SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

DIPLOMSKI RAD br. 323

PROGRAM ZA AUTOMATSKU ANALIZU PROTEIN KODIRAJUĆIH GENA

Ana Bulović

Zagreb, srpanj 2012.

Mnogi su put već oni opjevani

(makar od strane mene)

ali i ovi put ne mogu da odolim, a da ih ne spomenem.

Fala

Mamuli, Papuli,

Ivana, Lukaka,

Marijoot!

Sadržaj

Uvod

Evolucija svoj rad zasniva na nekolicini jednostavnih načela, poput mutacije, nasljeđivanja i prirodne selekcije, čija je opetovana primjena kroz dugi niz godina rezultirala širokim spektrom organizama koje se danas može pronaći na Zemlji, s novima u nastajanju. Usprkos razlikama koje se mogu vidjeti među živim svijetom, sličnosti su, iako manje uočljive na prvi pogled, velike. Usporedi li se 3.7 milijardi godina, za koje se pretpostavlja da je prošlo otkako je evolucija započela na Zemlji, sa 80-ak milijuna godina kad se pretpostavlja da je došlo da razdvajanja između čovjeka i miša, jasno je da se među tim organizmima može očekivati velika sličnost. Sličnost je najjasnije vidljiva na razini gena i proteina, što je postalo okosnica mnogih grana biologije, poput komparativne proteomike i genomike. Najpoznatiji model organizmi su tako postali Danio Rerio i Mus Musculus, proučavanje kojih služi boljem razumijevanju funkcioniranja čovjeka.
Povećanjem količine dostupnih sekvenciranih genoma (29 Mammalian Genomes Project) kao rezultat smanjenja cijene korištene tehnologije te podataka o strukturi proteina, potrebno je učiniti te informacije dostupnima na pregledan i sistematičan način. Ono što se zasad može smatrati obradom podataka dostupnih iz genoma je, među ostalim, obilježavanje regija koje sadrže gene, te u genima dijelove koji se prevode u protein. Za većinu vrsta nije dostupno mnogo dokaza na proteinskoj ili cDNK razini o postojanju pojedinog proteina, ali upravo zbog velike sličnosti među vrstama moguće je koristiti proteine ili cDNK druge vrste da bi se pronašao mogući gen.

Problem obilježavanja gena u čitavom genomu nipošto nije jednostavan. Potrebno je prvo pronaći na genomu gene koje odgovaraju poznatim proteinima te nakon toga ispravno identificirati dijelove gena koji se prevode u protein. Ovaj dio nije jednostavan zbog tog što se samo 2% genoma prevodi u proteine, i potrebno je, poetično govoreći, pronaći iglu u plastu sijena. Odnos veličina je sljedeći: ljudski genom (Human Genome project) je veličine 3,194 milijuna nukleotida, prosječni ljudski gen dug je oko 3,000 nukleotidnih baza, od kojih se u prosjeku 1,500 prevodi u protein. Drugi problem je priroda dostupnih genoma, čija kvaliteta još uvijek nije na razini koja bi ih učinila korisnima u analizi proteina. U ovom radu opisana

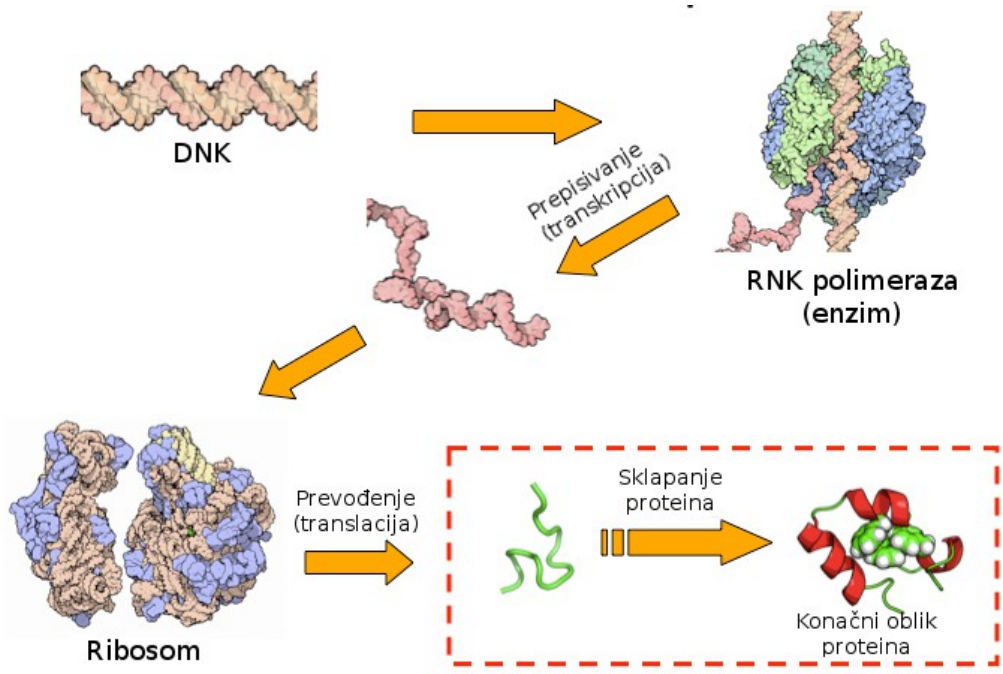
je metoda kojom je moguće postići bolju kvalitetu obilježavanja gena koristeći sličnost vrsta i optimalno poravnanje, iz kog je razloga manje ostjetljiva na šum izazvan greškama u sekvenciranju i slaganju genoma od trenutno dostupnih metoda koje se većinom oslanjaju na heurističke, a ne optimalne, metode poravnanja.

1. Genom, gen i protein

Svaka stanica, bilo više, bilo jednostaničnog organizma, sadrži sve genetske informacije o čitavom organizmu. Informacija je kodirana u molekuli DNK (DeoksiriboNukleinska Kiselina) koja je stastavljena od slijeda nukleotida. U DNK mogu se pronaći četiri vrste nukleotida označenih slovima – A (adenin), T (timin), G (guanin) i C(citozin). Svo naslijeđe jedinke, kodirano u DNK (ili RNK u slučaju virusa) naziva se genom. Veličina genoma proteže se od samo 2000 nukleotida kod nekih virusa do 149 milijardi nukleotida za slučaj vrste *Paris Japonica*, zasad navećeg pronađenog genoma. Za usporedbu, ljudski genom sastoji se od 3.4 milijarde nukleotida.

DNK je nositelj informacije, a mehanizmi kojima stanica obavlja potreban rad, koji među ostalim reguliraju njenu smrt, kretanje, rast i reakcije na vanjske podražaje u molekule koje nastaju prepisivanjem informacije iz molekule DNK. Te molekule su razne vrste RNK molekula (mRNK, tRNK, rRNK) i proteini. Dok je RNK sačinjena od nukleotida (s tom razlikom da umjesto timina ima uracil), protein se sastoji od aminokiselina. Način na koji protein nastaje iz molekule DNK opisan je procesom koji je u molekularnoj biologiji popularno znan kao središnja dogma.

1.1. Središnja dogma



Sl. 1: Središnja dogma biologije - lanci DNK molekule se razdvajaju te se ona prepisuje u molekulu mRNK. mRNK se nakon naknadne obrade prevodi u protein, koji potom poprima svoju konačnu strukturu

Da bi se lakše razumjela problematika obilježavanju protein-kodirajućih regija na genomu, potrebno je ukratko opisati proces nastanka proteina iz molekule DNK. Kratkim pojašnjenjem tog procesa postaje jasno koje su moguće zapreke pri pronalasku regija na genomu koje kodiraju proteine i koja je priroda preslikavanja DNK u protein. Također, poznavanje ovog procesa pomaže razumijevanju dostupnih dokaza za postojanje proteina i njihovo korištenje za pronalazak DNK niza koji ih kodira. Najprije, potrebno je pojasniti pojmove gena, eksona, introna i proteina.

Gen je dio lanca DNK koji sadrži "signale" potrebne da ga se prepozna kao nosioca informacije o proteinu, to jest da se na tom mjestu podijeli DNK lanac i započne prepisivanje u lanac RNK.

Gen je sačinjen od eksona i introna koji se redaju naizmjence, počevši i završivši eksonom. I eksoni i introni se prevode u RNK. Lanac RNK koji sadrži i introne i eksone naziva se nezrela

mRNK. Osim introna i eksona, gen sadrži i prethodno spomenute signale koji reguliraju proizvodnju proteina u pojedinoj stanici. Ti signali nisu binarne prirode (protein se ili prevodi ili ne) već adaptivno u reakciji s proteinima i drugim tvarima u stanici potiču ili sprečavaju proizvodnju proteina.

Nakon prevođenja eksona i introna u nezrelu mRNK dolazi do takozvanog srastanja. Srastanjem se iz molekule nezrele mRNK uklanjaju dijelovi koji neće činiti budući protein. To su regije početnih eksona (UTR – untranslated region) te introni.

Jedan gen može imati više rezultirajućih proteina na način da se prilikom srastanja iz pre-mRNK uklone i pojedini eksoni. Uklanjanjem različitih eksona nastaju različiti proteinski lanci. Ova se mogućnost naziva alternativno srastanje.

Lanac RNK iz kog je uklonjen sav materijal koji se neće prevesti u protein se naziva zreli mRNK. Kako su abecede DNK i RNK gotovo iste, nastanak RNK iz DNK lanca naziva se prepisivanje. Kako je abeceda proteina drugačija i sačinjava ju 20 aminokiselina, nastanak protiena iz RNK naziva se prevođenje.

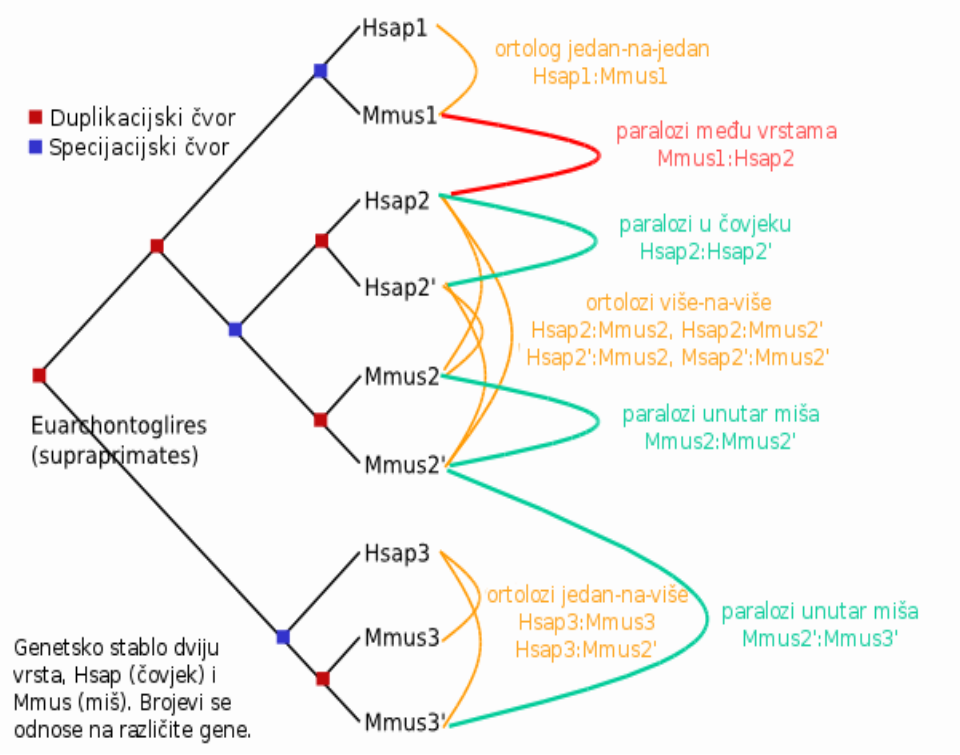
Za svaki slijed od tri nukleotida postoji odgovarajuća aminokiselina. Broj mogućih kombinacija od tri nukleotida jednak je $4^3 = 64$. Kako je različitih aminokiselina samo 20, preslikavanje iz DNK u protein je surjektivno. Drugim riječima, uz poznat proteinski slijed, nije moguće sa sigurnošću ustvrditi koji je bio izvorni slijed DNK iz kog se preveo dani protein, već će takvih mogućih izvornih slijedova biti više.

