG dinukleotidima (CpG otoci) najčešće su smješteni u promotorima gena ključnima za ekspresiju gena. Uz to, nerijetko se mogu naći i u egzonskim, intronskim te intergenskim regijama. Tipično obuhvaćaju sekvence duljine 300 do 3000 baza gdje je CpG dinukleotid prisutni u više od 50%.

Nemetilirani CpG otoci pružaju mogućnost regulaciji ekspresije gena te transkripcijskim faktorima i RNK polimerazi. U suprotnom dovode do regrutiranja proteina koji inhibiraju transkripciju, odnosno smanjuju ekspresiju gena.

Kovalentno dodavanje metilne skupine na citozin nukleotid prikazanog na slici 2.1 nazivamo DNK metilacija. DNK metilacija zapravo je epigenetička modifikacija prilikom koje se katalizacijom enzimima DNK metiltransferazama (DNMT) prenosi metilna skupina s S-adenozil metionima na ugljik broj 5 citozin nukleotida unutar CpG



Slika 2.1: Metilirani CpG otoci

dinukleotida čime se formira 5-metilcitozin.

Prilikom metilacije ne mijenja se sekvenca DNK, već se samo modificira citozin nukleotid što ima značajan utjecaj na regulaciju gena. Metilacija se kod sisavaca uglavnom događa na CpG dinukleotidima, no postoje rijetki slučajevi metilacije i nekih drugih dinukleotida.

Glavne uloge DNK metilacije su regulacija ekspresije gena gdje unutar promotorske regije gena obično rezultira takozvanim utišavanjem gena (engl. *gene silencing*). Regulacija se postiže inhibicijom vezanja transkripcijskih faktora i regrutiranjem proteina koji direktno kondenziraju kromatin te tako čine gen nedostupnim za transkripcijski proces.

Osim regulacije ekspresije gena, DNK metilacija direktno utječe na održavanje genomske stabilnosti supresiranjem transpozonskih elemenata koji inače izazivaju mutacije nad genomom.

Utjecaj nekih DNK metilacija moguće je povezati s brojnim bolestima poput mnogih vrsta karcinoma gdje promotori gena koji djeluju kao tumor-supresori uslijed hipermetilacije postanu "utišani" (engl. *silenced*) čime je stanicama tumora ili karcinoma omogućena abnormalna i nekontrolirana stanična proliferacija. U suprotnom, pretjerana hipometilacija dovodi do genomske nestabilnosti te aktivacije onkogena.

Također, neurodegenerativne bolesti poput Alzheimerove i Parkinsonove bolesti mogu se povezati s abnormalnim obrascima metilacije specifičnih regija na mozgu što vrlo vjerojatno utječe na ekspresiju gena uključenih u neuralnu funkciju i preživljavanje.

2.2. Zapisi podataka DNK metilacije

Za konstrukciju epigenetičkih satova prije svega potrebni su kvantificirani podatci DNK metilacije uzeti iz određenog uzorka. Prikupljeni podatci moraju biti prikladno digitalizirani i nad njima je za ispravno korištenje potrebno provesti određeno pretprocesiranje koje uglavnom zahtijeva detaljnu analizu i objašnjenje.

Intuitivno je da za samu DNK metilaciju najprije treba identificirati CpG otoke iz uzorka. CpG otoci najčešće se identificiraju uz pomoć bioinformatičkih alata koji analiziraju sekvence genoma i opažaju regije visoke gustoće CpG dinukleotida, a često pružaju mogućnost za odrediti specifične omjere promatranih i očekivanih dinukleotida.

Podatke o DNK metilaciji uglavnom se prikupljaju tehnologijama poput metilacijskih nizova (engl. *methylation arrays*) kao što su *Illumina Infinium Methylation EPIC array* ili disulfidnog sekvenciranja BS-seq (engl. *bisulfite sequencing*). Metilacijski nizovi generalno su brzi i jeftini za analize velikih uzoraka te pružaju mogućnost analiziranja metilacije s unaprijed odabranim CpG područjima genoma. Na primjer, poznati Horvathov sat počiva na podatcima prikupljenih metodom metilacijskih nizova. Nadalje, disulfidno sekvenciranje zlatni je standard za analizu DNK metilacije jer se podatci metilacije mapiraju direktno na pojedinačne nukleotide. U procesu analize metodom disulfidnog sekvenciranja DNK se tretira natrijevim bisulfitom koji nemetilirane citozin nukleotide transformira u uracil, a nemetilirani citozin nukleotidi ostaju nepromijenjeni te tako metilirani CpG nukleotidi postaju lako uočljivi.

Podatci prikupljeni navedenim tehnologijama detaljno se obrađuju. Podatci se najprije filtriraju kako bi se uklonili podatci niske kvalitete ili niske pokrivenosti čime se ojačava pouzdanost podataka. Kao i kod većine podataka korištenih u statističkim i matematičkim modelima kao što su to i epigenetički satovi, podatci se zatim normaliziraju metodama poput BMIQ (engl. *beta-mixture quantile normalization*) ili FunNorm (engl. *functional normalization*) kako bi se uklonile varijacije nepovezane s biološkim promjenama genoma. Filtrirani i normalizirani podatci se konačno mapiraju na referentni genom pazeći na transformacije citozin nukleotida u uracil nukleotide za koji se često koriste bioinformatički alati poput alata Bismark ili BS-Seeker.

Prikupljeni podatci kvantificiraju se na razini pojedinačnih CpG područja te se izražavaju u obliku beta ili M-vrijednosti. Beta vrijednosti predstavljaju omjer intenziteta signala metiliranih i nemetiliranih područja na pojedinačnim CpG dinukleotidima čija se vrijednost izražava u intervalu od 0 (potpuno nemetilirano) do 1 (potpuno metilirano). M-vrijednosti su logaritamske transformacije beta vrijednosti s ciljem kvalitetnije statističke obrade podataka. Iako su beta vrijednosti intuitivne, često su sklone varijabilnosti kod ekstremnih vrijednosti, dok M-vrijednosti bolje podliježu normalnoj distribuciji, što ih čini pogodnijima za većinu statističkih analiza.

Tako kvantificirane vrijednosti uglavnom se koriste kao ulazna točka većine konstruiranih epigenetičkih satova. Dobivene beta vrijednosti DNK metilacije također se mogu integrirati i s drugim genomskim podatcima u svrhu boljeg razumijevanja njihovog ponašanja i uloge u regulaciji gena i patogenezi raznih bolesti.

Podatci DNK metilacije intuitivno su veliki te računalno zahtjevni za pohranu. Uglavnom se koriste za to specijalizirane baze podataka od kojih su najpoznatiji GEO (Gene Expression Omnibus) te ENCODE. Podatci su uobičajeno pohranjeni u BAM (za sekvencirane podatke) ili IDAT (za podatke nastale metodom metilacijskih nizova) formatima, no često se za lakšu računalnu implementaciju epigenetičkih satova korist i CSV format.

2.3. Matematički modeli epigenetičkih satova

Epigenetički satovi uglavnom počivaju nad implementacijama raznih matematičkih modela koji se koriste za analizu DNK metilacije i, efektivno, predikciju biološke starosti. Neki od matematičkih modela obično su:

1. Linearni regresijski modeli

Linearni regresijski modeli uglavnom su najkorišteniji modeli za konstrukciju epigenetičkih satova. Biološka starost procijenjena je kao linearna kombinacija dobivenih beta vrijednosti uzorka nad odabranim CpG područjima. Linearni regresijski model glasi:

$$\text{Biološka starost} = \beta_0 + \sum_{i=1}^{n} \beta_i \cdot X_i$$

Gdje je:

- β_0 konstanta
- β_i koeficijent regresije za i-to CpG područje

– X_i beta vrijednost na i-tom CpG području

– n broj CpG područja uključenih u model

Najpoznatiji primjer epigenetičkog sata konstruiranog kao linearni regresijski model je Horvathov sat.

2. Elastic Net model

Model regularizirane linearne regresije, popularno nazvan Elastic Net, još jedan je od često korištenih matematičkih modela za konstrukciju epigenetičkih satova. Elastic Net model koristi kombinaciju L1 i L2 regularizacija, poznatiju kao Lasso-Ridge regresija, što pruža mogućnost selekcije relevantnih CpG područja te onemogućuje prenaučenost modela. Ciljna funkcija Elastic Net modela je minimizator:

$$\min_{\beta} \left(\frac{1}{2N} \sum_{i=1}^{N} (y_i - X_i \cdot \beta)^2 + \lambda \left(\alpha \|\beta\|_1 + \frac{(1-\alpha)}{2} \|\beta\|_2^2 \right) \right)$$

Gdje je:

– y_i predviđena biološka starost

– X_i metilacijske vrijednosti na odabranim CpG mjestima

– λ regularizacijski parametar koji kontrolira jačinu regularizacije

– α parametar koji balansira između L1 ($\|\beta\|_1$) i L2 ($\|\beta\|_2$) regularizacije

Najpoznatiji primjer epigenetičkog sata konstruiranog kao Elastic Net model je GrimAge epigenetički sat koji biomarkere korelirane sa starenjem integrira s podatcima DNK metilacije.

3. Strojno učenje i Random Forest

Za konstrukciju epigenetičkih satova često se koriste i metode strojnog učenja poput metode slučajnih šuma (engl. *random forest*). Metoda slučajnog šuma je kombinacija raznih metoda koje sudjeluju u konstruiranju stabla odluka s ciljem procjene biološke starosti. Svako tako konstruirano stablo odluku o procjeni donosi nad podskupom CpG područja, a konačna predikcija je prosjek svih procjena. Model metode slučajnih šuma je:

$$\text{Biološka starost} = \frac{1}{M} \sum_{m=1}^{M} \text{predikcija}_m(X)$$

Gdje je:

– M broj stabala u šumi

– predikcija$_m(X)$ predikcija biološke starosti m-tog stabla odluke

Modeli slučajnih šuma imaju sposobnost zapažanja složenih nelinearnih odnosa metilacije i biološke starosti, a ujedno su i robusni prema šumu i prenaučenosti.

4. Bayesovske metode

Probabilistički pristup modeliranja metilacijskih podataka te predikcije bioloških starosti konstruira se bayesovskim metodama. Bayesovske metode procjenjuju posteriorne distribucije parametara modela te pružaju mogućnost integracije prethodnih znanja o podatcima. Ti modeli posebno su korisni kada podatci imaju malen broj uzoraka ili je neizvjesnost u podatcima prevelika, a uglavnom se koriste u kombinaciji s drugim modelima s ciljem preciznije procjene biološke starosti.

5. Analiza glavnih komponenti (PCA) i analiza glavnih osi (PCOA)

PCA i PCOA su tehnike za smanjenje dimenzionalnosti koje se često koriste prije primjene regresijskih ili modela strojnog učenja. Te metode transformiraju podatke viših dimenzija u manje skupove glavnih komponenti koje zadržavaju većinu varijabilnosti u podacima. Ove glavne komponente zatim se koriste kao ulazne varijable u modelima predikcije.

3. Epigenetički satovi

3.1. Povijesni razvoj

Iako su epigenetički satovi relativno nova tehnologija u području biomedicine, od iznimnog su značaja. Kao kruna u istraživanju na području epigenetike, statistike i bioinformatike, njihov razvoj poboljšao je generalno razumijevanje starenja te otvorio nove mogućnosti u personaliziranoj medicini i istraživanju raznih bolesti.

Svoje početke epigenetički satovi pronalaze još kada je prvi puta uveden pojam epigenetike sredinom 20. stoljeća. Britanski biolog, Conrad Waddington, 1942. godine uvodi pojam epigenetike koji opisuje složene interakcije gena i fenotipskog okoliša organizma. Time započinje period istraživanja konkretnih molekularnih mehanizama koji polako uvrježuju sam pojam epigenetike i njezin značaj u biomedicini.

Nedugo nakon, 1950-ih godina prvi puta su identificirane metilne skupine vezane uz citozin nukleotid i time postavili temelje za razumijevanje DNK metilacije. DNK metilacija vrlo brzo iz obične molekularne zanimljivosti postaje ključnom za praćenje regulatornog mehanizma genoma i ekspresije gena, razvoju embrija te u održavanju stanične memorije.

Koncem 20. stoljeća DNK metilacija pronalazi svoj potpuni značaj u okvirima epigenetike i iz pukog slučajnog procesa postaje glavni akter u promatranju i očuvanju genomske stabilnosti. Tako je ustanovljeno da je DNK metilacija gena promotora povezana s njihovom represijom te da je demetilacija povezana s njihovom aktivacijom. To otkriće pridonijeli su boljem razumijevanju epigenetičkih promjena te kako točno DNK metilacija sudjeluje u regulaciji ključnih bioloških procesa te na koji točno način pridonosi patogenezi bolesti.

Znanstvenici su ustanovili da su razine DNK metilacije usko korelirane s procesom starenja te se DNK metilacija prvi puta počinje razmatrati kao potencijalni biomarker biološke starosti. Otkriveno je da tijekom starenja dolazi do promjene obrazaca DNK metilacije te je postavljena hipoteza da upravo te promjene održavaju biološku starost organizma neovisno o mjerilu tradicionalne kronološke starosti.

Početkom 21. stoljeća koncept biološke starosti postaje sve popularniji i uvriježava se mišljenje kako tradicionalni pristup izražavanja starosti kronološki ne preslikava uvijek stvarno zdravstveno stanje jedinke. Razlike zdravstvenih stanja dviju jedinki iste kronološke dobi dovode do jasne potrebe za razvojem pouzdanih biomarkera za starost koji imaju mogućnost preciznije procjene biološke starosti organizma za koju se DNK metilacija pokazala obećavajućom.

2011. godine dolazi do jedne od prvih konkretnih implementacija epigenetičkog sata kada Gregory Hannum i njegovi suradnici razvijaju sat koji za procjenu biološke starosti koristi DNK metilaciju. Hannumov epigenetički sat [6] dizajniran je kao linearni regresijski model koji nad uzorcima krvi povezuje točno određena CpG područja s kronološkom dobi. Njegovu univerzalnu primjenjivost ograničilo to što je model bio namijenjen za specifična tkiva.

S tom potrebom univerzalnog epigenetičkog sata s obzirom na uzorak tkiva, 2013. godine na Sveučilištu Kalifornija u Los Angelesu, profesor Steve Horvath objavljuje rad o novom univerzalnom epigenetičkom satu. Horvathov epigenetički sat [7] postaje jedan od najznačajnijih i najkorištenijih alata iz područja epigenetike. Sat je izgrađen na 353 CpG područja na temelju kojih procjenjuje biološku starost nad različitim tkivima i vrstama što ga je učinilo široko primjenjivim u raznim istraživačkim i kliničkim praksama.

Otkrićem revolucionarnog Horvathovog sata 2013. godine, nastaje cijeli niz novih epigenetičkih satova s ciljem usavršavanja, proširenja, specijalizacije ili unaprjeđenja na još višu univerzalnost epigenetičkih satova. Tako nastaju satovi koji su na primjer prilagođeni točno određenim tkivima ili su namijenjeni za procjenu biološke starosti s specifičnim bolestima ili životnim navikama. Sva nova otkrića uvelike su doprinijela istraživanju starenja i novim otkrićima u epidemiologiji.

Tako 2018. godine dolazi do razvoja epigenetičkog sata GrimAge [14], jednog od najvažnijih napredaka na području epigenetičkih satova. GrimAge epigenetički sat osim sposobnosti procjene biološke starosti ima mogućnost integracije raznih informacija o DNK metilaciji i biomarkerima usko povezanih sa starenjem ili rizikom smrtni. Ovaj model jedan je od prvih koji uspješno omogućuju procjenu preostalog životnog vijeka jedinke ili njenih rizika od smrtonosnih bolesti čime postaje jedan od najznačajnijih alata na području personaliziranje medicine. GrimAge epigenetički sat izgrađen je po uzoru na Horvathov epigenetički sat, ali uključuje dodatne složenosti u kojima uključuje razne biomarkere te se pokazao jednim od najpreciznijih epigenetičkih satova današnjice.

Posljednjih godina epigenetički satovi dobili su na iznimnoj važnosti te se koriste u

mnogim područjima biomedicinskih istraživanja kao i u raznim epidemiološkim studijima, posebno za proučavanje kauzalnosti utjecaja okoliša, životnih navika ili genetskih faktora na starenje. Također, epigenetički satovi pokazali su se korisnima u personaliziranoj medicini, ponajviše u prilagodbi terapije pacijenta s obzirom na pacijentovu biološku starost.

3.2. Vrste epigenetičkih satova

Uz značajan napredak razvoja epigenetičkih satova u posljednjem desetljeću, do danas se razvilo nekoliko obitelji epigenetičkih satova koji se koriste u različite svrhe.

Epigenetički satovi za procjenu kronološke starosti, još poznatiji kao opći epigenetički satovi, konstruirani su za procjenu kronološke i biološke starosti na temelju DNK metilacije različitih tkiva i vrsta živih organizama, među kojima je najpoznatiji Horvathov epigenetički sat [7]. Razvoj ove obitelji epigenetičkih satova omogućila je korištenje istih u raznim istraživačkim i kliničkim praksama zbog svoje primjene na različite specijalizirane vrste tkiva. Međutim, manjak univerzalnosti pojedinih općih epigenetičkih satova uvelike ograničava njihovu primjenu u određenim kontekstima. Sljedeći najpoznatiji epigenetički sat iz ove obitelji je već spomenut Hannumov epigenetički sat [6], a još neki od satova iz ove obitelji su:

- Bocklandtov epigenetički sat [2]
- Garagnanijev epigenetički sat [5]
- Linov epigenetički sat [13]
- Vidal-Bralo epigenetički sat [22]
- Weidnerov epigenetički sat [23]
- Zhangov epigenetički sat [28]
- Zhangov poboljšani epigenetički sat [27]

Pored općih epigenetičkih satova, razvijena je i obitelj epigenetičkih satova za procjenu biološke dobi na temelju specifičnih tkiva te epigenetički satovi za procjenu smrtnosti. Zbog svoje specifičnosti procjene određenih tkiva, ova obitelj epigenetičkih satova pokazala se preciznijom u usporedbi s općim epigenetičkim satovima. Najpoznatiji primjeri iz ove obitelji epigenetičkih satova su DNAmPhenoAge epigenetički sat [12] te GrimAge epigenetički sat [14], oba razvijena 2018. godine. DNAmPhenoAge epigenetički sat koristi DNK metilaciju specifičnih CpG područja koje su odabrane kao najrelevantnije za procjenu kronološke dobi i zdravstvenih rizika te je iznimno

precizan u predviđanju rizika od bolesti usko povezanih sa starenjem jedinke kao na primjer kardiovaskularne bolesti. Nažalost, primjena DNAmPhenoAge epigenetičkog sata ograničena je isključivo na krvne uzorke. GrimAge epigenetički sat primjer je modela koji uz DNK metilaciju kombinira određene biomarkere starenje ili rizika od smrti te integrira te informacije s ciljem precizne predikcije ne samo biološke starosti, već i rizika od smrtnosti te kroničnih bolesti. Zbog svojih mogućnosti, GrimAge je postao jedan od najvažnijih alata u epigenetičkim i kliničkim istraživanjima.

Obitelj gestacijskih i pedijatrijskih epigenetički satova uključuje specijalizirane epigenetičke satove koji se koriste za procjenu biološke starosti fetusa tijekom trudnoće i jedinke u djetinjstvu. Ovi epigenetički satovi omogućuju istraživanje epigenetskih promjena koje se događaju u tim ključnim fazama razvoja jedinke, pružajući uvid u to kako čimbenici iz okoline jedinke mogu utjecati na njeno zdravlje i razvoj. Gestacijski epigenetički satovi koriste se za procjenu biološke starosti fetusa tijekom trudnoće, a uključuju analizu DNK metilacije iz uzoraka placente ili pupkovine. Takvi epigenetički satovi pomažu u razumijevanju utjecaja čimbenika majke poput prehrane ili izloženosti majke toksinima na razvoj fetusa. Pedijatrijski epigenetički satovi primjenjuju se za praćenje biološke starosti djece u ranim godinama života, koristeći uzorke krvi ili druge biološke uzorke kako bi se proučavale promjene u DNK metilaciji povezane s razvojem i zdravljem djeteta. Neki od epigenetičkih satova iz ove obitelji su:

- Bohlinov epigenetički sat [3]
- Knightov epigenetički sat [9]
- Leejevi epigenetički satovi [11]
- Mayneov epigenetički sat [15]
- PedBE epigenetički sat [17]

Obitelj koja uključuje epigenetičke satove za procjenu biološke dobi nad rakom pogođenim tkivima te nad staničnim diobama (mitoza) sadrži specijalizirane alate koji analizom DNK metilacije identificiraju promjene specifične za tumorske stanice. Ključni su za procjenu brzine rasta ili agresivnosti tumorskih i kancerogenih tkiva te pružaju mogućnost za lakše praćenje razvoja tih bolesti. Epigenetički satovi iz ove obitelji su:

- EpiTOC epigenetički sat [25]
- EpiTOC2 te HypoClock epigenetički satovi [20]
- MiAge epigenetički sat [26]

Osim navedenih obitelji epigenetičkih satova, važno je spomenuti još obitelj epigenetičkih satova koji svoju procjene biološke starosti konstruiraju nad specifičnim životnim navikama kao i epigenetičke satove koji se ne temelje na uzorcima krvi. Obitelji epigenetičkih satova koji svoju procjenu temelje na specifičnim životnim navikama pripada skupina epigenetičkih satova koju je razvio bioinformatički analitičar Daniel L. McCartney [16] i uglavnom uključuju procjenu biološke starosti s obzirom na životne navike kao što su konzumacija alkohola, pušenje i prekomjerna tjelesna masa. Satovi koji svoje procjene biološke starosti ne temelje na isključivo uzorcima krvi su epigenetički sat DNAmClockCortical [19] te Horvathov *Skin&Blood* epigenetički sat [8].

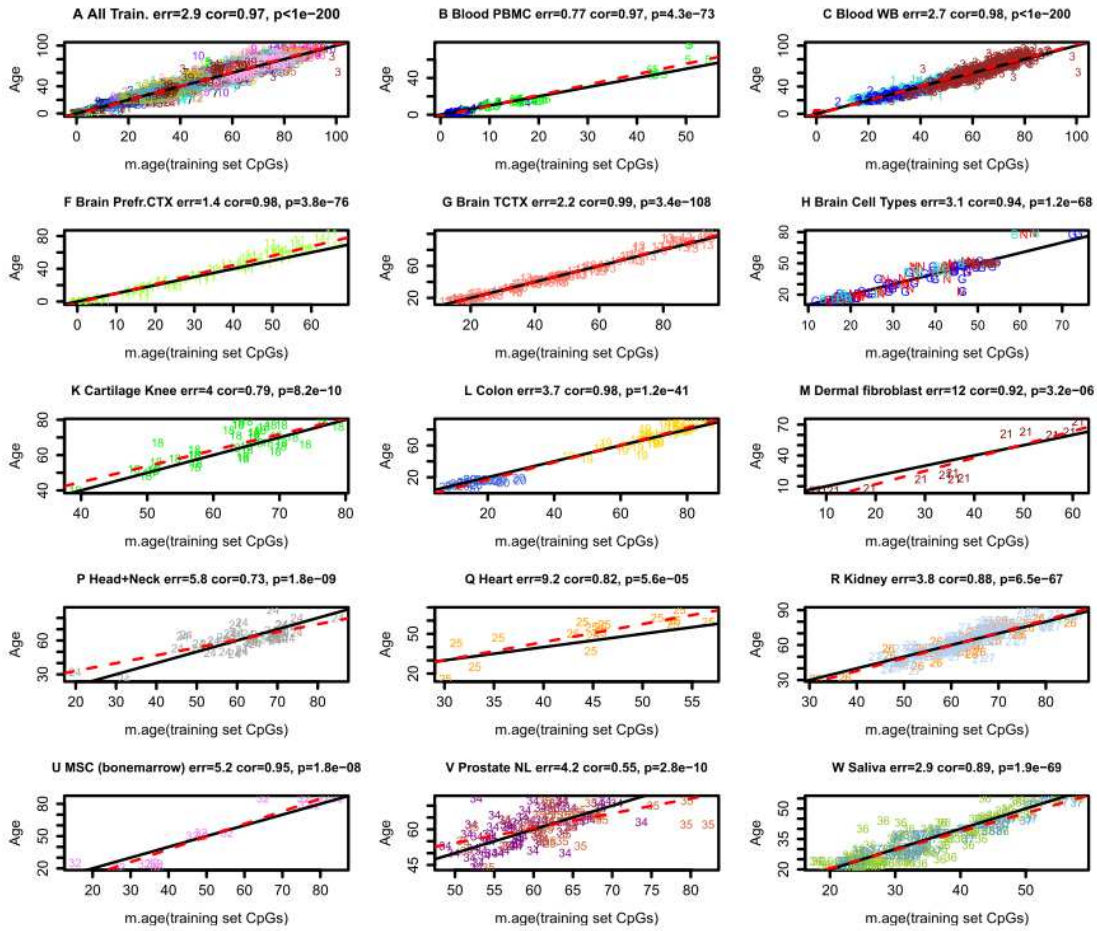
3.3. Horvathov sat

Horvathov epigenetički sat razvio je 2013. godine profesor na Sveučilištu Kalifornija u Los Angelesu, Steve Horvath. Motivacija za razvoj epigenetičkog sata bilo je razumijevanje mehanizma starenja i utjecaj epigenetičkih modifikacija na ovaj proces. Njegova istraživanja fokusirala su se na identificiranje epigenetičkih promjena koje su u uskoj korelaciji s kronološkom i biološkom dobi promatrane jedinke.