Osim proteina, mRNK se može prevesti u lanac cDNK, koji je točan slijed nukleotida iz kog je nastala mRNK, ali bez introna i UTR-a. Značaj ove molekule u kontekstu ovoga rada je taj što se za mnoge pretpostavljene proteine kao dokaz postojanja uzima dokaz postojanja cDNK koja ih kodira. U većini slučajeva ovo je istina, ali ne mora biti – cDNK može biti jedini proizvod iz mRNK, bez odgovarajućeg proteina.

Direktna povezanost ova tri podatka – gena, mRNK i proteina da naslutiti da bi jedan od njih bio dovoljan da se dobije slika o ostala dva. Poznavajući gen, može li se znati kako će izgledati rezultirajući protein? Poznavajući mRNK ili cDNK, može li se pronaći gen na genomu iz kog se prepisuju? U sljedećem su potpoglavlju opisani podaci s kojima danas biolozi i bioinformatičari raspolažu i neke od grešaka koje se mogu očekivati kada se radi s biološkim podacima.

1.2. Homologija

Ono što se može očekivati u prirodi je da ona ne odbacuje uspješna rješenja. Funkcijske jedinice stanice, proteini, odnosno geni koji ih kodiraju, se iz tog razloga nalaze pod velikim selekcijskim pritiskom. Selekcijski pritisak podrazumijeva da većina jedinki vrste koje imaju mutaciju na genu zbog koje će protein postati manje ili potpuno nefunkcionalan, neće preživjeti embrionalnu fazu. Težnja očuvanja genetskog slijeda nije prisutna samo u jedinkama iste vrste.



Sl. 2: Prikaz mogućih odnosa homologije među genima: ortolozi i paralozi (jedan na jedan, jedan na više, više na više)

Ono što može značajno olakšati analizu bioloških podataka iz jedne vrste je prethodno znanje o njoj srodnim vrstama. Ta sličnost među vrstama koja je posljedica toga što su nastale iz zajedničkog pretka naziva se homologija. O homologiji se može pričati i na razini gena i proteina i tada se kaže da su dva ili više proteina (gena) homolozi. Kada je riječ o proteinima (genima), razlikuju se dvije vrste homologa: ortolozi i paralozi. Geni u različitim vrstama nastali kao rezultat specijacije (razdvajanja vrsta iz zajedničkog pretka) nazivaju se

ortolozima. Geni iste vrste koji su nastali duplikacijom (umetanje kopije gena na novo mjesto u genomu) nazivaju se paralozi. Na slici 2 mogu se vidjeti odnosi među genima koji su nastali kao rezultat specijacije i duplikacije.

Budući da su nužno rezultat specijacije, ortolozi se uvijek odnose na barem dvije različite vrste. Iako nije nužno, ortolozi često obavljaju istu ili sličnu funkciju u različitim vrstama. S druge strane, iako su nazivno rezultat duplikacije, paralozi ne moraju biti u istoj vrsti. Ako rezultat duplikacije slijedi specijacija kojom se u oba organizma pojavljuju oba gena (izvorni i rezultat duplikacije), onda je riječ o paralozima u različitim vrstama. Paralozi često obavljaju slične funkcije. Ipak, zbog smanjenoj selekcijskog pritiska na duplicirani gen, on brže mutira i može obavljati nove funkcije.

Mnogo je bilo pokušaja utvrđivanja homologije korištenjem bioinformatičkih alata. Nekolicina bioloških baza nudi rezultate predviđanja odnosa homologije, kao što su primjerice Ensembl (EnsemblCompara GeneTrees [21]), COGs [14], eggNOG [21], InParanoid [24], OrthoDB [25], OrthoMCL [26] i OrthoMam [27].

2. Obilježavanje genoma

2.1. Motivacija

Ljudski je genom prvi puta objavljen 2003. godine, kao rezultat projekta čije se trajanje proteglo preko jednog desetljeća i sa cijenom od tri milijarde dolara. Od tada je sekvencirano više desetaka eukariotskih genoma te su se pojavili brojni projekti s ciljem obilježavanja na stotine, tisuće, čak i desetne tisuća genoma. Primjerice, cilj Genome 10k projekta [10] je sekvencirati 10000 genoma kralješnjaka. Projekt 1000 Genomes [11] usmjeren je na mapiranje i identificiranje razlika među ljudskim genomima. 29 Mammals projekt [12] započet je s ciljem bolje interpretacije genoma kroz njihovu usporedbu. Količina dostupnih "sirovih" informacija raste sve brže (s padom cijene sekvenciranja). Za ručnu obradu svih dostupnih genoma – analiza kodirajućih i nekodirajućih regija, homologije, strukture i funcije proteina, bio bi potreban dugi niz godina. Razvojem prikladnih bioinformatičkih alata ovaj težački posao može se olakšati, ubrzati te dobrim dijelom automatizirati.

Informacije koje potencijalno možemo dobiti iz genoma odnose se, primjerice, na pronalazak regulativnih nekodirajućih regija, kodirajućih regija i rezultirajućih proteina (njihove strukture i funkcije). Informacije koje dobivamo uspredbom genoma mogu nam ukazati na evolucijske mehanizme uključene u razvoj vrsta, te dijelova posebice interesantnih za kralješnjake poput prilagodljivog imunološkog sustava i ljudima nadasve interesantnog neurološkog sustava. Time dobivamo bolju sliku odnosa između genotipa – onog što je kodirano u genomu i fenotipa – obilježja kog uočavamo u jedinci.

Najčešći pristup automatiziranom obilježavanju genoma je sljedeći – na osnovu poznatih proteinskih, cDNK ili mRNK slijedova pronaći njhove polazišne regije na genomu korištenjem algoritama za poravnanje sljedova. Najveći problem obilježavanja genoma danas je nepostojanje prikladnih alata koji mogu obrađivati genome brzinom kojom oni pristižu iz alata za sekvenciranje. Iz tog se razloga u većim projektima obilježavanja genoma koriste heuristički bioinformatički alati koji su zbog svoje brzine prikladni za velike količine podataka kakve se susreću u genomu. Heuristički alati o kojima je riječ koriste se za

usporedbu sljedova, postupak koji je okosnica obilježavanja genoma. Heuristički alati za poravnanja sljedova koji se najčešće koriste u obilježavanju genoma su razne inačice BLAST-a [13]. Korištenjem heurističkog, za razliku od optimalnog poravnanja, žrtvuje se preciznost. U vječnoj bitci između preciznosti i brzine izvođenja, kad je genom u pitanju, pobjedu odnosi brzina. Rezultat toga su predviđeni proteinski sljedovi lošije kvalitete, što direktno utječe na mogućnost otkrivanja homologije među vrstama, strukture i funkcije proteina te usporedbe proteina u različitim vrstama

2.2. Ensembl projekt obilježavanja genoma

Jedan od većih poduhvata s ciljem sistematičnog obilježavanja genoma poduzet je od strane Ensembl tima ([1], [2]). Kako se ovaj rad djelimično oslanja na njihov proces obilježavanja, te nudi poboljšanja za isti, u ovom će poglavlju ukratko opisan niz koraka koje oni koriste s napomenama o mogućim pogreškama koje svaki od koraka potencijalno unosi.

Prvi korak (imenom *Raw Computes*) uključuje procesiranje sirovih genoma – iz slijeda DNK se, među ostalim, predviđaju moguće lokacije gena korištenjem Genscan alata [5]. Ono što će Genscan alat dati kao izlaz su lokacije kodirajućih regija na genomu i rezultirajući proteini.

Ostala dva koraka *(Targetted Stage, Similarity Stage)* služe se dokazima na razini cDNK ili proteina iz same vrste (iz za to referentnih baza), kao i njoj srodnih vrsta za pronalazak protein-kodirajućih regija na genomu. Za ovo se oslanjaju na sličnost nizova.

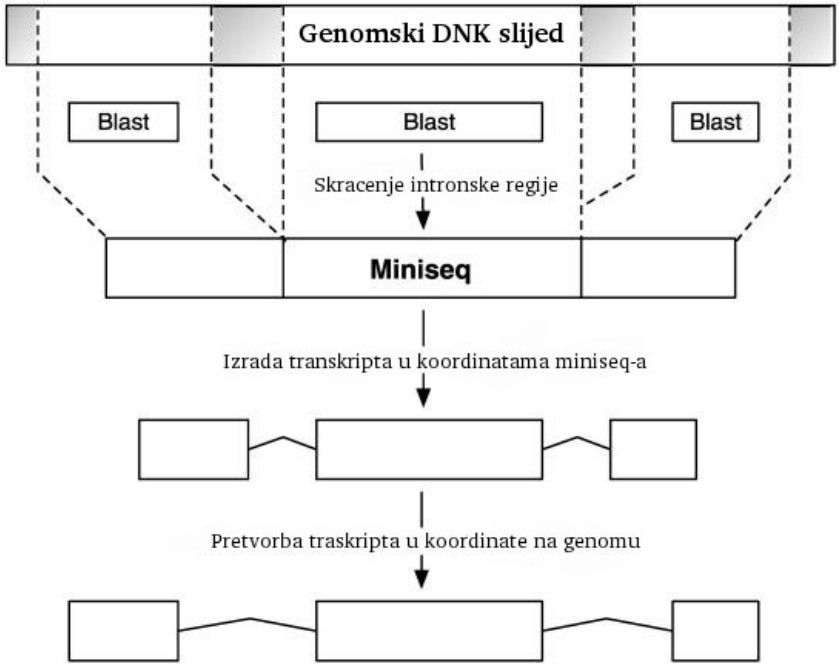
Drugi korak (*Targetted Stage*) koristi poznate cDNK i proteinske slijedove iz vrste za koju se radi obilježavanje genoma. Ako postoji značajna sličnost između proteina predviđenog Genscan alatom i već poznatog proteina vrste u nekoj od referentnih baza (RefSeq, UniprotKB), odbacuje se rezultat predviđanja i svrhu obilježavanja koristi se poznati protein. Ovi su proteini obilježeni kao *pep:known* u Ensembl anotaciji.

Za proteine koji nisu uspješno obilježeni u drugom koraku provjerava se postoji li sličan protein u srodnim vrstama. Ako je pronađen protein zadovoljavajuće sličnosti sljeda s predviđenim proteinom, odbacuje se rezultat predviđanja i taj se protein koristi za daljne obilježavanje gena. Ovi su proteini obilježeni kao *pep:novel* u Ensembl notaciji.