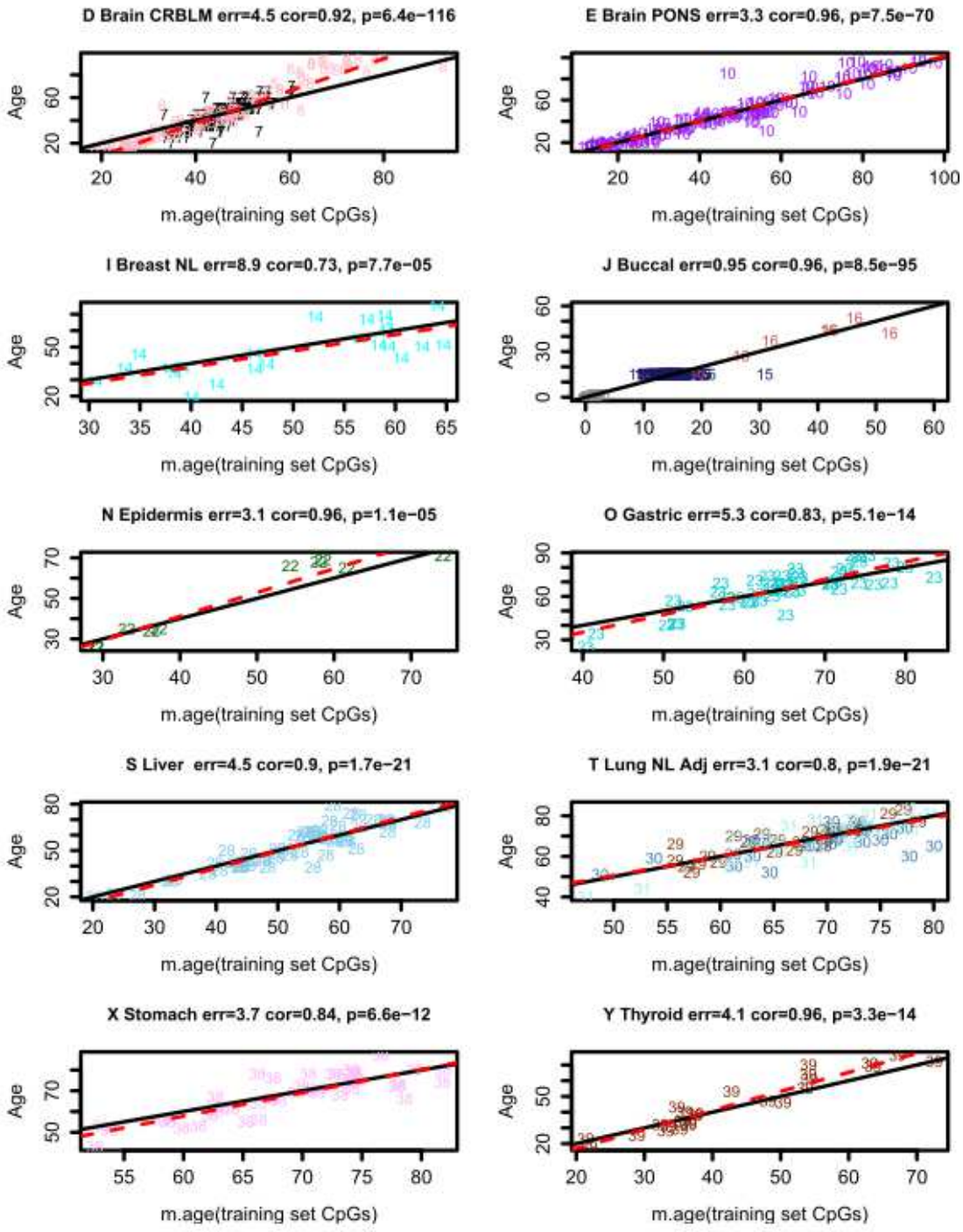
Horvathov epigenetički sat [7] jedan je od najvažnijih i najutjecajnijih alata razvijenih na području epigenetike, posebnog značaja za proučavanje biološkog starenja. Procjena biološke starosti organizma koristeći Horvathov epigenetički sat počiva na specifičnim obrascima DNK metilacije. Zbog svoje preciznosti, Horvathov epigenetički sat postao je zlatni standard u istraživanju starenja, epigenetike te je do danas ostao kao referentna točka za daljnji razvoj novih epigenetičkih satova.

Za razvoj epigenetičkog sata, profesor Horvath je analizirao DNK metilacijske podatke iz više od 8000 uzoraka tkiva i stanica koji su, između ostalog, uključivali 51 različitu vrstu tkiva i stanica ljudskog tijela. Cilj je bio razviti univerzalni epigenetički model koji može precizno procijeniti biološku starost različitih tkiva koristeći zajednički skup specifičnih CpG područja. Nakon opsežnih analiza, profesor Horvath identificirao je 353 specifična CpG područja genoma koje su pokazivale predvidljive promjene metilacije u korelaciji s kronološkom dobi jedinke.

Horvathov epigenetički sat konstruiran je po uzoru na model linearne regresije, međutim, Horvathov model ne opisuje direktno kronološku dob iz predočenih uzoraka DNK metilacije, već model opisuje njihove log vrijednosti. Log vrijednosti uzete su za predodžbu zato što frekvencije DNK metilacije opadaju starenjem pa log vrijednosti bolje opisuju taj nelinearan odnos. Ta predodžba može biti prikazana ovako:



Slika 3.1: Prikaz odnosa kronološke dobi te procijenjene biološke dobi Horvathovog epigenetičkog sata preuzetog iz [7] gdje je prikazano kako je kod gotovo svih vrsta tkiva korelacija veća ili jednaka 0.97 te je standardna devijacija manja ili jednaka 2,9 godina



Slika 3.2: Nastavak prikaza sa slike 3.1 odnosa kronološke dobi te procijenjene biološke dobi Horvathovog epigenetičkog sata preuzetog iz [7]

$$\text{Transformirana dob} = log(\text{Kronološka dob} + 1)$$

Horvathov epigenetički sat tada može biti opisan kao:

$$\text{Transformirana dob} = \beta_0 + \sum_{i=1}^{n} \beta_i \cdot X_i$$

Gdje je:

- β_0 konstanta
- β_i koeficijent regresije za i-to od 353 CpG područja
- X_i beta vrijednost na i-tom od 353 CpG područja
- n broj CpG područja uključenih u model (Horvathov model 353)

Na ovako dobivene vrijednosti potrebno je primijeniti anti-transformacijsku metodu kojom se poništava bilo kakva transformacija nad dobnim podatcima učinjena prilikom izračuna. U slučaju Horvathovog sata, konačnu kronološku dob moguće je dobiti primjenom sljedeće funkcije:

$$\text{Procjena Kronološke dobi} = exp(\text{Transformirana dob} - 1)$$

Odabrana CpG područja korištena za konstrukciju Horvathovog sata posebna su zbog svoje konzistentnosti i predvidljivosti u različitim vrstama tkiva što je krucijalno za univerzalnost ovog epigenetičkog sata. Za razliku od prethodnih modela poput Hannumovog epigenetičkog sata [6], koji su uglavnom specijalizirani za točno određena tkiva, Horvathov epigenetički sat primjenjiv je na širok opseg različitih tkiva i stanica, što ga čini izuzetno korisnim u raznim istraživačkim i kliničkim kontekstima.

Osim univerzalnosti Horvathovog epigenetičkog sata, njegova iznimno važna karakteristika je točnost i preciznost procjene, gdje Horvathov model pokazuje izuzetno visoku korelaciju s kronološkom dobi, pri čemu je standardna devijacija pogreške procjene (engl. *mean average error*) u različitim tkivima bila oko 3,6 godina, što znači da model precizno procjenjuje kronološku dob unutar nekoliko godina. Konkretni podatci o preciznosti prikazani su na slikama 3.1 i 3.2

Široka primjena Horvathovog epigenetičkog sata leži u istraživanjima procesa starenja, epidemiološkim istraživanjima, kliničkim istraživanjima, personaliziranoj medicini te forenzici. Najvažnija primjena je procjena biološke starosti nad različitim populacijama te identificiranje pojedinaca čija biološka starost značajno odstupa od njihove kronološke starosti. Takvi pojedinci su potencijalno pod povećanim rizikom

obolijevanja od bolesti povezanih sa starenjem poput kardiovaskularnih bolesti, dijabetesa, raka ili neurodegenerativnih poremećaja.

3.4. Postojeće tehničke implementacije

Većina postojećih programskih rješenja i implementacija raznih sustava za rad s epigenetičkim satovima ili za analizu DNK metilacije pisana je uglavnom u programskom jeziku R. Iako postoje komercijalni sustavi za rad s epigenetičkim satovima, postoji značajan broj javno dostupnih alata dostupnih za istraživanje i ispitivanje određenih funkcionalnosti.

1. Javno dostupni R paketi

Jedna od najvećih javno dostupnih platformi za bioinformatičke analize pisana u programskom jeziku R je platforma Bioconductor. Bioconductor nudi brojne knjižnice za analizu DNK metilacije ili za primjenu epigenetičkih satova od kojih su najpoznatiji ENmix [24], wateRmelon, methylclock [18] i minfi [1]. ENmix paket osmišljen je za obradu i analizu podataka dobivenih iz Illumina DNK metilacijskih nizova poput sesame450K te EPIC. U paketu podržane su funkcionalnosti za preprocesiranje podataka, normalizaciju podataka, identifikaciju diferencijalne metilacije te funkcionalnost za integraciju podataka s raznim epigenetičkim satovima. Paket wateRmelon u manjem opsegu podržava funkcionalnosti za kontrolu i normalizaciju Illumina metilacijskih podataka. Methylclock paket specijaliziran je za rad s raznim popularnim epigenetičkim satovima te pruža podršku za implementaciju satova u vlastite projekte, a paket minfi, iako idejno nije razvijen za rad s epigenetičkim satovima, širok spektar korisnih funkcionalnosti te laka integracija s ostalim dostupnim Biocondutor paketima ga čini iznimno korisnim.

Za ovaj rad, najvažniji javno dostupan paket pisan u programskom jeziku R koji podržava rad s epigenetičkim satovima je paket methylCIPHER [21]. MethylCIPHER paket razvijen je s ciljem lakog korištenja unaprijed podržanih epigenetičkih satova. Paket podržava više od 20 epigenetičkih satova koji kao ulazne vrijednosti primaju ciljane beta vrijednosti za koje se želi napraviti starosna procjena, unaprijed poznate podatke o fenotipu te, opcionalno, imputacijske podatke za beta vrijednosti koje nedostaju u uzorku za određeni sat. MethylCIPHER između ostaloga podržava i najpoznatije satove poput Horvathovog i Hannumovog epigenetičkog sata, no najveće ograničenje samog paketa je njegova robusnost i teška razumljivost što su ujedno i glavni motivi za pisanje ovog rada.

2. Javno dostupne Python knjižnice

Iako među javno dostupnim alatima uglavnom prevladavaju paketi implementirani u programskom jeziku R, postoji nekolicina knjižnica pisanih u programskom jeziku Python koje podržavaju rad s pojedinačnim ili određenim podskupom poznatih epigenetičkih satova. Najpopularnije knjižnice s podrškom za rad s epigenetičkim satovima su ComputAgeBench [10] te AltumAge [4]. ComputAgeBench je platforma koja podržava usporedbu i ocjenjivanje preciznosti i performansi raznih epigenetičkih satova te pruža standardizirani okvir za evaluacija procjena epigenetičkih satova. Također, ComputAgeBench ima javno dostupan huggingface repozitorij s podatcima potrebnim za podržanu validaciju. AltumAge s druge strane je Python knjižnica u kojoj je implementiran novi istoimeni epigenetički sat temeljen na dubokom učenju izgrađen po uzoru na Horvathov epigenetički sat.

3. Komercijalni sustavi i web-platforme

Osim javno dostupnih repozitorija, paketa i knjižnica za podršku i rad s epigenetičkim satovima, također postoje komercijalni sustavi kao i web-platforme za obradu metilacijskih podataka ili procjenu biološke starosti. Najpoznatija javno dostupna platforma za procjenu biološke starosti je Horvathov Online Kalkulator koju je razvio profesor Steve Horvath. Kalkulator pruža besplatan alat za računanje i procjenu biološke dobi na temelju dobivenih beta vrijednosti iz DNK metilacije. Važno je da su podatci prije unosa u kalkulator prikladno obrađeni. Od komercijalnih rješenja, važno je spomenuti Illumina Genome Studio, softver koji pruža podršku za analizu podataka generiranih Illumina DNK metilacijskim platformama. Iako primarna zadaća ovog softvera nisu epigenetički satovi, njihova implementacija i rad s njima, softver omogućuje obradu i normalizaciju DNK metilacijskih podataka koji se onda mogu prikladno koristiti u postojećim javno dostupnim paketima za procjenu biološke starosti.

4. Epygenetics

Cilj ovog rada je opisati implementaciju radnog okvira za podršku i laku integraciju rada s epigenetičkim satovima u obliku knjižnice implementirane u programskom jeziku Python. Konačno rješenje utjelovljeno je u obliku Python knjižnice imena *epygenetics* koja prati dobre prakse programiranja u programskom jeziku Python, posljednji PEP8 standard programiranja, a za lakšu integraciju knjižnice u svrhu implementiranja novih epigenetičkih satova u budućnosti utilizira se dobra praksa nasljeđivanja s dovoljnim razinama apstrakcije.

4.1. Motivacija

Ponajveća motivacija za izradu radnog okvira *epygenetics* za rad i laku implementaciju epigenetičkih satova je ta što su javno dostupni paketi koji pružaju željene funkcionalnosti uglavnom pisani nezgrapno u programskom jeziku R. Paketi su dosta robusni na promjene i na laku implementaciju potencijalnih novih satova, a sam programski jezik R, iako moćan, nije dovoljno zastupljen u nekoj komercijalnijoj uporabi među populacijom razvojnih inženjera.

Za implementaciju radnog okvira, programski jezik Python pokazao se najboljim izborom među ostalim programskim jezicima koji podržavaju objektno-orijentiranu paradigmu zbog bogatog izvora knjižnicama za lako manevriranje velikim podatcima kao što je knjižnica pandas te brzim i učinkovitim knjižnicama za kompleksne matematičke operacije te operacije strojnog učenja kao što su NumPy, SciPy te scikit-learn knjižnice.

Kao uzor za izradu radnog okvira *epygenetics* uzet je javno dostupan paket razvijen u programskom jeziku R imena *methylCIPHER*. Iako *methylCIPHER* pruža poprilično širok spektar epigenetičkih satova koji se mogu koristiti, testirati i istraživati u njihovom paketu (njih preko 20), sam paket je dosta neintuitivan za korištenje, slabo je dokumentiran i objašnjen te je dosta inertan na promjene ili integracije novih epigenetičkih satova.

Stoga je ponajveći cilj ovog rad da radni okvir, osim funkcionalnih epigenetičkih satova, podržava i lako i intuitivno korištenje knjižnice za implementaciju ili razvijanje potencijalnih novih epigenetičkih satova.

4.2. Arhitektura i korištene tehnologije

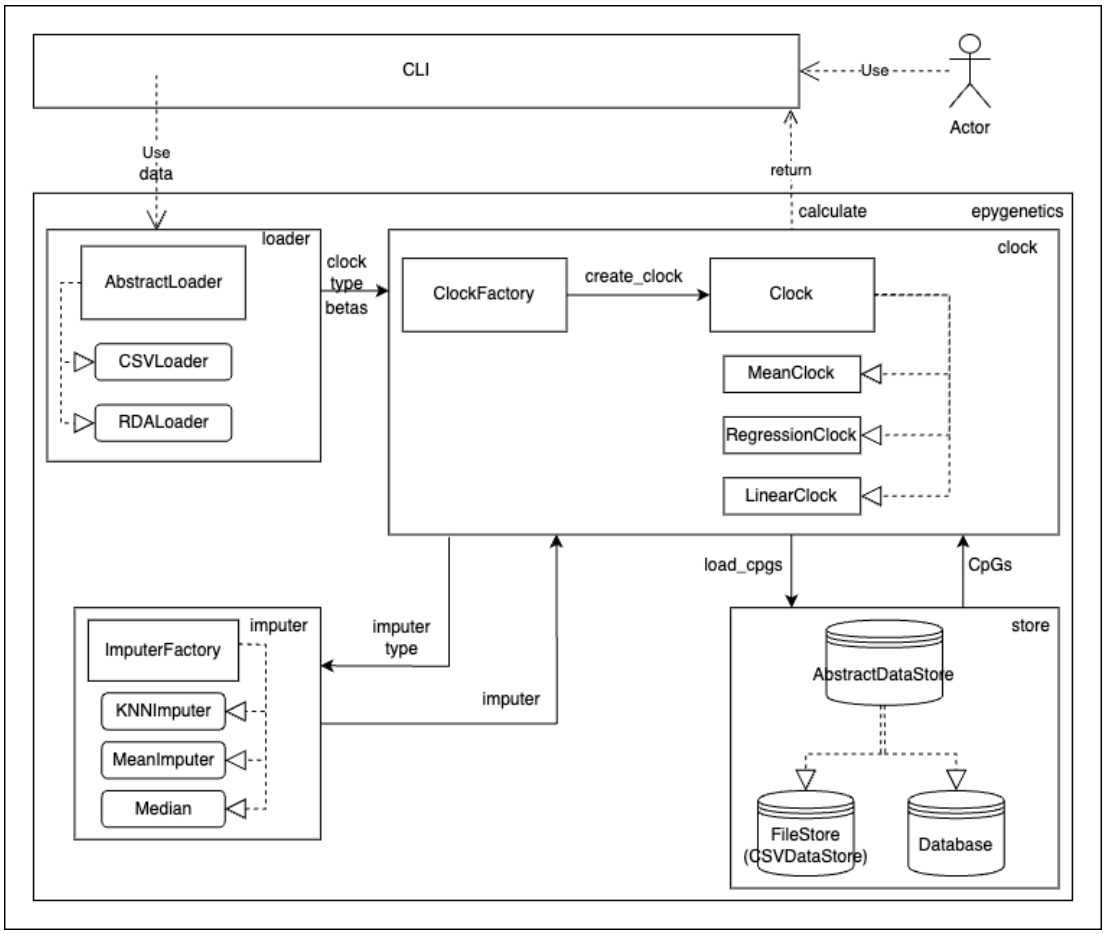
Prilikom razvoja programske potpore za radni okvir *epygenetics* stavljen je naglasak na praćenje dobrih programerskih načela i praksi kao što su načelo jedinstvene odgovornosti, načelo lake nadogradivosti te čist kod (engl. *clean code*).

Na slici 4.1 prikazan je pogled na logičku arhitekturu prilikom korištenja unaprijed implementiranih epigenetičkih satova za čiji rad implementirani radni okvir pruža podršku. Pogled osim logičke domene implementacije, implicitno prikazuje i modularnost implementirane programske podrške pa je tako vidljivo da je programska potpora razvijena koristeći module programskog jezika Python. Svaki modul (imena programskih modula navedena su u gornjim desnim kutovima njihovih okvira) odgovoran je za ispravno rukovođenje svojim funkcionalnostima. Tako, na primjer, programski modul *clock* rukovodi instanciranjem željenog razreda te izvođenjem programskog koda za procjenu kronološke starosti.

Korisniku je ulaz u korištenje sustava otvoren putem naredbenog korisničkog sučelja (engl. *command line interface (CLI)*) gdje je korisnik dužan unijeti potrebne podatke kako bi ispravno koristio funkcionalnosti radnog okvira. Korisničko sučelje zatim predaje potrebne podatke radnom okviru koji najprije parsira ulazne argumente te koristeći konkretnu implementaciju apstraktnog razreda *AbstractLoader* učitava podatke o beta vrijednostima DNK metilacije iz nekog od podržanih formata datoteka.

Učitani podatci tada se šalju razredu *ClockFactory* koji instancira željeni unaprijed podržani epigenetički sat iz argumenta naziva epigenetičkog sata. Konkretna implementacija epigenetičkog sata prilikom instanciranja komunicira s razredom *ImputerFactory* kako bi se koristila željena imputacijska metoda te komunicira s konkretnom implementacijom razreda *AbstractDataStore* iz koje učitava potrebne podatke o skupu CpG otoka koji se koriste prilikom procjene kronološke starosti. Konačno, izlaz radnog okvira korisniku je prikazan putem naredbenog korisničkog sučelja.

Dobra programerska načela najviše su utilizirana koristeći oblikovne obrasce strategije i okvirne metode te dobre modularnosti same programske potpore. Tako, na primjer, u baznom razredu *Clock* okvirnu metodu predstavlja metoda *execute* koja u točno određenom redoslijedu poziva apstraktne metode *validate* i *calculate* čija implementacija ovisi o satovima koji će u budućnosti nasljeđivati bazni razred *Clock*.



Slika 4.1: Prikaz logičke arhitekture radnog okvira *epygenetics*

Metoda *validate* u naravi je zadužena za provjeru i validaciju parametara koje epigenetički sat koristi pri izračunu. Tako će se u toj metodi provjeravati dostupnost informacije DNK metilacije za pojedina CpG područja čiju prisutnost sat zahtijeva za ispravan rad te će validirati i priložene beta vrijednosti nad kojima se radi izračun. U slučaju nedostatka informacije ili nepravilnosti u podatcima, ovisno o instrukciji zadanoj pozivom, metoda *validate* zadužena je i za imputaciju podataka po navedenoj metodi u instrukciji.

Nakon uspješno obavljene validacije, u metodi *calculate* potrebno je implementirati logiku kojom implementirani epigenetički sat obavlja procjenu starosti s obzirom na podatke poslane kroz parametre.

Te dvije metode dovoljne su da opišu većinu do sada dostupnih implementacija epigenetičkih satova koje mogu naslijediti vršni razred *Clock*. Kako je većinu satova koje paket *methylCIPHER* podržava moguće dosta generično izraziti jer uglavnom prate tri modela implementacije, tako su u knjižnici *epygenetics* podržane tri vrste satova prikazani na slici 4.2 koje nasljeđuju vršni razred *Clock*:

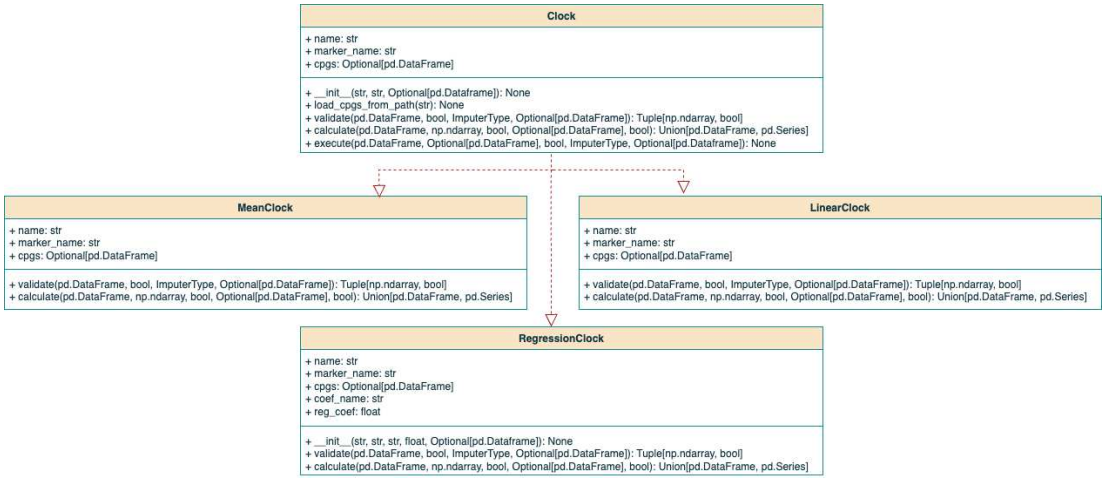
- Razred *LinearClock* koji prati model običnog linearnog mapiranja ulaznih vrijednosti
- Razred *MeanClock* koji prati model linearnog mapiranja srednjih ulaznih vrijednosti
- Razred *RegressionClock* koji prati model (regularizirane) linearne regresije

Svaki od te tri vrste epigenetičkih satova implementira svoje funkcionalnosti metoda *validate* i *calculate* iz vršnog razreda *Clock*. Konkretne implementacije unaprijed podržanih epigenetičkih satova koje pruža radni okvir *epygenetics* uglavnom nasljeđuju neki od te tri vrste implementacija vršnog razreda čime je zadovoljeno načelo oblikovnog obrasca strategija.

Metoda *validate*, u slučaju nepotpunosti ili nepravilnosti u podatcima o DNK metilaciji ili njenim beta vrijednostima, zahtijeva imputaciju podataka koji nedostaju bez koje sat ne može napraviti procjenu biološke starosti.