Uvjet da predviđeni protein bude zamjenjen ili proteinom iz vrste ili proteinom srodne vrste je da je na referentnim bazama proteina obilježen klasom dokaza o postojanju proteina PE 2 [8]. To znači da za njegovo postojanje postoji dokaz na proteinskoj ili cDNK razini.

Proteini pronađeni Genscan alatom za koje ne postoji odgovarajući protein iz dotične vrste ili njoj srodnih vrsta odbacuju se u daljnjoj obradi, ali ih je moguće preuzeti s Ensembl baze. Obilježeni su kao *pep:genscan* u Ensembl notaciji.

U procesu pronalazaka gena i eksona se, osim proteina, koriste i cDNK sljedovi. Da bi se pomoću poznate cDNK ili proteina došlo do gena, potrebno je na genomu pronaći regiju koja im odgovara po slijedu. U Ensembl projektu pronalazak regije od interesa radi se pomoću alata pmatch (Durbin, neobjavljeno). Rezultat pmatch-a smatra se grubom procjenom lokacije gena. Ono što je sljedeće potrebno učiniti je locirati eksone i introne u toj grubo izdvojenoj



Sl. 3: Generiranje Miniseq slijeda - regija gena predviđena pomoću pmatch alata se dodatno smanjuje poravnavanjem proteinskog slijeda na predviđenu regiju čime je moguće grubo lokalizirati eksone. Skraćenjem intronskih regija se smanjuje veličina slijeda koju Genewise mora obraditi.

regiji. Za to se koristi alat GeneWise [7] zasnovan na HMM-u (skriveni Markovljevi modeli). Problem obrade sekvence pomoću GeneWise alata je njegova iznimna sporost. Iz tog se razloga prva, gruba lokacija pronađenog gena dodatno smanjuje.

Način na koji se obavlja smanjenje slijeda koji će se obraditi vidljiv je na slici 3. Proteinski slijed vrste čiji se genom obilježava ili protein iz srodne vrste se poravna na predviđenu regiju gena korištenjem tBLASTn alata u slučaju proteina, ili BLASTn alata u slučaju cDNK. Ovo rezultira grubim lokacijama eksona, koje se naknadno prošire 200 nukleotida s obje strane u svrhu modeliranja intronskih regija. Tako generiran slijed naziva se *miniseq* i služi kao ulaz u GeneWise. Izlaz iz GeneWise alata su točne lokacije eksona i introna, predviđanje kojih je zasnovano na HMM modelu koji koristi metapodatke o karakterističnim signalnim sljedovima na početku i kraju gena, eksona i introna. Za eukariotske se gene pretpostavlja da je velik broj takvih signalnih sljedova još uvijek nepoznat.

Problem ovoga pristupa je upravo korak u kom se regija od interesa smanjuje upotrebom tBLASTn alata. Zbog heurističke prirode alata, ovaj je korak podložan pogrešci (ne prepoznavanje dijela sekvence) što može rezultirati propuštanjem pojedinih eksona, posebice onih kraćih.

3. Podaci

3.1. Dostupni biološki podaci

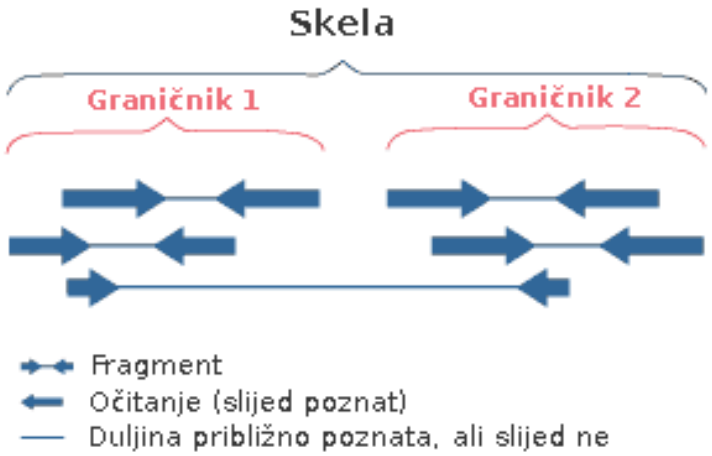
Biološki podaci koji su danas dostupni kreću se od potpunih genoma, dijelova genoma, cDNK i mRNK nizova, proteinskih nizova do proteinskih struktura koje oni tvore u stanici. Baze podataka koje nude informacije o genomu su Ensembl (obilježeni genomi brojnih kralješnjaka i eukariota) te brojne specijalizirane baze za pojedine organizme: FlyBase *(Drosophila melanogaster* model organizam), EcoCyc (*E.coli*), Wormbase (*C.Elengans*). Među bazama koje nude proteinske nizove su UniProt, SwissProt, Pfam itd. Od baza na kojima se mogu preuzeti podaci o strukturi proteina najpoznatija je PDB (Protein DataBank).

U ovom radu korišteni su genomi 36 eukariotskih vrsta preuzeti sa Ensembl baze podataka. Popis korištenih vrsta može se naći u dodatku C. Za svaku od vrsta dostupni su na Ensembl stranicama podaci o sekvenciranju i sklapanju genoma. Da bi se razumjelo s kakvim se podacima radi, potrebno je ukratko pojasniti ta dva postupka.

Sekvenciranje DNK je očitanje nukleotidnih baza iz dijelova DNK. Točnost očitanja te duljina niza koji se očitava ovise o korištenoj tehnologiji. Prosječne duljine očitanih fragmenata kreću se od 25 za nove, do 1000 nukleotida za starije tehnologije. Kako je prosječna duljina gena (kod čovjeka) oko 3000 nukleotida, jednim se nizom ne obuhvaća čitav gen. Nove tehnologije su jeftinije i brže, ali su fragmenti kraći od onih dobivenih korištenjem starijih tehnologija, premda se ta situacija brzo mijenja. Nakon sekvenciranja dostupi su podaci o slijedovima DNK, ali ne i o njihovim točnim lokacijama, kao ni o njihovom položaju u odnosu na druge očitane sljedove. Drugi parametar koji određuje kvalitetu genoma je pokrivenost – parametar koji opisuje u koliko se prosječno fragmenata može očekivati da je svaki od nukleotida sadržan. Što je veća pokivenost i duljina fragmenata, može se očekivati bolja kvaliteta rezultirajućeg genoma. Idealan rezultat sklapanja genoma je ispravan poredak svih očitanja – potrebno je pronaći ispravan redoslijed i smjer fragmenata, što postaje sve teže što su fragmenti kraći i što je manje redundancije (manja pokrivenost).

Oba postupka, sekvenciranje i sklapanje genoma, unose pogrešku koja može sezati od nedostajućeg ili pogrešno očitanog nukleotida do nedostajućih čitavih regija.

Od 38 vrsta koje su korištene, 12 ima genom sklopljen u kromosome s visokom pokrivenošću sekvenciranja, dok su ostale rezultat sekvenciranja niske pokrivenosti i za njih podaci o kromosomima nisu dostupni.



Sl. 4: Skela (scaffold) se sastoji od graničnika (contig) i praznina među graničnicima čije se duljine procjenjuju poznavanjem očekivane duljine fragmenta. Graničnici su sljedovi nukleotida rekonstruirani s visokom razinom pouzdanosti.

Na taj način sekvencirani genomi se sklapaju u takozvane skele (scaffold). Primjer skele može se vidjeti na slici 4, i sastoji se od graničnika (contig) koji su regije u kojima je redoslijed nukleotida pouzdano rekonstruiran i od praznina. Duljina praznina procjenjuje se iz poznate očekivane duljine fragmenta koji se sekvencira. U ovom postupku rekonstrukcije postoje mnogi mogući problemi, i broj im se povećava što je očitani fragment kraći. Najveći problem predstavljaja sekvenciranje *de novo* genoma – bez referentnog genoma na kog bi se sljedovi mogli mapirati [3].

3.2. Lokalna kopija Ensembl baze

Lokalna kopija Ensembl baze sastoji se od FASTA datoteka. Njihovo je nazivlje objašnjeno je u README datotekama koje se nalaze u svakoj od mapa. Ukratko, ako je riječ o DNK sljedovima, ime datoteke bit će:

```
<species>.<assembly>.<release>.<sequence type>.<id type>.<id>.fa
```

Sequence type može biti:

- *dna* – nemaskirani DNK slijed
- *dna_rm* – maskirani DNK slijed. Maskirane su regije niske složenosti i regije s ponavljanjima. Maskirane regije obilježene su slovima N umjesto izvornim slijedom.

Id type može biti:

- *chromosome* – sljedovi DNK složeni u kromosome. ID se odnosi samo na kromosome.
- *nonchromosomal* – sljedovi DNK koji nisu dodijeljeni nijednom kromosomu. Na neke vrste ovo su jedini dostupni podaci.

Ako je riječ o proteinskim sljedovima, ime datoteke će biti:

```
<species>.<assembly>.<release>.<sequence type>.<status>.fa
```

Sequence type će uvijek biti *pep.*

Status može biti:

- *all* – svi proteini koji pripadaju klasi *pep:known* ili *pep:novel* klasi za tu vrstu.
- *abinitio* – svi proteini koji su rezultat Genscan predviđanja gena (i odgovarajućih proteina)

Fasta datoteke formatirane su korištenjem alata *formatdb* [18]. Za dohvat slijedova iz generiranjih baza korišten je alat *fastacmd* [18]. Eskoni se ne mogu dohvatiti preko lokalnije kopije Ensembl-a, već sa udaljenje baze korištenjem Biomart alata [17]. Potrebno je napomenuti da nije moguće za sve proteine dohvatiti odgovarajuće eksone. Oni su dostupni isključivo za proteine klasa *pep:known* i *pep:novel.* Iz tog su razloga za proteine klase *pep:genscan* naknadno predviđene lokacije eksona upotrebom Genewise alata.

U radu s eksonima s Ensembl baze podataka, treba znati da su oni nekad ispresijecani kratkim regijama umetanja koje Ensembl naziva intronima pomaka okvira (*frameshift introns)* [19]. Riječ je o greškama u sekvenciranju koje umeću nekolicinu nukleotida tamo gjde se oni zapravo ne nalaze. Ovo remeti redoslijed prepisivanja DNK u protein i zbog toga su ove greške u Ensembl bazi označene kao introni. Ovaj pomak može također biti rezultat mutacije

u jedinki vrste koja je sekvencirana. Intron pomaka okvira može biti dugačak 1, 2, 4 ili 5 nukleotida.

3.2.1. Ensembl klasifikacija proteina i gena

Kako Ensembl nije baza koja se primarno bavi proteinima, već obilježavanjem genoma, podaci o proteinima dostupni na Ensembl bazi su preuzeti sa za to referentnih baza, kao što su SWISS-PROT/TrEMBL [30] i NCBI RefSeq [31]. Genscan alatom se predviđaju lokacije na čitavom genomu vrste.

Pri obilježavanju vrste, s referentnih baza proteina se preuzmu proteini te vrste i mapiraju na genom. Ovi proteini imaju oznaku *pep:known*, te njima korespondentni geni također imaju oznaku *known*. Iz skupa proteina predviđenih Genscan-om se uklanjaju oni koji slijedom i lokacijom odgovaraju ovim proteinima. Za ostatak proteina se traži ortolog u srodnim vrstama. Ako se pronađe ortolog, onda se taj protein označava kao *pep:novel*, a njemu korespondentni gen također statusom *novel*.