Imputacijske metode su unaprijed određene i korisnik je dužan u slučaju potrebne imputacije odabrati jednu od metoda koje radni okvir podržava. Trenutno implementirane imputacijske metode su:

- Imputacijska metoda temeljena na medijanu vrijednosti uzorka *MedianImpute*
- Imputacijska metoda temeljena na srednjoj vrijednosti uzorka *MeanImpute*
- Imputacijska metoda temeljna na algoritmu K najbližih susjeda *KNNImpute*



Slika 4.2: Dijagram vršnog razreda *Clock* i osnovne tri podvrste podržanih epigenetičkih satova

– Regularna imputacijska metoda mapiranja vrijednosti iz unaprijed prikupljenih relevantnih podataka *RegularImpute*

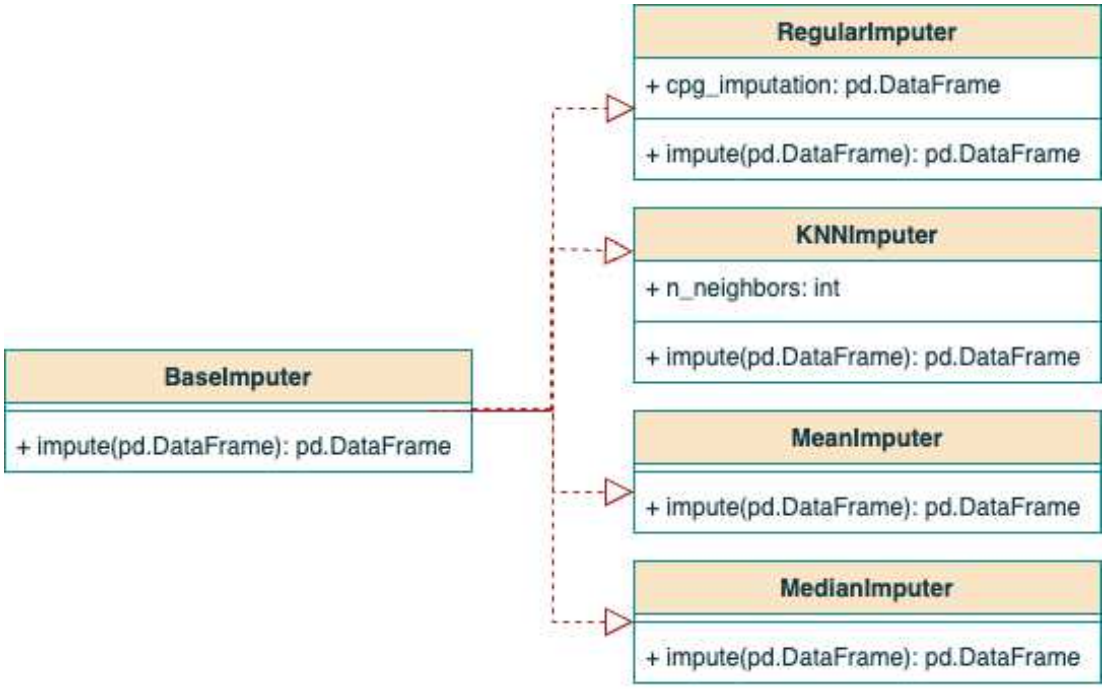
Ako se prilikom pokretanja nekog od satova metoda imputacije ne zada unaprijed, standardno je svakom satu dodijeljena regularna imputacijska metoda. Svaka od imputacijskih metoda, kao i kod implementacije podržanih epigenetičkih satova, implementira vršni razred *BaseImputer* koji otvara sučelje za implementaciju apstraktne metode *impute* čime se efektivno utilizira oblikovni obrazac strategija prikazan na slici 4.3.

Mehanizam biranja imputacijske metode po zadanom parametru prilikom pokretanja nekog od epigenetičkih satova postignut je implementacijom oblikovnog obrasca tvornica. Na sličan način kao i imputacijske metode implementirani su i modul za učitavanje podataka *loader* te modul za bazu podataka *store*.

S obzirom na to da se u tom procesu samog proučavanja softverskog koda i prikupljanja potrebnih podataka za implementaciju radnog okvira *epygenetics* u nekoliko navrata trebalo prevoditi datoteke snimljene u formatu ekstenzije .rda u datoteke formata CSV, važno je spomenuti Python knjižnicu pyreadr za čitanje datoteka formata .rda u podatkovni tip DataFrame pandas programske knjižnice. Koristeći pyreader knjižnicu implementiran je razred *RDALoader* za učitavanje podataka iz datoteka tog formata, a uz to radni okvir pruža implementaciju razreda *CSVLoader* koji utilizira programsku knjižnicu pandas za čitanje datoteka formata .csv (engl. *comma separated value*).

Pri izradi radnog okvira *epygenetics* korištena je razvojna okolina PyCharm od distributera JetBrains, a za izradu dijagrama korištena je web-aplikacija diagram.net.

Za implementaciju brzih i učinkovitih izračuna matematičkih operacija korištenih



Slika 4.3: Dijagram vršnog razreda *BaseImputer* i osnovnih imputacijskih metoda

pri implementaciji epigenetičkih satova te ostalih pomoćnih funkcija korištena je Python knjižnica NumPy, SciPy i scikit-learn koje su uobičajen podskup knjižnica u ovom kontekstu.

S obzirom na to da se prilikom same implementacije, ali i kasnije uporabe, gotovo uvijek koriste podatci velikih memorijskih zahtjeva, za lakšu obradu i prikaz podataka prilikom automatskog testiranja korištena je Python knjižnica pandas i njezin matrični tip DataFrame.

Pri prikupljanju referentnih podataka iz knjižnice *methylCIPHER* za kasniju validaciju i testiranje implementiranog radnog okvira korišten je programski jezik R i njegova razvojna okolina RStudio.

4.3. Primjeri korištenja

Radni okvir *epygenetics* trenutno je javno dostupan u obliku repozitorija na platformi GitHub. Repozitorij je moguće povući za testiranje i korištenje, a u planu je da radni okvir bude objavljen na najpopularnijem sustavu za dohvaćanje programskih knjižnica pisanih u programskom jeziku Python imena PyPI. S tim proširenjem radni okvir će biti moguće koristiti unutar privatnih projekata i slično.

Trenutna implementacija radnog okvira izgrađena je po uzoru na programski pa-

ket *methylCIPHER* te sadrži oko 26 unaprijed implementiranih epigenetičkih satova podijeljenih u 6 skupina:

- Opći epigenetički satovi
    - Bocklandtov epigenetički sat [2]
    - Garagnanijev epigenetički sat [5]
    - Hannumov epigenetički sat [6]
    - Horvathov epigenetički sat [7]
    - Linov epigenetički sat [13]
    - Vidal-Bralo epigenetički sat [22]
    - Weidnerov epigenetički sat [23]
    - Zhangov epigenetički sat [28]
- Epigenetički satovi za procjenu biološke dobi i smrtnosti
    - PhenoAge epigenetički satovi [12] (PhenoAge, PRCPhenoAge, Non-PRCPhenoAge, HRSInCHPhenoAge)
- Epigenetički satovi za procjenu nad rakom pogođenim tkivima te staničnim diobama
    - EpiTOC epigenetički sat [25]
    - HypoClock epigenetički sat[20]
- Gestacijski i pedijatrijski procjenitelji
    - Bohlinov epigenetički sat [3]
    - Knightov epigenetički sat [9]
    - Leejevi epigenetički satovi [11] (LeeControlClock, LeeRobustClock, LeeRefinedRobustClock)
    - Mayneov epigenetički sat [15]
    - PedBE epigenetički sat [17]
- Epigenetički satovi za procjenu dobi nad specifičnim životnim navikama
    - McCartneyjevi epigenetički satovi [16] (Alcohol, BMI, Smoking)
- Epigenetički satovi za procjenu dobi nad uzorcima koji nisu iz krvi
    - Horvathov Skin&Blood epigenetički sat [8]
    - DNAmCorticalClock [19]

U nastavku neće biti prikazano korištenje svih podržanih epigenetičkih satova, no njihovi rezultati i analiza prikazani su u usporedbi rezultata s onima dobivenim iz programskog paketa *methylCIPHER*.

4.3.1. Uporaba podržanih satova

Ulazna točka implementiranog radnog okvira je funkcija main glavnog modula knjižnice. U trenutnoj implementaciji, radni okvir moguće je pokrenuti izvršavanjem naredbe:

```
python -m epygenetics.main [args]
```

Izlaz pokretanjem ulazne naredbe uz *help* prikazan je na slici 4.4. Na slici je vidljivo kako najprije definiranjem argumenta *clock* korisnik odabire jedan od unaprijed podržanih epigenetičkih satova. Ako korisnik unese ime oznake sata koji nije podržan, izlaz naredbe ispisat će mu listu oznaka podržanih satova.

```
thesis@fer: python -m epygenetics.main --help
usage: main.py [-h] -c CLOCK -d DNAM [-p PHENO] [-i] [-f IMPUTATION_FILE] [-m IMPUTATION_METHOD] [-v]

Clock Execution Script

options:
  -h, --help                                show this help message and exit
  -c, --clock CLOCK                         Name of the clock to execute
  -d, --dnam DNAM                           Path to DNA methylation betas file
  -p, --pheno PHENO                         Path to pheno file
  -i, --imputation                          Impute missing CpG values
  -f, --imputation-file IMPUTATION_FILE     Path to CpG imputation file
  -m, --imputation-method IMPUTATION_METHOD Imputation method to use
  -v, --verbose                             Show traceback if an error occurs
```

Slika 4.4: Prikaz pokretanja ulazne naredbe uz argument *help*

Svakom podržanom satu pripada unaprijed određena oznaka putem konkretne implementacije enumeracijskog tipa podatka. Jednom kada korisnik unese točnu labelu sata uz sve ostale nužne ili potrebne argumente, ispod haube djeluje mehanizam oblikovnog obrasca tvornica kao i kod odabira imputacijske metode. Zatim se argumenti parsiraju, obrađuju i odabrani sat koji tvornica vrati izvrši svoju procjenu starosti.

Osim oznake podržanog sata, nužan argument za pokretanje procjene starosti je i zapis beta vrijednosti DNK metilacije argument *dnam* ili skraćeno *d*. Kao ulaz tog argumenta predaje se putanja do datoteke koja sadrži te vrijednosti zapisane u datoteci formata CSV (engl. *comma separated values*). Nažalost, trenutna implementacija radnog okvira još uvijek ne podržava neke druge formate datoteka, no CSV format za zapisivanje beta vrijednosti DNK metilacije je najčešći.

Svi ostali argumenti naredbe su opcionalni i ako nisu predane prilikom izvršavanja naredbe pridjenuti će im se predodređene vrijednosti.

Argument *pheno* mora biti putanja do datoteke formata CSV u kojoj je zapisana informacija o nekom predodređenom fenotipu koji korisnik želi koristiti prilikom pro-

```
thesis@fer: python -m epygenetics.main --clock=UnsupportedClock
usage: main.py [-h] -c CLOCK -d DNAM [-p PHENO] [-i] [-f IMPUTATION_FILE] [-m IMPUTATION_METHOD] [-v]
main.py: error: the following arguments are required: -d/--dnam
```

Slika 4.5: Primjer izvođenja ulazne naredbe bez definirane putanje do datoteke koja sadržava DNK metilacijske beta vrijednosti

```
thesis@fer: python -m epygenetics.main --clock=UnsupportedClock --dnam='data/examples/exampleBetas.csv'
Provided clock type is not recognized. Please choose from the following list:
HRSInChPhenoAge
non_prcPhenoAge
prcPhenoAge
PhenoAge
EpiTOC
hypoClock
MiAge
Bocklandt
Garagnani
Hannum
Horvath1
Lin
VidalBralo
Weidner
Zhang
Bohlin
Knight
LeeControl
LeeRobust
LeeRefinedRobust
Mayne
PEDBE
DNAmClockCortical
Horvath2
Alcohol_McCartney
BMI_McCartney
Smoking_McCartney
An error occurred: Clock type UnsupportedClock is not implemented
```

Slika 4.6: Primjer izvođenja ulazne naredbe s labelom nepodržanog epigenetičkog sata

cjene starosti. Kao i kod datoteke u kojoj su sadržane beta vrijednosti, CSV format datoteke je trenutno jedini podržan.