Proteini još mogu imati i status *genscan* što znači da su predviđeni Genscanom, ali takav protein još nije poznat ni za izvornu vrstu, ni za njoj srodne vrste. Geni mogu imati status *merged* što znači da je taj gen rezultat Ensembl cjevovoda za obilježavanje, ali i da je ručno obilježen od Havana tima [4]. Od korištenih vrsta ručne anotacije dostupne su samo za čovjeka i miša.

3.3. Korišteni formati

3.3.1. Format proteinskih i nukleotidnih sljedova

FASTA format je tekstualni format često korišten u bioinformatici, koji služi za predstavljanje poptidnih ili nukleotidnih sljedova. Sastoji se od zaglavlja koje započinje znakom ">", nakon čega slijedi identifikator slijeda. U zaglavlju se, osim identifikatora, mogu nalaziti i druge informacije. Različite baze koriste različite formate zaglavlja. Informacije su, ipak, najčešće razdvojene znakom "|". Primjer ove konvencije je:

```
>2642682|2642746|ENSAMET00000014187|ENSAMEE00000136184|1
GGTTCAAGGAATTTTTCTGCAAACAGTTCTAAGAGCAGTACAGCCAGAACTGGTGGCTTTCTCCT
```

3.3.2. Format izlaznih datoteka poravnanja

Kako su u aplikaciji korištena dva tipa poravnanja, riječ je o dva izlazna formata.

Za poravnanja generirana pomoću BLAST alata, izlazni format je XML jeziku. Specifikacije ovog izlaznog formata mogu se naći na [20]. XML jezik je nadasve prikladan za raščlanjivanje teksta. Kako se format BLAST izlaznih datoteke često mijenja, postupak raščlanjivanja testa potrebno je napraviti sa što više robusnosti. Za to se pokazao najbolji XML format, među ostalim zbog već postojećih alata za automatsko učitavanje rezultata poravnanja u ovom formatu. Ovim je datotekama dan nastavak *blastout*.

Alat korišten za optimalno poravnanje SW# ima jedan predefinirani izlazi format. Ovim datotekama je dan nastavak *.swout*. SW# ima jednostavan i intuitivan izlazni format, opisan u [9].

4. Metode

4.1. Korišteni alati

4.1.1. SW#

SW# [9] alat je implementacija Smith-Waterman algoritma za poravnanje sljedova. Smith-Waterman [28] je algoritam optimalnog lokalnog poravnanja sljedova s prazninama. Optimalno poravnanje znači da će od svih mogućih poravnanja, pod zadanim uvjetima (matrica zamjena i cijena umetanja / brisanja), biti pronađeno ono najbolje. Ovo ne mora značiti da je to poravnanje biološki najznačajnije. Implementacija algoritma zasniva se na algoritmu dinamičkog programiranja. U ovom konkretnom slučaju, algoritam je implementiran na grafičkim karticama s CUDA arhitekturom. Iz tog razloga ova implementacija nudi značajno ubrzanje u odnosu na standardnu implementaciju na središnjoj procesnoj jedinici. Ubzranje postaje tim veće što su veće duljine sljedova koji se poravnanjavaju.

4.1.2. BLAST

BLAST [13] je alat za usporedbu bioloških (proteinskih i DNK) sljedova. BLAST ne koristi optimalno poravnanje, to jest usporedbu na razini čitavih nizova, već heurističkom metodom locira kratka podudaranja između dva slijeda. Za nukleotidne sljedove ta podudaranja su najčešće duljine 11 baza, dok za proteine taj broj najčešće iznosi 3. Te riječi duljine k uzimaju se slijedno iz niza. Primjerice, ako je slijed aminokiselina SMCRRL uz $k = 3$, riječi s kojima će se raditi usporedba bit će SMC, MCR, CRR i RRL. Izbjegavanjem poravnanja čitavih nizova se vrijeme izvođenja značajno smanjuje. Iz pronađenih poravnanja ciljnog slijeda i na ovaj način izdvojenih riječi od k slova, uklanjaju se ona koja su ocijenjena rezultatom koji je ispod predefiniranog praga kvalitete T. Ova mala poravnanja se nakon tog proširuju u HSP-ove (High Scoring Pairs), koji su parovi poravnanja visoke kvalitete (score) koji se nalaze u predefiniranom susjedstvu. Nakon ocjene broja i kvalitete HSP-ova, regije poravnanja se lokalnim poravnanjem proširuju na dva ili više HSP-ova. Za svako poravnanje se računa e[1] vrijednost, te se odbacuju ona kojima je ona veća od zadane vršne e vrijednosti.

U ovoj je aplikaciji korišten alat naredbene linije *blastall v2.2.21*. Ovaj alat nudi sljedeće vrste poravnanja:

- DNK upit na DNK bazu (blastn)
- proteinski upit na proteinsku bazu (blastp)
- proteinski upit na bazu prevedene DNK (tblastn)
- upit prevedene DNK na proteinsku bazu (blastx)
- upit prevedene DNK na bazu prevedene DNK (tblastx)

4.1.3. Genscan

Genscan ([5], [6]) je alat zasnovan na modelu skrivenih Markovljevih lanaca za predviđanje gena i njima odgovarajućih proteinskih proizvoda na genomu. Omogućava predviđanje lokacija ekson – intron regija na slijedu genomske DNK. Izlaz iz programa je, osim predviđenih lokacija, proteinski proizvod. Ne omogućava predviđanje *alternative splicing*-a.

[1]Detalji o *e* vrijednosti BLAST algoritma mogu se naći u [53] i [54]

Ovo je trenutno jedan od najboljih programa [29] za predviđanje ekson-intron strukture i Ensembl tim ga koristi za inicijalnu fazu cjevovoda *(Raw computes*) u kojoj se predviđaju geni na čitavom genomu.

4.1.4. Genewise i Exonerate

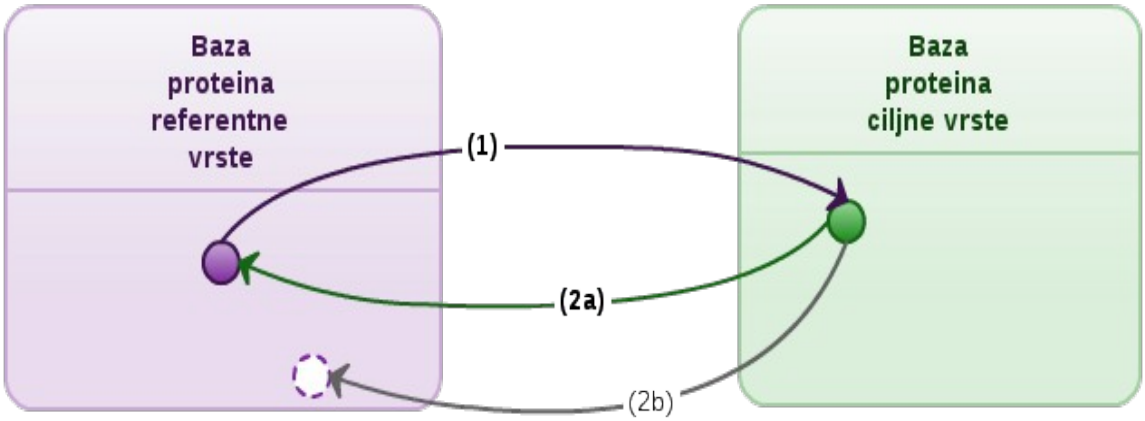
Ponešto drugačiji problem je predviti lokacije eksona i introna ako je poznat proteinski proizvod gena ili ortologni proteinski proizvod iz srodne vrste. Potrebno je pronaći lokacije na lancu DNK koje odgovaraju sličnom proteinskom proizvodu, te ujedno zadovoljavaju modele eksona i introna. U ovu svrhu korišten je program Genewise [7], alat zasnovan na skrivenim Markovljevim modelima. Za model se može reći da je dvoslojan: između poznatog proteina i DNK slijeda postoji sloj koji odgovara "predviđenom" proteinu. DNK slijed se prevodi u predviđeni protein, a ovaj predviđeni protein je takav da u najvećoj mogućoj mjeri pod zadanim uvjetima (modeli eksona i introna) odgovara poznatom proteinu.

Exonerate[Error: Reference source not found] je alat koji nudi rješenje za najveći problem Genewise-a, a to je njegovo veliko vrijeme izvršavanja. Oba programa napisana su korištenjem Dynamite [34] prevodioca koji služi jednostavnom prevođenju skrivenih Markovljevih modela u algoritme dinamičkog programiranja (izlaz je C kod).

4.2. Metoda pronalaska ortologa

Za svaki proteina teoretski je moguće tražiti ortologe u drugim vrstama. Ortologija, kao i većina toga u biologiji, nije potpuno jasno i razriješeno pitanje, što hoće reći da ne postoji popis svih ortolognih proteina među različitim vrstama. Razlog tomu je što je ortologiju teško ustvrditi. Mnogo je načina predviđanja ortologije, ali problem svih pristupa je što ne postoji temeljna dokazana istina s kojom bi se mogli usporediti njihovi rezultati. Jedan primjer referentne baze za ustvrđivanje ortologa među proteinima je KOG [14], ali je broj dostupnih vrsta samo sedam, od kojih je samo čovjek kralješnjak i time je za potrebe ovog rada neiskoristiva.

Pristup koji se koristi u ovom radu je poznat pod imenom BLAST RBH (*Reciprocal Best Hit*) i sastoji se od dva koraka, kao što se može vidjeti na slici 5.



Sl. 5: Ilustracija RBH-a. (1) Protein iz referentne vrste se BLASTp-om poravna na sve proteine ciljane vrste. (2a) Protein s najboljim rezultatom poravnanja se poravna na sve proteine referentne vrste. Rezultat je izvorni protein i pretpostavlja se ortologija. (2b) Rezultat nije izvorni protein. Pretpostavka o ortologiji se odbacuje.

1. Protein iz referentne se korištenjem BLASTp alata pokuša poravnati na sve dostpune proteine iz ciljane vrste.
2. Protein iz ciljane vrste koji daje najbolje poravnanje se korištenjem BLASTp-a pokuša poravnati na sve proteine iz referentne vrste (dakle, čovjeka).

Ako je protein iz referentne vrste koji je dao najbolje poravnanje ujedno i izvorni protein, onda se za taj par proteina pretpostavlja da su ortolozi. Ovaj pristup oslanja se isključivo na sličnost proteinskih sljedova, što se pokazalo kao dovoljno dobra metoda ustvrđivanja ortologije [15], posebice uzevši u obzir njenu jednostavnost. Zanimljiva opaska na račun RBH-a može se naći u [16].

4.2.1. Prevođenje poravnanja s prazninama u protein

Jedan od problema prilikom razvoja aplikacije ukazao se s potrebom dobivanja proteinskog slijeda iz poravnanja na razini DNK. Za srodnu vrstu (iz koje su uzete informacije o eksonima) na raspolaganju stoji proteinski slijed. Poravnanje eksona srodne vrste na regiju gena ciljne vrste može izgledati primjerice ovako:

```
16859  CTGGCTGCAAAAATATCATCCCTG-----AGAGTGTTAAGACATTGCTGT  16903
 1315  CCGG------TCATACGAATGCTGTTCCAACAGATGCAAGAAATTGCTAT  1358

16904  TAAAAATAAAATTCTGGCAAAGTGTCTCAGGACTTCAAAAGGGACAGAC-  16952
 1359  TAAAATTAAAATGCCAGCGAACTATCTGAGGGCTTCAAAAGGGACAGACC  1408

16952  -----TG--------------TAAACTGTCAAGAACATATCAATAAAAGT  16983
 1409  ATAATTGCAAATGAGGCCTGATAAACTGCCAAGAATATATCACCCAAAAA  1458
```

Iz ovog se primjera može vidjeti mogući problemi pri prevođenju ovakvog slijeda u protein – praznine se javljaju u oba slijeda. Ono što praznine uzrokuju je promjena okvira. Okvir diktira od koje nukleotidne baze počinje prevođenje u protein. Kako se trojke baza prevode u jednu aminokiselinu, jasno je da nije svejedno od koje baze će započeti prevođenje. Ovo se može vidjeti na jednostavnom primjeru: ako je nukleotidni slijed

`ATTCTGGCAAAGTGTCTCAGGACTTCAAA,`
prijevodi će biti:

- za okvir 0 (počinje se od prve baze) ILAKCLRTS
- za okvir 1 (prva baza se preskače, prevođenje počinje od T) FWQSVSGLQ
- za okvir 2 (prve dvije baze se preskaču, prevođenje počinje od drugog T) SGKVSQDFK

Promjenom okvira očigledno se dobiju posve različiti proteinski sljedovi. Svaka od praznina u poravnanju može uzrokovati promjenu okvira prevođenja. Zbog tog je potrebno nakon svake praznine ponovno ustanoviti ispravan okvir.

Pretpostavka je da se eksoni iz referentne vrste uvijek mogu ispravno prevesti u već poznat protein. Zbog tog taj već poznat protein služi kao predložak za prevođenje DNK ciljne vrste u protein. Kada se za određenu regiju iz poravnanja pronađe ispravan okvir za prevođenje u već poznat protein, tada se njoj poravnata regija također može prevesti u protein. Ako se u toj regiji pak nalaze praznine, one će u proteinu biti prevedene u X.

4.3. Opis zadatka

Zadatak ovog rada bio je razviti aplikaciju koja na sistematičan način nudi informaciju o proteinskom slijedu upotrebom homologije među srodnim eukariotskim vrstama. Za razliku od pristupa korištenog od strane Ensembl tima, opisanog u prethodnom poglavlju, lokacije eksona se ne predviđaju korištenjem proteinskog proizvoda srodne vrste već eksona iz te vrste koji kodiraju dotični protein. Korištenjem optimalnog poravnanja povećava se mogućnost pronalaska protein kodirajućih regija, posebice u slučajevima kada je riječ o genomu niske pokrivenosti[2].

Postoje slučajevi kada dio poravnanja, usprkos očiglednoj sličnosti sljedova, nedostaje zbog heurističke prirode korištenih alata. Kao alternativa tom pristupu, u ovom radu koristi se SW# alat koji nudi implementaciju Smith-Waterman algoritma na grafičkim karticama s CUDA arhitekturom. Zahvaljujući ovoj implementaciji, koja nudi značajno ubrzanje u usporedbi sa brzinom izvođenja na središnjoj procesnoj jedinici, moguće je gotovo eliminirati jedini argument protiv korištenja optimalnog poravnanja – njegova sporost. Izvođenje je još uvijek sporije od izvođenja heurističkih algoritama, ali preciznost nije žrtvovana.

Opis posla može se opisati u devet koraka:

1. Definiranje referentnih vrsta za svaku vrstu. Referentne vrste su one čiji su genomi dobro sekvencirani i obilježeni (u trenutnoj verziji aplikacije kao referentna vrsta koristi se samo čovjek)
2. Sklapanje ispitnog skupa. Proteini koji čine ispitni skup su ih proteini iz referentnih vrsta koji zadovoljavaju određene uvjete opisane u nastavku teksta.
3. Pronalazak ortologa za proteine iz ispitnog skupa u ostalim vrstama pomoću RBH pristupa.
4. Dohvat potrebnih podataka za daljnju obradu – proteinskih sljedova, regija gena, proširenih regija gena te eksona sa Ensembl baze.
5. Poravnanja sljedova. Uključuje četiri vrste poravnanja:

[2]Niska pokrivenost se odnosi na pokrivenost manju od 2x.

- poravnanje proteinskog slijeda iz referentne vrste na spojene eksone vrste (preuzete s Ensembl-a) korištenjem tBLASTn-a
- poravnanje eksona referentne vrste na proširenu regiju gena vrste korištenjem BLASTn-a
- poravnanje spojenih eksona referentne vrste (takozvani cDNK) na eksone vrste koji su preuzeti sa Ensembl-a koristeći
- poravnanje eksona referentne vrste na proširenu regiju gena vrste koristeći optimalno poravnanje (SW# aplikacija).

Prva tri poravnanja služe za usporedbu dobivenih rezultata i za analizu slučajevima u kojima su pronađeni pojedini eksoni koji nisu obilježeni na Ensembl-u.

6. Postanaliza poravnanja koja uključuje odbacivanja poravnanja koja nisu u pravom redoslijedu (pretpostavka je da se eksoni moraju pojavljivati slijedno u poravnanju) te uklanjanje preklapajućih poravnanja.
7. Rekonstrukcija proteinskog slijeda iz postignutih poravnanja. Nukleinski slijed dobiven poravnanjem eksona referentne vrste na proširenu regiju gena prevodi se u protein. Način na koji je ovo ostvareno opisan je naknadno u tekstu.
8. Generiranje statistika koje sadrže informaciju o tome koliki je postotak pojedinog eksona pronađen svakim od opisana četiri tipa poravnanja.
9. Generiranje poravnanja tri proteina – proteina iz referentne vrste, proteina iz vrste iz Ensembl baze te proteina koji je dobiven na način opisan u koraku X. Ovo služi olakšanoj vizualnoj inspekciji dobivenog rezultata.

Ukratko, ideja je popraviti informaciju o proteinskom slijedu koja je trenutno dostupna. Jednu vrstu uvijek uzimamo kao referentnu. To znači da proteine te vrste uzimamo kao ispravne, kao i obilježene gene iz kojih nastaju. Za drugu vrstu koja ima ortologan protein cilj je pokušati poboljšati, ili može se reći, "popraviti" proteinski niz sa informacijom koja se nalazi u genomu, ali nije pronađena zbog načina na koji se genom pretražuje. Mana dostupnih metoda pretraživanja genoma je to što se za poravnanja koriste algoritmi s heuristikom koji ne garantiraju pronalazak najboljeg poravnanja. Ovaj nedostatak posebice postaje očigledan ako je riječ o genomu koji je nekvalitetno sekvenciran. U ovom radu se ne radi takav kompromis,

već se koristi optimalno poravnanje. Eksoni iz referentne vrste poravnaju se na regiju gena ciljne vrste i iz te se informacije rekonstruira protein. U dosta slučajeva ovakvim se pristupom mogu popraviti greške koje su rezultat zlosretne kombinacije lošeg sekvenciranja i heurističkih algoritama poravnanja.

5. Implementacija

5.1. Aplikacija SuperExonRetriever2000

Postupak opisan u poglavlju 4.2 implementiran je u aplikaciji kreativno nazvanoj SuperExonRetriever2000. Implementacija je u programskom jeziku Python, a izvorni kod može se preuzeti sa internet stranica https://github.com/abulovic/SuperExonRetriver2000/tree/master/ExoLocator. Sve potrebno za instalaciju i konfiguraciju aplikacije nalazi se na Wiki stranicama aplikacije (https://github.com/abulovic/SuperExonRetriver2000/wiki/Installation-Instructions).

Funkcijski se aplikacija može podijeliti na dva dijela: prvi dio, ili **podatkovni cjevovod** (data pipeline), služi dohvaćanju i generiranju svih podataka potrebnih za daljnju obradu. U ovom dijelu aplikacije se obavljaju koraci 3, 4 i 5 kao što su opisani u poglavlju 4.3 (dohvaćanje podataka s Ensembl baze i generiranje potrebnih poravnanja). U drugom se dijelu aplikacije, ili **cjevovodu za analizu** (analysis pipeline), obavljaju koraci od 6. do 9., dakle postanaliza poravnanja, rekonstrukcija proteinskog slijeda iz poravnanja, generiranje statistike i poravnanja proteina. Podatkovni cjevovod nalazi se u direktoriju *pipeline*, dok se cjevovod za analizu nalazi u direktoriju *data_analysis*.

Osim ova dva funkcijski zasebna dijela programa, u vršnom se direktoriju aplikacije može naći i direktorij imena *utilities*. U tom se direktoriju mogu pronaći pomoćne klase i funkcije koje se koriste u ostatku programa, s funkcionalnostima većinom usmjerenima na učitavanje konfiguracije aplikacije i unificiranom načinu pristupa datotekama. Dostupne klase iz *utilities* paketa su:

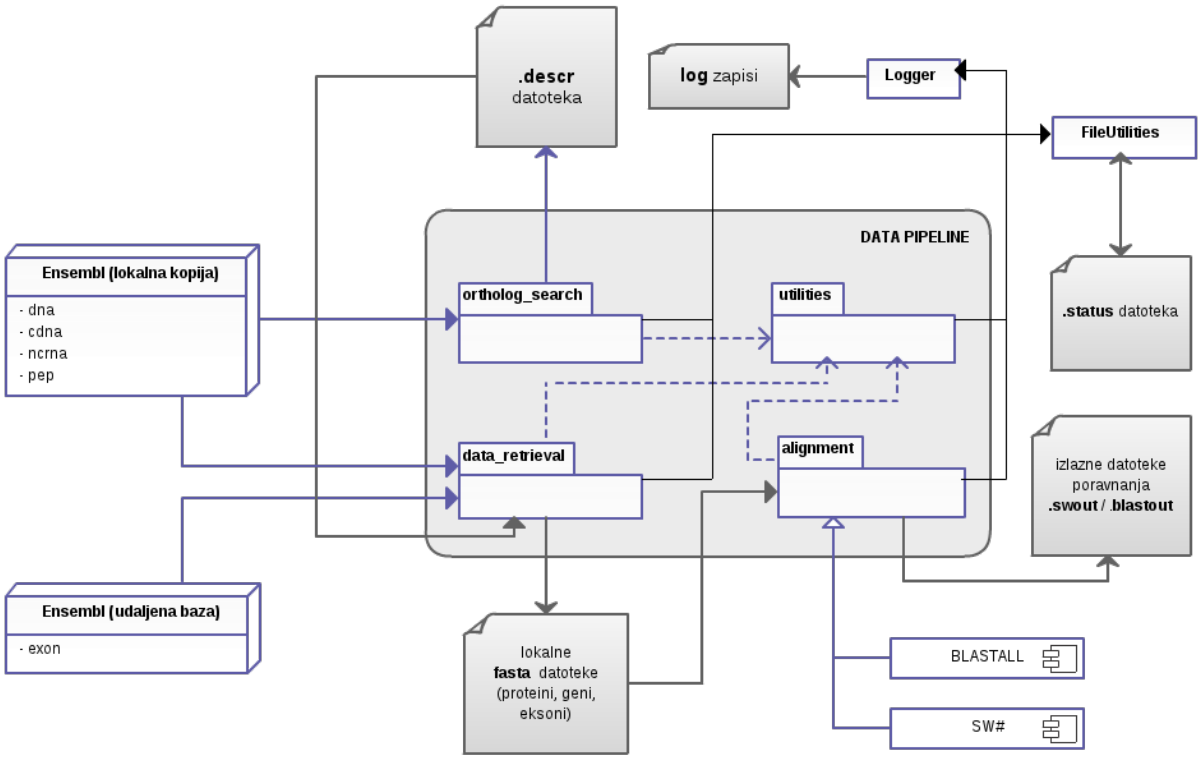
- **ConfigurationReader**: implementiran kao singleton oblikovni obrazac. Služi dinamičkom učitavanju konfiguracija iz svih konfiguracijskih datoteka (nastavak .cfg) dostupnih u *cfg* direktoriju (osim datoteke logging.cfg)
- **DescriptionParser**: Služi učitavanju datoteka iz direktorija proteina s popisom ortologa u srodnim vrstama. Nudi informacije o svakom od pronađenih ortolognih

proteina: mogu se dohvatiti sve dostupne informacije ili zasebno samo identifikatori proteina, regije gena, pronađene vrste ili informacije o lancu DNK (1 ili -1).