Argumente *imputation-file* te *imputation-method* nije moguće definirati ako prije njih nije definiran argument *imputation* kao što je prikazano na slikama 4.7 i 4.8.

```
thesis@fer: python -m epygenetics.main --clock=Bohlin --dnam='data/examples/exampleBetas.csv' --imputation-method=knn
An error occurred: CpG Check failed and imputation is not enabled or feasible.
```

Slika 4.7: Primjer korištenja KNN imputacije bez definiranog argumenta *imputation*

```
thesis@fer: python -m epygenetics.main --clock=Bohlin --dnam='data/examples/exampleBetas.csv' --imputation
Imputation of missing CpG Values occurred for Bohlin
An error occurred: Necessary CpG is missing and no imputation data provided!
```

Slika 4.8: Primjer korištenja regularne metode imputacije bez datoteke s imputacijama

Jednom kad je taj argument definiran, argument *imputation-method* (upotreba prikazana na slici 4.9) moguće je nadjenuti jednom od unaprijed definiranih imputacijskih metoda: metoda imputacije srednjom vrijednosti, metoda imputacije medijan vrijednosti, metoda imputacije K najbližih susjeda ili regularna imputacijska metoda. Ako nijedna metoda nije definirana, a argument *imputation je podignut*, unaprijed definirana vrijednost za metodu imputacije je regularna imputacija. Argument *imputation-file* potrebno je definirati isključivo ako je odabrana metoda imputacije eksplicitno regularna ili nije zadana kao što je prikazano na slici 4.10, u suprotnom ju nije potrebno definirati.

```
thesis@fer: python -m epygenetics.main --clock=Bohlin --dnam='data/examples/exampleBetas.csv' --imputation --imputation-method=knn
Imputation of missing CpG Values occurred for Bohlin
0    270.290722
1    267.816494
2    269.495242
3    269.346981
4    270.377479
dtype: float64
```

Slika 4.9: Primjer imputacije korištenjem KNN metode

Verbose argument korisnik može definirati ako želi imati bogatiji izlaz iz radnog okvira, posebno u slučaju kakve neočekivane pogreške ili iznimke.

Na slici 4.11 prikazano je jednostavno korištenje implementacije Horvathovog epigenetičkog sata za koji nije potrebna imputacija.

```
thesis@fer: python -m epygenetics.main --clock=Bohlin --dnam='data/examples/exampleBetas.csv' --imputation --imputation-file='data/imputes/sesame_450k_median.csv'
Imputation of missing CpG Values occurred for Bohlin
0   292.193728
1   291.000691
2   291.426289
3   290.996663
4   291.456610
dtype: float64
```

Slika 4.10: Primjer regularne imputacijske metode

```
thesis@fer: python -m epygenetics.main --clock=Horvath1 --dnam='data/examples/exampleBetas.csv'
0    56.205439
1    47.457539
2    48.241724
3    51.535883
4    45.054478
dtype: float64
```

Slika 4.11: Korištenje implementacije Horvathovog epigenetičkog sata u radnom okviru *epygenetics*

4.3.2. Dodavanje novih satova u knjižnicu

Ono što ponajviše razlikuje radni okvir *epygenetics* od referentnog programskog paketa *methylCIPHER* je upravo to što radni okvir implementiran u obliku knjižnice pruža mogućnost daljnjih proširenja i lakog postavljanja novih satova uz pomoć knjižnice.

Primjerice, na ispisu 4.1 prikazano je kako korisnik može izabrat jednu od tri postojeće vrste implementiranih satova (u ovom slučaju epigenetički sat po modelu linearne regresije) te mu za konačno postavljanje samo promijeni parametre.

```
1 from typing import Optional
2 import pandas as pd
3 from epygenetics.clocks.base_clocks.regression_clock import
      RegressionClock
4
5 class CustomRegressionClock(RegressionClock):
6     def __init__(self, cpgs: Optional[pd.DataFrame] = None) -> None:
7         super().__init__("CustomRegressionClock", 'CpG', 'Coef',
      0.333, cpgs)
```

Ispis 4.1: Implementacija novog epigenetičkog sata temeljenog na modelu linearne regresije

Na ispisu 4.2 prikazano je kako unutar posebnog dijela programske potpore koristiti novi epigenetički sat. Najprije je potrebno učitati potrebne razrede za korištenje implementiranog sata prikazanog u odsječku ispisa 4.2 od linije 1 do 5. Zatim u funkciji *main*, korisnik najprije učitava potrebne podatke koristeći konkretnu implementaciju apstraktnog razreda *AbstractLoader*, a zatim instancira novu implementaciju epigenetičkog sata s učitanim potrebnim podatcima te pozivanjem metode *execute* pokreće sat kao što je i prikazano na odsječku programskog koda između linije 8 te linije 12.

```
1  import pandas as pd
2  from custom_regression_clock import CustomRegressionClock
3  from epygenetics.clocks.base_clocks.clock import Clock
4  from epygenetics.loader.loader import AbstractLoader
5  from epygenetics.loader.strategies.csv import CSVLoader
6
7  def main():
8      loader: AbstractLoader = CSVLoader()
9      cpgs: pd.DataFrame = loader.load_data('../data/examples/
       exampleCpGs.csv')
10     clock: Clock = CustomRegressionClock(cpgs)
11     dna_m: pd.DataFrame = loader.load_data('../data/examples/
       exampleBetas.csv')
12     clock.execute(dna_m)
13
14 if __name__ == '__main__':
15     main()
```

Ispis 4.2: Korištenje implementiranog epigenetičkog sata

S druge strane, korisnik je slobodan implementirati potpuno novi sat u cijelosti. U tom procesu korisnik će koristiti razrede i metode kao i pomoćne funkcije čije je učitavanje prikazano na ispisu 4.3.

```
1  from typing import Optional, Tuple, Union
2  import numpy as np
3  import pandas as pd
4  from epygenetics.clocks.base_clocks.clock import Clock
5  from epygenetics.imputers import ImputerType
6  from epygenetics.imputers.base_imputer import BaseImputer
7  from epygenetics.imputers.factory import ImputerFactory
8  from epygenetics.utils.anti_trafo import anti_trafo
9  from epygenetics.utils.trafo import trafo
```

Ispis 4.3: Učitavanje pomoćnih razreda, metoda i funkcija za implementaciju novog epigenetičkog sata

Na ispisu 4.4 prikazana je početna definicija novog epigenetičkog sata te njegov konstruktor. Korisnik je slobodan modificirati konstruktor i definiciju svakog novog sata onako kako mu najviše odgovara, što pridodaje načelu lakog rada s implementiranim radnim okvirom.

```
class CustomClock(Clock):
    def __init__(self,
                 reg_coef: float = 50.412,
                 coef_name: str = 'Coef',
                 cpgs: Optional[pd.DataFrame] = None
                 ) -> None:
        self.reg_coef: float = reg_coef
        self.coef_name: str = coef_name
        super().__init__("CustomClock", 'CpG', cpgs)
```

Ispis 4.4: Implementacija potpuno novog epigenetičkog sata

Ispis 4.5 prikazuje implementaciju metode *validate* za novi epigenetički sat. U ovom primjeru, metoda će najprije izlučiti podskup CpG otoka iz podataka beta vrijednosti DNK metilacije koji su potrebni za rad implementiranog epigenetičkog sata. Zatim se postavlja zastavica *cpg_check* koja indicira jesu li u danim podatcima doista podržane vrijednosti svih potrebnih CpG otoka. Ako postoje rupe u podatcima (neke informacije nedostaju) ili je podignuta zastavica *is_imputation*, metoda *validate* obavit će imputaciju podataka koji nedostaju koristeći željenu imputacijsku metodu.

```
def validate(self,
             dna_m: pd.DataFrame,
             is_imputation: bool = False,
             imputer_type=ImputerType.REGULAR,
             cpg_imputation: Optional[pd.DataFrame] = None
             ) -> Tuple[np.ndarray, bool]:
    if self.cpgs is None:
        raise ValueError("CpGs not loaded.")

    present_cpgs: np.ndarray = np.intersect1d(self.cpgs[self.
    marker_name], dna_m.columns)
    cpg_check: bool = len(self.cpgs[self.marker_name]) == len(
    present_cpgs)

    if not cpg_check and is_imputation:
        # Impute missing CpG values
        print(f"Imputation of missing CpG Values occurred for {self.
    name}")
        for cpg in self.cpgs[self.marker_name]:
```

```
            if cpg not in dna_m.columns:
                dna_m[cpg] = np.nan

        imputer: BaseImputer = ImputerFactory.create_imputer(
    imputer_type, cpg_imputation)
        imputer.impute(dna_m)

        present_cpgs = self.cpgs[self.marker_name].values

    return present_cpgs, cpg_check
```

Ispis 4.5: Implementacija metode *validate*

Nakon same validacije podataka slijedi i sam izračun pa je tako potrebno implementirati i metodu *calculate* implementiranog epigenetičkog sata, čija je implementacija prikazana na ispisu 4.6. Najprije se provjerava jesu li podatci ispravni te ako nisu funkcija signalizira pogrešku u postpuku. No, ako su podatci ispravni ili imputirani, provodi se algoritam procjene te kao izlaz dobivamo procjenu kronološke dobi.

```
def calculate(self,
              dna_m: pd.DataFrame,
              common_cpgs: np.ndarray,
              cpg_check: bool,
              pheno: Optional[pd.DataFrame],
              is_imputation: bool
              ) -> Union[pd.DataFrame, pd.Series]:
    if cpg_check or is_imputation:
        beta_values: pd.DataFrame = dna_m[common_cpgs]
        coefficients: pd.Series = self.cpgs.set_index(self.
    marker_name).loc[common_cpgs, self.coef_name]
        tt: np.ndarray = np.dot(beta_values, coefficients) + self.
    reg_coef
        tt: anti_trafo(trafo(tt))

        if pheno is not None:
            pheno[self.name] = tt
            return pheno
        else:
            return pd.Series(tt, index=dna_m.index)

    else:
        raise Exception("CpG Check failed and imputation is not
    enabled or feasible.")
```

Ispis 4.6: Implementacija metode *calculate*

Novi epigenetički sat čija je programska potpora prikazana na ispisima 4.3, 4.4, 4.5 i 4.6 implementiran je po uzoru na implementaciju Horvathovog epigenetičkog sata u radnom okviru *epygenetics*. Implementirani satovi razlikuju se, naravno, u skupu CpG otoka potrebnih za izračun i procjenu same kronološke starosti te se razlikuju u regresijskim koeficijentima, no oba sata implementirana su po uzoru na model linearne regresije, specifično *elastic net* model uz korištenje anti-transformacijske funkcije *anti_trafo* kako bi se podatci konačno prilagodili iz log vrijednosti stvarnoj kronološkoj dobi.

Ulazna točka svih novo-implementiranih epigenetičkih satova poziva se na isti način kao što je to prikazano na ispisu 4.2 na primjeru epigenetičkog sata implementiranog po modelu linearne regresije. Korisniku je, dakle, dovoljno na željenom mjestu pozvati metodu *execute* implementiranog epigenetičkog sata te obaviti procjenu. Izlaz procjene sata kreiranog po uzoru na regresijski model prikazan je na slici 4.12, a izlaz potpuno implementiranog sata prikazan je na slici 4.13.

```
/opt/homebrew/anaconda3/envs/epygenetics/bin/python
0    1.553232
1    1.144177
2    1.218705
3    1.330745
4    1.031189
dtype: float64
```

Slika 4.12: Prikaz izlaza epigenetičkog sata konstruiranog po modelu linearne regresije

```
/opt/homebrew/anaconda3/envs/epygenetics/bin/python
0    51.632232
1    51.223177
2    51.297705
3    51.409745
4    51.110189
dtype: float64
```

Slika 4.13: Prikaz izlaza potpuno nanovo konstruriranog epigenetičkog sata

4.4. Potpora za automatsko ispitivanje

Radni okvir *epygenetics* pruža potporu za automatsko ispitivanje programske potpore koja najviše služi u svrhu validacije unaprijed implementiranih epigenetičkih satova.

Ta validacija je ujedno i garancija da radni okvir radi kako treba te da su izlazi satova korelirani s izlazima epigenetičkih satova podržanih u programskom paketu *methylCIPHER*.

Potporu za automatsko ispitivanje uključuje skup od oko 105 unit testova od kojih 22 testa nažalost nisu u potpunosti implementirana s obzirom na trenutno stanje radnog okvira dok preostalih 83 testa zadovoljavaju trenutnu implementaciju radnog okvira.