- **FileUtilities**: Pristup datotekama koje koriste oba cjevovoda kao što su popis identifikatora proteina iz referentne vrste (ispitni skup), popis vrsta za koje je potrebno tražiti ortologe, te *.status* datoteke u direktorijima proteina. Čitanje i pisanje u fasta datoteke.
- **Logger**: nudi unificirano sučelje za bilježenje mogućih grešaka pri radu sustava. Učitava konfiguracijsku datoteku *logging.cfg* smještenu u *cfg* direktorij aplikacije. Ovisno o konfiguracijskoj datoteci, nudi različite formate zapisa u dnevnike do kojih se putevi također definiraju u spomenutoj konfiguracijskoj datoteci.

5.2. Podatkovni cjevovod

Podatkovni cjevovod služi dohvatu svih podataka potrebnih za kasniju analizu poravnanja i sintezu proteina iz generiranih poravnanja. Kao ulaz u podatkovni cjevovod prvenstveno služi lokalna kopija Ensembl baze, na kojoj se nalaze podaci o čitavim genomima i proteinima. Podatke o eksonima potrebno je dohvatiti s udaljene baze korištenjem Biomart sučelja (pristup je ostvaren pomoću *BiomartRemoceAccess.pl* skripte). Izlaz iz podatkovnog cjevovoda su *.descr* datoteke (Dodatak A), potom lokalne kopije gena, proteina i eksona u fasta datotekama i konačno, potrebna poravnanja. Implementacija svakog od ovih koraka bit će pobliže opisana u nastavku poglavlja. Kao što je prikazano na slici 6, podatkovni se cjevovod sastoji se od četiri modula – *utilities, ortholog_search, data_retrieval* i *alignments*. Osim *utilities* modula, svi ostali moduli komuniciraju isključivo preko datoteka koje generiraju i učitavaju. Ovime su koraci cjevovoda potpuno odvojeni i moguće ih je pokretati zasebno. Ovakav je pristup od velike važnosti pri izgradnji cjevovoda, posebice ako je riječ o



Sl. 6: Podatkovni cjevovod sastoji se od četiri modula: *ortholog_search, data_retrieval, alignment* i *utilities.* Ovi moduli služe pronalasku ortologa ljudskih proteina u drugim vrstama te dohvatu svih potrebnih podataka za daljnju analizu (sljedovi proteina, gena, proširenih gena i eksona) te generiranju potrebnih poravnanja. Svi moduli zapisuju status vlastitog izvršavanja u .status datoteke, te rezultate izvršavanja u log zapise.

biološkim podacima, s kojima je nemoguće predvidjeti sve iznimke koje se mogu dogoditi. Ovakvim je dizajnom moguće iterativno mijenjati implementaciju svakog od modula u odgovoru na nove zahtjeve, bez potrebe za ponovnim izvršavanjem već odrađenog posla. Osim toga, svaki od modula zapisuje rezultate izvršavanja u zaseban dnevnik pomoću kojeg se onda mogu iščitati mogući problemi i iznimke.

5.2.1. Statusne datoteke

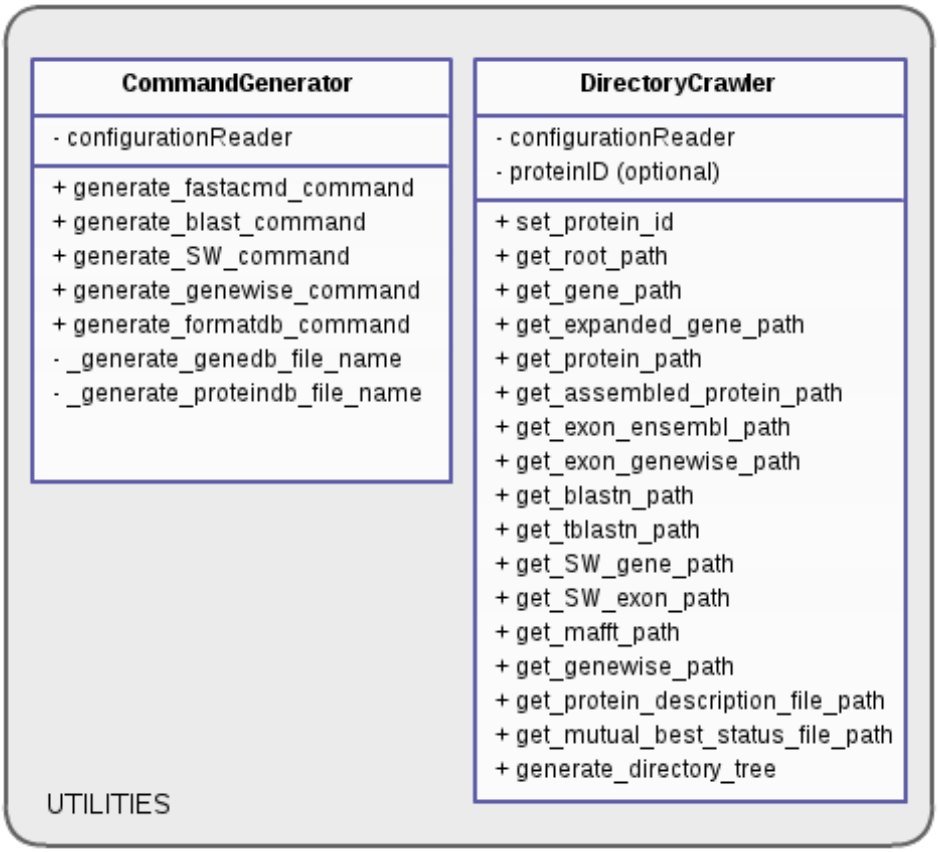
Svaki od modula zapisuje status izvršavanja u .status datoteku koja se nalazi u svakom od direktorija proteina. Statusi mogu biti OK, PARTIAL ili FAILED. OK status znači da se korak uspješno obavio za sve vrste. PARTIAL status znači da postoje neke vrste za koje nije bilo moguće izvršiti određeni korak. Ovisno o koraku koji se izvršava, u prikladnom se direktoriju nalazi popis vrsta za koje je izvršavanje bilo neuspješno. FAILED status znači da korak nije uspio ni za jednu vrstu. Moguće je odabrati da se pojedini modul ponovno pokrene samo za one proteine za koje nije ocijenjeno da je uspješno izvršen. U slučaju PARTIAL statusa, to znači da će se korak izvršavati samo za one vrste za koje nije uspio u prethodnoj iteraciji, dok u slučaju FAILED statusa to znači da će se korak ponoviti za sve vrste.

Popis statusnih varijabli nalazi se u konfiguracijskom direktoriju aplikacije (cfg) i zove se *status_file_keys.txt.* Pregledom statusne datoteke proteina dobiva se sadržajan uvid u uspješnost izvođenja svakog od koraka aplikacije.

5.2.2. utilities modul

Utilities modul podatkovnog cjevovoda služi:

- generiranju naredbi alata naredbenog retka koji se koriste u aplikaciji, a to su blastall, sw#, mafft i genewise.
- Dohvatu apsolutnih puteva do datoteka iz direktorija proteina. Na slici 7 može se vidjeti popis svih dostupnih puteva. Iz konfiguracijske se datoteke čita apsolutni put direktorija s rezultatima, na što se dodaje identifikator proteina, te konačno potrebni direktorij (podaci, poravnanja). Za detaljnu sliku svih direktorija, pogledati Dodatak B.

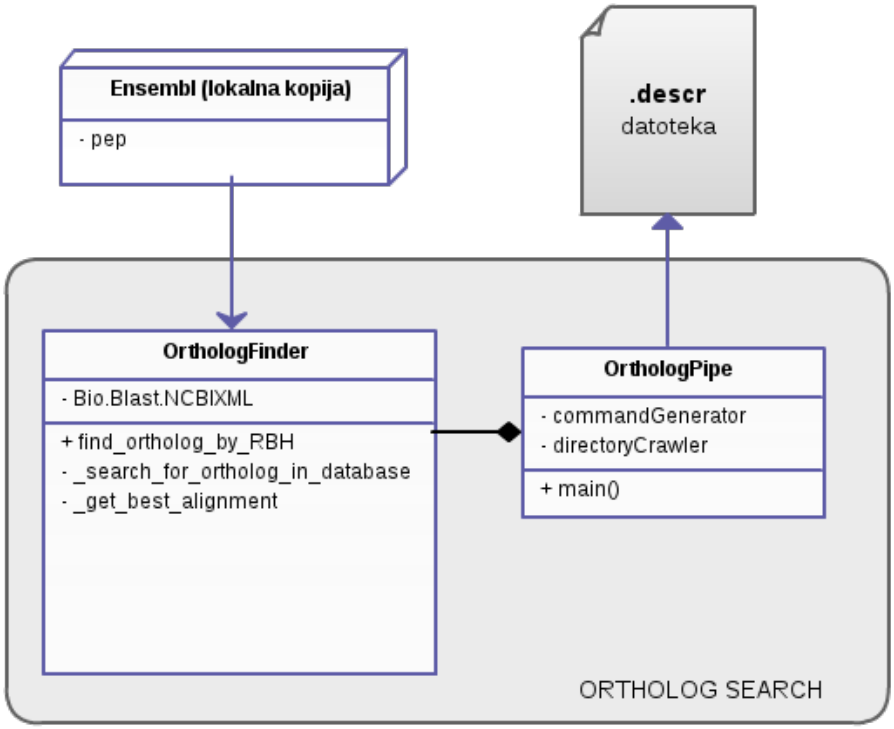


Sl. 7: Utilities modul. DirectoryCrawler služi generiranju puteva do svih mapa i datoteka pojedinog proteina. Nudi funkciju za generiranje čitavog stabla direktorija za pojedini protein. CommandGenerator služi za generiranje naredbi aplikacija naredbenog retka koje se koriste u aplikaciji.

5.2.3. ortholog_search modul

Ortholog_search modul nudi implementacije funkcija za pronalaženje ortolognih proteina na način kao što je to opisano u poglavlju 4.2. U svrhu raščlanjivanja izlaza iz BLASTp programa korištena je dostupna implementacija iz BioPython programskog sučenja dostupna u paketu Bio.Blast.NCBIXML. Kao što je prikazano u dijagramu klasa na slici 8, modul se sastoji od dvije klase. Ortholog_finder nudi funkcionalnost izvršavanja BLASTp upita nad proteinskom bazom, raščlanjivanja izlaza te identificiranja ortologa. Ortholog_pipe služi za pokretanje postupka pronalaska ortologa za sve proteine iz ispitnog skupa. Svaki pronađeni ortolog zapisuje u datoteku *ENSP00000xxxxx.descr* u formatu koji je detaljno opisan u

Dodatku A. Informacije o toku izvođenja se zapisuju u zapisnik definiran u konfiguracijskoj datoteci *logging.cfg*. Za svaki protein zapisuje se status pronalaska ortologa u skrivenu *.status* datoteku. Status može biti OK, PARTIAL ili FAILED.



Sl. 8: Ortholog search modul. Ortholog finder pronalazi najbolji RBH za pojedini protein. Ortholog pipe služi za pronalazak ortologa za sve proteine (i sve vrste) i zapisivanje informacija i mogućih grešaka u log datoteku. Rezultati RBH-a se za svaki protein zapisuju u .descr datoteku.

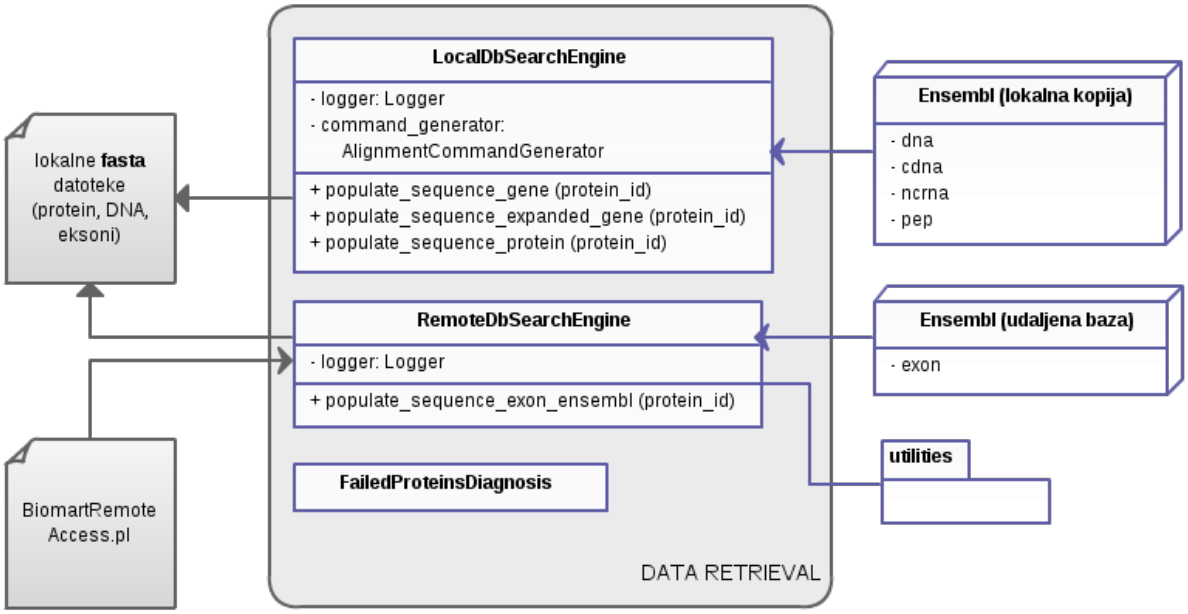
Izlaz iz modula za pronalazak ortologa je za svaki protein popis vrst aza koje je uspješno pronađen ortolog s odgovarajućim informacijama o lokaciji i tipu gena. Opis datoteke u kojoj su te informacije sadržane nalazi se u Dodatku A.

5.2.4. data_retrieval modul

Rad data_retrieval modula može se objasniti na primjeru jednog proteina u sljedećim koracima:

1. učita se .status datoteka i provjeri koji je status pronalaska ortologa. Ako je MUTUAL_BEST status FAILED, u .status datoteku se zapisuje da je dohvat podataka također neuspio. To je ujedno i kraj rada.

2. Za slučaj da je MUTUAL_BEST status OK ili PARTIAL, učitava se opisna datoteka (*ENSP00000XXXXX.descr).* Iz opisne datoteke se iščitavaju sve informacije o genima, transkriptima i proteinima.
3. Radi se dohvat proteinskih sljedova. Popunjava se direktorij *sequence/protein*.
4. Radi se dohvat sljedova regija gena. Popunjava se direktorij *sequence/gene.*
5. Radi se dohvat sljedova proširenih regija gena. Regije su proširene s obje strane i to za broj baza definiran u konfiguracijskoj datoteci *command_line.cfg*. Predefinirana vrijednost (i korištena u ovom projektu) je 150,000 baza. Popunjava se direktorij *sequence/expanded_gene.*
6. Radi se dohvat dostupnih eksona (*pep:known* i *pep:novel* proteini) sa udaljene Ensembl baze koristeći BioMart sučelje. Popunjava se direktorij *sequence/exon/ensembl*.
7. Za proteine za koje nisu dostupni obilježeni eksoni, pokreće se GeneWise na preuzetoj regiji gena. Tako pronađeni eksoni spremaju se u direktorij *sequence/exon/genewise.*
8. Za svaki od dohvata osvježava se status u *.status* datoteci. Klasni dijagram data_retrieval modula može se vidjeti na slici 9. Kao i kod ostalih modula, slijed izvršavanja se bilježi u zapisniku koji je definiran u *logging.cfg* datoteci.



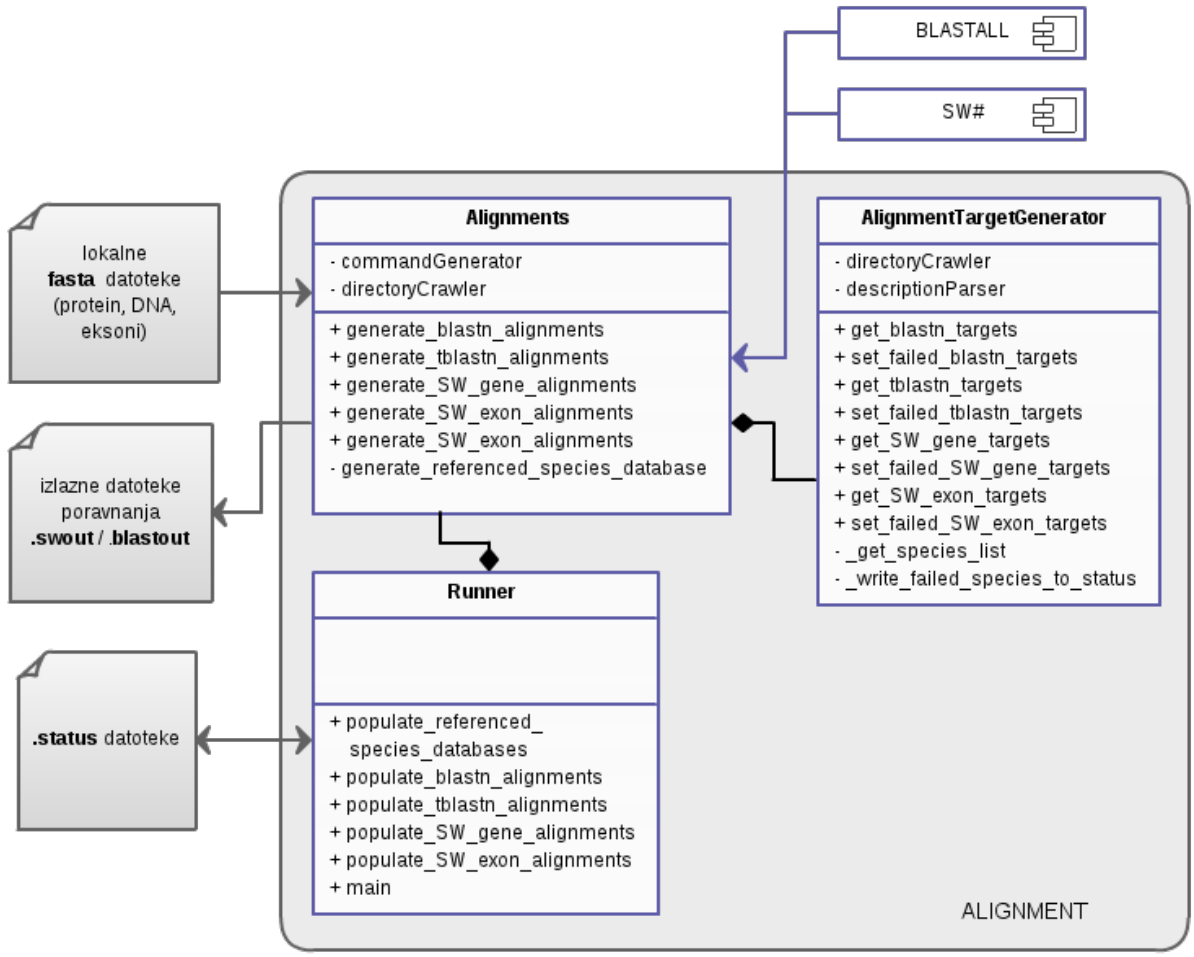
Sl. 9: Data retrieval modul. Služi pristupu lokalnoj kopiji Ensembl baze (LocalDbSearchEngine) te pristupu udaljenoj bazi preko Biomart sučelja. Rezultati se zapisuju u za to predodređen direktorij lokalno na računalu.

5.2.5. alignments modul

Za slučaj da su statusi prijašnjih koraka OK ili PARTIAL, pomoću funkcija alignment modula mogu se generirati sljedeća poravnanja:

- poravnanje eksona referentne vrste na proširenu regiju gena ciljne vrste koristeći BLASTn. Popunjava se direktorij *alignment/blastn.*
- Poravnanje proteina ciljne vrste na eksone referentne vrste koristeći tBLASTn. Popunjava se direktorij *alignment/tblastn.*
- Poravnanje eksona referentne vrste na proširenu regiju gena ciljne vrste koristeći SW# alat. Popunjava se direktorij *alignment/sw/gene.*
- Poravnanje eksona referentne vrste na spojene eksone ciljne vrste. Popunjava se direktorij *alignment/sw/exon*.

Za svako od poravnanja osvježavaju se za to prikladni statusi u *.status* datoteci. Organizacija modula može se vidjeti u klasnom dijagramu na slici 10.



Sl. 10: Alignment modul. Služi za generiranje poravnanja (blastn, tblastn, SW). Vrste se učitavaju iz .descr datoteka. Prije generiranja se provjerava status pojedinog poravnanja u .status datoteci i ovisno o tome pozivaju se prikladne funkcije.