Potporu za automatsko ispitivanje moguće je u potpunosti pokrenuti s naredbom:

```
python -m epygenetics.testing [args]
```

Naredbi je moguće dodati argumente *unit* s ciljem pokretanja isključivo implementiranih unit testova, a izlaz tako pokrenute naredbe prikazan je na slici 4.14.

```
thesis@fer: python -m epygenetics.testing --unit
============================================================================================= test session starts ==============================================================================================
collected 105 items

epygenetics/test/clocks/base_clocks/linear_test.py::LinearClockTestCase::test_calculate_cpg_check_fail_without_imputation SKIPPED (Not implemented yet)                                                  [  0%]
epygenetics/test/clocks/base_clocks/linear_test.py::LinearClockTestCase::test_calculate_with_cpg_check PASSED                                                                                          [  1%]
epygenetics/test/clocks/base_clocks/linear_test.py::LinearClockTestCase::test_calculate_with_pheno SKIPPED (Not implemented yet)                                                                       [  2%]
epygenetics/test/clocks/base_clocks/linear_test.py::LinearClockTestCase::test_check_cpgs_missing_imputation_data PASSED                                                                                [  3%]
epygenetics/test/clocks/base_clocks/linear_test.py::LinearClockTestCase::test_check_cpgs_no_imputation PASSED                                                                                           [  4%]
epygenetics/test/clocks/base_clocks/linear_test.py::LinearClockTestCase::test_check_cpgs_with_imputation PASSED                                                                                         [  5%]
epygenetics/test/clocks/base_clocks/linear_test.py::LinearClockTestCase::test_initialization PASSED                                                                                                     [  6%]
epygenetics/test/clocks/base_clocks/mean_test.py::MeanClockTestCase::test_calculate_cpg_check_fail_without_imputation PASSED                                                                            [  7%]
epygenetics/test/clocks/base_clocks/mean_test.py::MeanClockTestCase::test_calculate_with_cpg_check PASSED                                                                                               [  8%]
epygenetics/test/clocks/base_clocks/mean_test.py::MeanClockTestCase::test_calculate_with_pheno PASSED                                                                                                   [  9%]
epygenetics/test/clocks/base_clocks/mean_test.py::MeanClockTestCase::test_check_cpgs_missing_imputation_data PASSED                                                                                     [ 10%]
epygenetics/test/clocks/base_clocks/mean_test.py::MeanClockTestCase::test_check_cpgs_no_imputation PASSED                                                                                               [ 11%]
epygenetics/test/clocks/base_clocks/mean_test.py::MeanClockTestCase::test_check_cpgs_with_imputation PASSED                                                                                             [ 12%]
epygenetics/test/clocks/base_clocks/mean_test.py::MeanClockTestCase::test_initialization PASSED                                                                                                         [ 13%]
epygenetics/test/clocks/base_clocks/regression_test.py::RegressionClockTestCase::test_calculate_cpg_check_fail_without_imputation PASSED                                                               [ 14%]
epygenetics/test/clocks/base_clocks/regression_test.py::RegressionClockTestCase::test_calculate_with_cpg_check PASSED                                                                                  [ 15%]
epygenetics/test/clocks/base_clocks/regression_test.py::RegressionClockTestCase::test_calculate_with_pheno PASSED                                                                                      [ 16%]
epygenetics/test/clocks/base_clocks/regression_test.py::RegressionClockTestCase::test_check_cpgs_no_imputation PASSED                                                                                  [ 17%]
epygenetics/test/clocks/base_clocks/regression_test.py::RegressionClockTestCase::test_check_cpgs_with_imputation PASSED                                                                                [ 18%]
epygenetics/test/clocks/base_clocks/regression_test.py::RegressionClockTestCase::test_initialization PASSED                                                                                             [ 19%]
epygenetics/test/clocks/biological_age_and_mortality_predictors/hrc_in_ch_pheno_age_test.py::HRSInCHPhenoAgeClockTestCase::test_initialization PASSED                                                    [ 20%]
epygenetics/test/clocks/biological_age_and_mortality_predictors/hrc_in_ch_pheno_age_test.py::HRSInCHPhenoAgeClockTestCase::test_initialization_csv_missing PASSED                                       [ 20%]
epygenetics/test/clocks/biological_age_and_mortality_predictors/non_prc_pheno_age_test.py::NonPRCPhenoAgeClockTestCase::test_initialization PASSED                                                       [ 21%]
epygenetics/test/clocks/biological_age_and_mortality_predictors/non_prc_pheno_age_test.py::NonPRCPhenoAgeClockTestCase::test_initialization_csv_missing PASSED                                          [ 22%]
epygenetics/test/clocks/biological_age_and_mortality_predictors/pheno_age_test.py::PhenoAgeClockTestCase::test_initialization PASSED                                                                     [ 23%]
epygenetics/test/clocks/biological_age_and_mortality_predictors/pheno_age_test.py::PhenoAgeClockTestCase::test_initialization_csv_missing PASSED                                                        [ 24%]
epygenetics/test/clocks/biological_age_and_mortality_predictors/prc_pheno_age_test.py::PRCPhenoAgeClockTestCase::test_initialization PASSED                                                             [ 25%]
epygenetics/test/clocks/biological_age_and_mortality_predictors/prc_pheno_age_test.py::PRCPhenoAgeClockTestCase::test_initialization_csv_missing PASSED                                                [ 26%]
```

Slika 4.14: Prikaz pokretanja automatske potpore za isključivo implementirane UNIT testove

4.4.1. Usporedba rezultata s knjižnicom *methylCIPHER*

Pokretanje potpore za automatsko testiranje s naredbom *comparison* ispisat će izlaz u obliku tablica koji sadrže informaciju o usporedbi izlaza podržanih epigenetičkih satova radnog okvira *epygenetics* s izlazom satova podržanih u programskom paketu *methylCIPHER*.

Usporedbe izlaza podržanih satova prikazane su u tablicama 4.1, 4.2, 4.3, 4.4, 4.5, 4.6, 4.7 i 4.8 .

Iz prikazanih usporedbi vidljivo je nekoliko stvari o radnom okviru *epygenetics*. Prije svega, važno je adresirati kako većina satova kojima u ulaznim podatcima nije potrebna imputacija podataka davaju uglavnom slične izlaze kao i njihove srodne implementacije u programskom paketu *methylCIPHER*.

Izlazi epigenetičkih satova koji se u danoj usporedbi znatno razlikuju od izlaza svojih srodnih implementacija daju takve rezultate ponajviše zbog različitih metoda

Clock	epygenetics	methylCIPHER
HRSInChPhenoAge	46.175572	55.521770
	40.475084	55.125490
	45.658330	55.584410
	49.907506	55.679620
	48.901753	55.722320
PhenoAge	52.293152	52.293150
	41.058674	41.058670
	43.544603	43.544600
	43.966974	43.966970
	NaN	40.352420

Tablica 4.1: Usporedba izlaza satova za procjenu biološke dobi i smrtnosti

Clock	epygenetics	methylCIPHER
EpiTOC	0.124080	0.12408041
	0.114720	0.11472045
	0.088143	0.08814333
	0.094679	0.09467855
	0.093666	0.09366637
HypoClock	0.844387	0.8443867
	0.858785	0.8587854
	0.865991	0.8659912
	0.866919	0.8669189
	0.871404	0.8714039

Tablica 4.2: Usporedba izlaza satova za procjenu biološke dobi s obzirom na rakom pogođena tkiva te staničnu diobu

Clock	epygenetics	methylCIPHER
Bocklandt	0.482857	0.4828566
	0.452261	0.4522614
	0.476842	0.4768421
	0.549446	0.5494463
	0.465779	0.4657794
Garagnani	0.695129	0.6951288
	0.610208	0.6102083
	0.584879	0.5848794
	0.596986	0.5969861
	0.650683	0.6506833
Hannum	62.334222	62.33422
	53.901388	53.90139
	53.597325	53.59733
	59.537669	59.53767
	58.801074	58.80107
Horvath Multi-Tissue	56.205439	56.20544
	47.457539	47.45754
	48.241724	48.24172
	51.535883	51.53588
	45.054478	45.05448

Tablica 4.3: (1) Usporedba izlaza općih epigenetičkih satova

Clock	epygenetics	methylCIPHER
Lin	45.585684	47.46955
	34.323300	36.32625
	24.506547	26.39302
	29.620897	31.48122
	30.734696	32.54199
Vidal-Bralo	50.518861	27.17847
	53.530620	28.71482
	51.836565	28.46388
	52.400827	29.35216
	50.715486	28.32386
Weidner	58.265374	107.39252
	54.144983	126.26082
	53.667892	48.26884
	62.243928	104.53928
	50.704644	118.65014
Zhang	-3.013470	-3.01347
	-2.867893	-2.867893
	-2.942713	-2.942713
	-3.260276	-3.260276
	-2.93625	-2.936251

Tablica 4.4: (2) Usporedba izlaza općih epigenetičkih satova

Clock	epygenetics	methylCIPHER
Bohlin	292.193728	290.5586
	291.000691	289.3656
	291.426289	289.7912
	290.996663	289.3616
	291.45661	289.8215
Knight	46.462954	43.75382
	45.889277	43.18014
	46.121079	43.41194
	45.976312	43.26718
	46.073900	43.36476
Lee Control	18.261555	12.14854
	18.661235	12.16179
	18.133249	12.01177
	18.079320	12.04271
	17.830622	11.96609
Lee Robust	23.612459	24.95762
	23.532638	24.96283
	23.531571	24.87859
	23.601707	24.93005
	23.635202	24.92568

Tablica 4.5: (1) Usporedba izlaza gestacijskih i pedijatrijskih satova za procjenu biološke dobi

Clock	epygenetics	methylCIPHER
Lee Refined Robust	33.049577	30.02084
	33.242190	30.02200
	33.000116	29.96719
	33.027419	30.03654
	32.993148	30.08753
Mayne	11.703282	18.81865
	11.746959	19.31211
	11.992752	19.11796
	12.305900	19.33233
	12.753932	19.58006
PedBE	6.337964	8.667792
	5.263821	7.397710
	4.969639	6.868020
	5.253728	7.210960
	5.360848	7.287029

Tablica 4.6: (2) Usporedba izlaza gestacijskih i pedijatrijskih satova za procjenu biološke dobi

Clock	epygenetics	methylCIPHER
DNAmAgeCortical	43.764001	51.11870
	39.839688	47.19439
	39.089534	46.44424
	41.307839	48.66254
	40.003865	47.35857
Horvath Skin&Blood	57.080572	58.70330
	49.333591	51.05889
	48.737121	50.36209
	52.772401	54.37485
	52.522762	54.07953

Tablica 4.7: Usporedba izlaza satova za procjenu biološke dobi s uzorcima ne nužno krvne naravi

Clock	epygenetics	methylCIPHER
Alcohol McCartney	NaN	-11.77366
	NaN	-11.77065
	-12.148776	-12.14878
	NaN	-12.21418
	NaN	-12.22958
BMI McCartney	NaN	-0.3433347
	NaN	-0.6141560
	NaN	-0.6782138
	NaN	-0.3257721
	NaN	-0.5696143
Smoking McCartney	3.993508	3.993508
	4.501657	4.501657
	NaN	3.173744
	3.216788	3.216788
	NaN	4.414541

Tablica 4.8: Usporedba izlaza satova za procjenu biološke dobi s praćenjem životnih navika

korištenih prilikom imputacije podataka koji nedostaju u ulaznim podatcima. Različite imputacijske metode za srodne epigenetičke satove korištene su ponajviše iz razloga što je poprilično teško razlučiti koje točno imputacijske metode se koriste za koje točno satove u programskoj knjižnici *methylCIPHER*. U konačnici, u radnom okviru *epygenetics* korištena je ona imputacijska metoda koja daje najsličnije rezultate.

Međutim, najslabija točka implementiranih satova radnog okvira su epigenetički satovi koji u svojim izlazima, za neke od 5 primjera nad kojim je rađena usporedba svih satova, daju izlazne vrijednosti tipa NaN, to jest *not-a-number*. Pretpostavka je da se takve vrijednosti nalaze u nekim izlazima jer funkcije korištenih vanjskih knjižnica dobiju pogrešku na izlaz kada se prilikom izračuna skalarnog umnoška matrica u matricama nalaze jako male vrijednosti.

Ovaj dio implementacije ostaje za popravljanje pri budućem razvoju ovog radnog okvira.

5. Zaključak

Epigenetički satovi iznimno su napredna i revolucionarna tehnologija iz područja biomedicine koja svoj značajan zamah u razvoju doživljava u posljednjem desetljeću. Iako postoji puno vrsta epigenetičkih satova otkrivenih do danas te, iako se mnogi satovi među sobom razlikuju po razini svoje univerzalnosti ili specijaliziranosti, svi satovi pokazali su se iznimno korisnim u mnogim znanstvenim i kliničkim istraživanjima, posebno u personaliziranoj medicini s ciljem lakšeg praćenja patogeneze bolesti.

Implementirani radni okvir imena *epygenetics* temeljito opisan ovim radom ima za cilj popularizaciju rada s epigenetičkim satovima, a u tome mu mnogo pridonosi to što je pisan u programskom jeziku Python, koji je poprilično popularan i općeprihvaćen programski jezik. U prilog ovoj odluci ide to što je većina postojećih tehničkih implementacija koji pružaju podršku za rad s epigenetičkim satovima pisana u programskom jeziku R poput programskog paketa *methylCIPHER*.

Radni okvir pruža podršku s unaprijed implementiranih 26 epigenetičkih satova izgrađenih po uzoru na one koji su implementirani u programskom paketu *methylCIPHER*. Bitna razlika, a ujedno i najveća motivacija za implementiranje radnog okvira u obliku knjižnice pisane u programskom jeziku Python je ta da radni okvir pruža podršku za implementiranje potencijalnih novih epigenetičkih satova koje korisnik može osmisliti koristeći razrede i pomoćne funkcije koje knjižnica pruža.

Iako se kod nekih implementacija epigenetičkih satova izlaz u nekoj mjeri razlikuje od izlaza koje daju epigenetički satovi implementirani u programskom paketu *methylCIPHER*, većina implementiranih satova daje rezultate obećavajuće točnosti, što pokazuje da je ideja ove knjižnice poprilično na mjestu te da je njezin cilj donekle ostvaren.

U nadi da će knjižnica koristiti i pomoći znanstvenoj zajednici pri daljnjem istraživanju, nedostatci ove knjižnice nastavljat će se popunjavati u budućnosti s ciljem veće fleksibilnosti, prilagodljivosti, točnosti i preciznosti.

LITERATURA

[1] Martin J Aryee, Andrew E Jaffe, Hector Corrada-Bravo, Christine Ladd-Acosta, Andrew P Feinberg, Kasper D Hansen, i Rafael A Irizarry. Minfi: a flexible and comprehensive bioconductor package for the analysis of infinium dna methylation microarrays. *Bioinformatics*, 30(10):1363–1369, 2014.

[2] Sven Bocklandt, Wen Lin, Mary E Sehl, Francisco J Sánchez, Janet S Sinsheimer, Steve Horvath, i Eric Vilain. Epigenetic predictor of age. *PloS one*, 6(6):e14821, 2011.

[3] Jon Bohlin, Siri Eldevik Håberg, Per Magnus, Sarah E Reese, Håkon K Gjessing, Maria Christine Magnus, Christine Louise Parr, CM Page, Stephanie J London, i Wenche Nystad. Prediction of gestational age based on genome-wide differentially methylated regions. *Genome biology*, 17:1–9, 2016.

[4] Lucas Paulo de Lima Camillo, Louis R Lapierre, i Ritambhara Singh. A pan-tissue dna-methylation epigenetic clock based on deep learning. *npj Aging*, 8(1): 4, 2022.

[5] Paolo Garagnani, Maria G Bacalini, Chiara Pirazzini, Davide Gori, Cristina Giuliani, Daniela Mari, Anna M Di Blasio, Davide Gentilini, Giovanni Vitale, Sebastiano Collino, et al. Methylation of elovl 2 gene as a new epigenetic marker of age. *Aging cell*, 11(6):1132–1134, 2012.

[6] Gregory Hannum, Justin Guinney, Ling Zhao, LI Zhang, Guy Hughes, SriniVas Sadda, Brandy Klotzle, Marina Bibikova, Jian-Bing Fan, Yuan Gao, et al. Genome-wide methylation profiles reveal quantitative views of human aging rates. *Molecular cell*, 49(2):359–367, 2013.

[7] Steve Horvath. Dna methylation age of human tissues and cell types. *Genome biology*, 14:1–20, 2013.

[8] Steve Horvath, Junko Oshima, George M Martin, Ake T Lu, Austin Quach, Howard Cohen, Sarah Felton, Mieko Matsuyama, Donna Lowe, Sylwia Kabacik, et al. Epigenetic clock for skin and blood cells applied to hutchinson gilford progeria syndrome and ex vivo studies. *Aging (Albany NY)*, 10(7):1758, 2018.

[9] Anna K Knight, Jeffrey M Craig, Christiane Theda, Marie Bækvad-Hansen, Jonas Bybjerg-Grauholm, Christine S Hansen, Mads V Hollegaard, David M Hougaard, Preben B Mortensen, Shantel M Weinsheimer, et al. An epigenetic clock for gestational age at birth based on blood methylation data. *Genome biology*, 17:1–11, 2016.

[10] Dmitrii Kriukov, Evgeniy Efimov, Ekaterina A Kuzmina, Ekaterina E Khrameeva, i Dmitry V Dylov. Computagebench: Epigenetic aging clocks benchmark. *bioRxiv*, stranice 2024–06, 2024.

[11] Yunsung Lee, Sanaa Choufani, Rosanna Weksberg, Samantha L Wilson, Victor Yuan, Amber Burt, Carmen Marsit, Ake T Lu, Beate Ritz, Jon Bohlin, et al. Placental epigenetic clocks: estimating gestational age using placental dna methylation levels. *Aging (Albany NY)*, 11(12):4238, 2019.

[12] Morgan E Levine, Ake T Lu, Austin Quach, Brian H Chen, Themistocles L Assimes, Stefania Bandinelli, Lifang Hou, Andrea A Baccarelli, James D Stewart, Yun Li, et al. An epigenetic biomarker of aging for lifespan and healthspan. *Aging (albany NY)*, 10(4):573, 2018.

[13] Qiong Lin, Carola I Weidner, Ivan G Costa, Riccardo E Marioni, Marcelo RP Ferreira, Ian J Deary, i Wolfgang Wagner. Dna methylation levels at individual age-associated cpg sites can be indicative for life expectancy. *Aging (Albany NY)*, 8(2):394, 2016.

[14] Ake T Lu, Austin Quach, James G Wilson, Alex P Reiner, Abraham Aviv, Kenneth Raj, Lifang Hou, Andrea A Baccarelli, Yun Li, James D Stewart, et al. Dna methylation grimage strongly predicts lifespan and healthspan. *Aging (albany NY)*, 11(2):303, 2019.

[15] Benjamin T Mayne, Shalem Y Leemaqz, Alicia K Smith, James Breen, Claire T Roberts, i Tina Bianco-Miotto. Accelerated placental aging in early onset preeclampsia pregnancies identified by dna methylation. *Epigenomics*, 9(3):279–289, 2017.

[16] Daniel L McCartney, Robert F Hillary, Anna J Stevenson, Stuart J Ritchie, Rosie M Walker, Qian Zhang, Stewart W Morris, Mairead L Bermingham, Archie Campbell, Alison D Murray, et al. Epigenetic prediction of complex traits and death. *Genome biology*, 19:1–11, 2018.

[17] Lisa M McEwen, Kieran J O'Donnell, Megan G McGill, Rachel D Edgar, Meaghan J Jones, Julia L MacIsaac, David Tse Shen Lin, Katia Ramadori, Alexander Morin, Nicole Gladish, et al. The pedbe clock accurately estimates dna methylation age in pediatric buccal cells. *Proceedings of the National Academy of Sciences*, 117(38):23329–23335, 2020.

[18] Dolors Pelegí-Sisó, Paula De Prado, Justiina Ronkainen, Mariona Bustamante, i Juan R González. methylclock: a bioconductor package to estimate dna methylation age. *Bioinformatics*, 37(12):1759–1760, 2021.

[19] Gemma L Shireby, Jonathan P Davies, Paul T Francis, Joe Burrage, Emma M Walker, Grant WA Neilson, Aisha Dahir, Alan J Thomas, Seth Love, Rebecca G Smith, et al. Recalibrating the epigenetic clock: implications for assessing biological age in the human cortex. *Brain*, 143(12):3763–3775, 2020.

[20] Andrew E Teschendorff. A comparison of epigenetic mitotic-like clocks for cancer risk prediction. *Genome Medicine*, 12:1–17, 2020.

[21] Kyra L Thrush, Albert T Higgins-Chen, Zuyun Liu, i Morgan E Levine. R methylcipher: a methylation clock investigational package for hypothesis-driven evaluation & research. *biorxiv*, stranice 2022–07, 2022.

[22] Laura Vidal-Bralo, Yolanda Lopez-Golan, i Antonio Gonzalez. Simplified assay for epigenetic age estimation in whole blood of adults. *Frontiers in genetics*, 7: 126, 2016.

[23] Carola Ingrid Weidner, Qiong Lin, Carmen Maike Koch, Lewin Eisele, Fabian Beier, Patrick Ziegler, Dirk Olaf Bauerschlag, Karl-Heinz Jöckel, Raimund Erbel, Thomas Walter Mühleisen, et al. Aging of blood can be tracked by dna methylation changes at just three cpg sites. *Genome biology*, 15:1–12, 2014.

[24] Zongli Xu, Liang Niu, Leping Li, i Jack A Taylor. Enmix: a novel background correction method for illumina humanmethylation450 beadchip. *Nucleic acids research*, 44(3):e20–e20, 2016.

[25] Zhen Yang, Andrew Wong, Diana Kuh, Dirk S Paul, Vardhman K Rakyan, R David Leslie, Shijie C Zheng, Martin Widschwendter, Stephan Beck, i Andrew E Teschendorff. Correlation of an epigenetic mitotic clock with cancer risk. *Genome biology*, 17:1–18, 2016.

[26] Ahrim Youn i Shuang Wang. The miage calculator: a dna methylation-based mitotic age calculator of human tissue types. *Epigenetics*, 13(2):192–206, 2018.

[27] Qian Zhang, Costanza L Vallerga, Rosie M Walker, Tian Lin, Anjali K Henders, Grant W Montgomery, Ji He, Dongsheng Fan, Javed Fowdar, Martin Kennedy, et al. Improved precision of epigenetic clock estimates across tissues and its implication for biological ageing. *Genome medicine*, 11:1–11, 2019.

[28] Yan Zhang, Rory Wilson, Jonathan Heiss, Lutz P Breitling, Kai-Uwe Saum, Ben Schöttker, Bernd Holleczek, Melanie Waldenberger, Annette Peters, i Hermann Brenner. Dna methylation signatures in peripheral blood strongly predict all-cause mortality. *Nature communications*, 8(1):14617, 2017.

Sažetak

Ovaj rad bavi se razvojem i primjenom radnog okvira *epygenetics* za razvoj i laku integraciju epigenetičkih satova. U kratkom pregledu povijesnog razvoja epigenetičkih satova detaljno su opisane neki od postojećih epigenetičkih satova. Također su opisani matematički modeli na kojim počiva većina do sada otkrivenih epigenetičkih satova. Pregled do sada postojećih tehničkih implementacija koji na bilo koji način podržavaju rad s analizom DNK metilacije ili epigenetičkih satova, uvod je u detaljan opis implementacije radnog okvira u obliku Python knjižnice po uzoru na programski paket *methylCIPHER* pisanog u programskom jeziku R. Implementirani radni okvir imena *epygenetics* omogućuje jednostavnu i intuitivnu integraciju različitih epigenetičkih satova sa svrhom podržavanja i popularizacije rada s epigenetičkim satovima. Cilj rada bio je stvoriti alat koji je fleksibilan, lako proširiv, koji nije inertan na promjene i koji je jednostavan za korištenje, čime se olakšava rad s velikim skupovima podataka i omogućuje preciznija procjena biološke starosti. Na kraju rada opisane su korištene tehnologije prilikom razvoja ovog radnog okvira, kao i upute za korištenje radnog okvira. radni okvir ima ugrađenu potporu za automatsko testiranje čiji su rezultati iskorišteni za evaluiranje knjižnice čime je pokazana visoka usklađenost s rezultatima dobivenim korištenjem *methylCIPHER* programskog paketa.

Ključne riječi: CpG područje, DNK metilacija, nukleotid, dinukleotid, Python, epygenetics, methylCIPHER, Horvath, model, epigenetika, epigenetički sat, radni okvir

Framework for Age Estimation Using Epigenetic Clocks Library

Abstract

This thesis focuses on the development and application of the *epygenetics* framework for the creation and easy integration of epigenetic clocks. A brief overview of the historical development of epigenetic clocks is provided, with detailed descriptions of some existing epigenetic clocks. The thesis also explores the mathematical models underlying most of the currently known epigenetic clocks. A review of existing technical implementations that support DNA methylation analysis or epigenetic clocks in any way serves as an introduction to a detailed description of the implementation of the framework in the form of a Python library modeled after the *methylCIPHER* package written in the R programming language. The implemented framework, named *epygenetics*, allows for simple and intuitive integration of various epigenetic clocks to support and popularize the use of epigenetic clocks. The goal of this work was to create a tool that is flexible, easily extensible, non-robust, and simple to use, thereby facilitating work with large datasets and enabling more accurate biological age estimation. The final section of the thesis describes the technologies used in the development of this framework, as well as instructions for its use. The framework includes built-in support for automated testing, the results of which were used to evaluate the library, demonstrating high consistency with results obtained using the *methylCIPHER* package.

Keywords: CpG site, DNA methylation, nucleotide, dinucleotide, Python, epygenetics, methylCIPHER, Horvath, model, epigenetics, epigenetic clock, framework