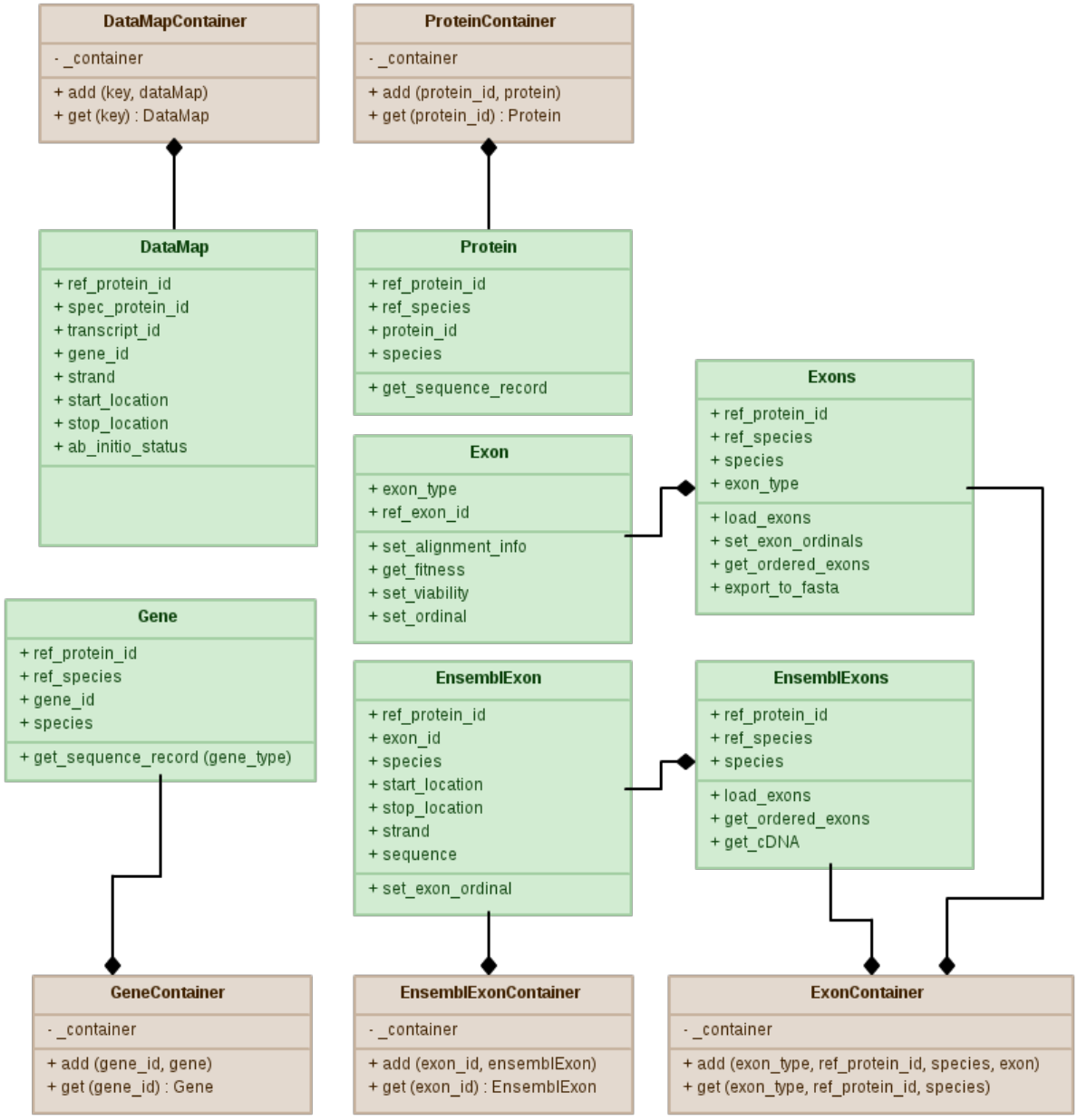
5.3. Cjevovod za analizu

Glavni proizvodi cjevovoda za analizu su protein sastavljen iz SW poravnanja i statistika u kojoj se može vidjeti koliki je postotak izvornog eksona (onog iz referentne vrste) pronađen korištenjem pojedinog poravnanja. Poravnanja koja se analiziraju nabrojana su u poglavlju 5.2.5.

Cjevovod za analizu nalazi se u *data_analysis* direktoriju aplikacije i može se reći da se sastoji od dva dijela: od objektnog modela bioloških podataka i od dijela koji vrši operacije

nad tim objektnim modelom. U sljedećem se potpoglavlju prvo biti objašnjen objektni model, a potom koje se to operacije nad njim obavljaju.

5.3.1. Objektni model



Sl. 11: Objektni model bioloških podataka. Zeleno su obojane klase koje modeliraju biološke podatke, dok su smeđe obojane klase koje po identifikatoru-ključu mapiraju odgovarajuće objekte.

Na slici 11 može se vidjeti objektni model čija je svrha sistematično učitavanje i dohvaćanje bioloških podataka. Zelenom su bojom prikazani objekti koji modeliraju biološke podatke, poput gena i proteina, a smeđom bojom spremnici u koji se po prikladnom identifikatoru-ključu pohranjuju odgovarajući objekti. Svaki od spremnika je implementiran po *Singleton* oblikovnom obrascu, tako da bi se, jednom nakon što je učitavanje objekata obavljeno, spremnik mogao koristiti u čitavoj aplikaciji.

Izvor informacije za svaki od proteina iz ispitnog skupa je *.descr* datoteka. Svaki zapis iz te opisne datoteke učitat će se u jedan **DataMap** objekt. Dakle, u objektu tipa DataMap bit će sadržane informacije o identifikatoru referentnog proteina, o vrsti, identifikatorima proteina, gena i transkripta, o lokaciji gena te *ab_initio* statusu koji govori je li riječ o predviđenom ili poznatom proteinu. DataMap objekti se pohranjuju u **DataMapContainter** po ključu (identifikator referentnog proteina, ciljna vrsta). Korištenjem DataMap objekata moguće je učiniti sve ostale objekte nezavisnima – protein ne mora znati za gen koji ga kodira, već može dohvatiti odgovarajući objekt tipa DataMap, iz kog se može iščitati identifikator gena, čime ga se može jednostavno dohvatiti iz njegovog (Gene) spremnika. Ovakva organizacija doprinosti nezavisnosti među objektima i omogućava lakše unošenje promjena u svaki od njih.

Klasa **Protein** služi kao model proteina. Svaki objekt tipa Protein sadrži informaciju o identifikatoru referentnog proteina, svom identifikatoru i o vrsti iz koje potiče. Nudi mogućnost *lijenog učitavanja* proteinskog slijeda, što znači da se slijed ne učitava pri generiranju objekta, već pri prvom pozivu metode *get_sequence_record*. Objekt tipa Protein se posprema u **ProteinContainer** spremnik po vlastitom indentifikatoru proteina. Za klasu **Gene** može se reći isto, osim što umjesto identifikatora proteina se mapira po identifikatoru gena u **GeneContainer** spremnik. Osim toga, gen ima tip – ili je običan ili je proširen – misli se naravno na regiju gena koja se učitava pozivom *get_sequence_record* metode.

Klasa **EnsemblExon** služi kao model eksona preuzetog s Ensembl-a, te sadrži identifikator referentnog proteina te sve informacije o samom eskonu – njegova lokacija, identifikator, slijed i vrsta. Objekti tip EnsemblExon pohranjuju se direktno u **EnsemblExonContainer**, gdje se mapiraju po indentifikatoru eksona. Osim toga, mogu se učitati u objekt tipa **EnsemblExons**, koji sadrži sve eksone koji se mogu dohvatiti s Ensembl baze za pojedini protein. Ova klasa nudi metodu *lijenog učitavanja* eksona, te slijednog dohvata eksona. Pri

preuzimanju eksona s Ensembl-a oni ne dobiju u pravom redoslijedu u kom se nalaze na genomu. Pri prvom pozivu metode get_ordered_exons se eksonima postavljaju redni brojevi, koji se naknadno koriste za slijedni dohvat eksona.

Klasa **Exon** služi kao model eksona koji je rezultat poravnanja izvornog eksona na regiju gena ciljne vrste. Korišteno poravnanje može biti *blastn, tblastn, sw_exon* i *sw_gene*. Svaki od eksona nastalih poravnanjem ima informaciju o identifikatoru referentnog eksona, te nudi metodu za postavljanje informacija o poravnanju (kao što je broj identičnih nukleotida, duljina poravnanja, slijed, izvorni slijed (referentnog eksona), broj praznina u poravnanju, lokacije početka i kraja poravnanja u oba slijeda). Svi eksoni jednog proteina nastali poravnanjem učitavaju se u objekt tipa **Exons**. Eksonima se dodjeljuje par rednih brojeva. Prvi broj odgovara rednom broju referentnog eksona. Drugi broj odgovara rednom broju dobivenog nepreklapajućeg poravnanja za taj ekson. I u ovom slučaju redni brojevi služe lakšem slijednom dohvatu eksona. Redni se brojevi postavljaju pri pozivu funkcije *get_ordered_exons* iz klase Exons.

Objekti tipa Exons i EnsemblExons se pohranjuju u **ExonContainer** po ključu vrste, tipa eksona i identifikatora referentnog proteina. Tip eksona može biti *ensembl, blastn, tblastn, sw_gene* i *sw_exon.*

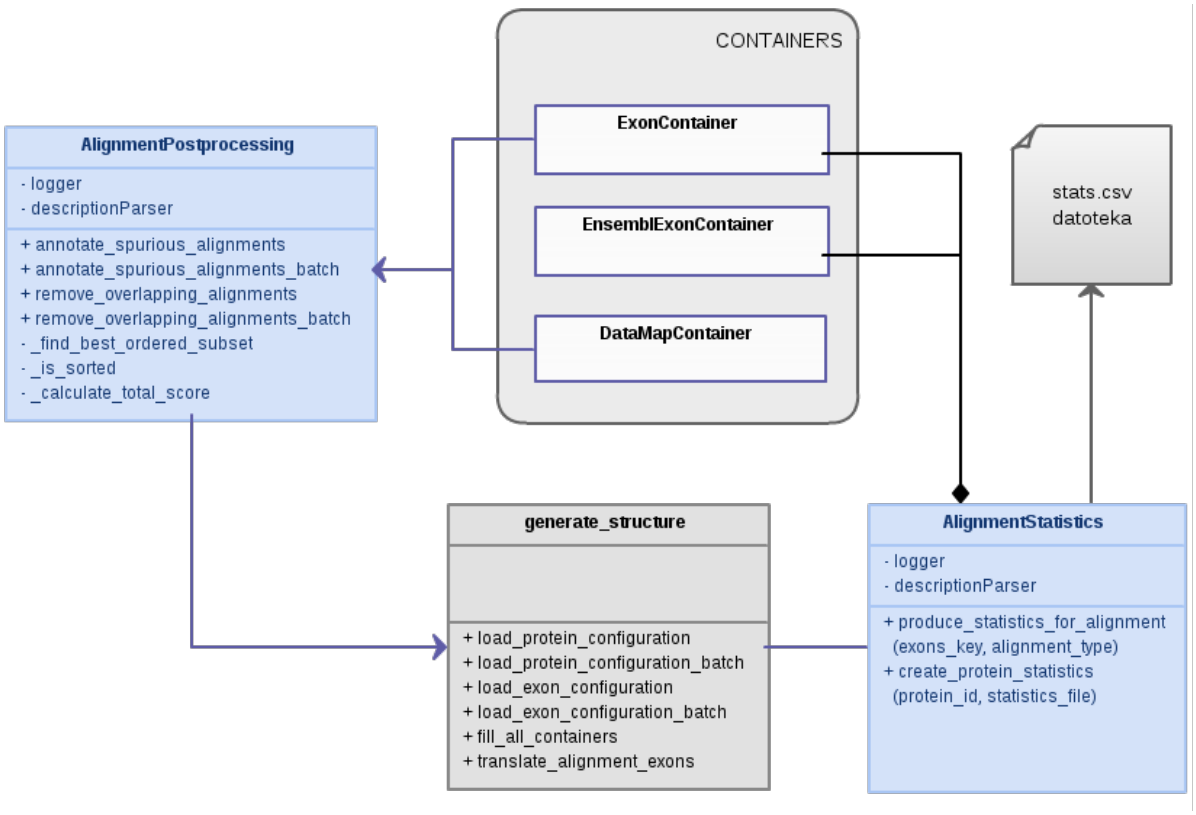
5.3.2. Naknadna analiza poravnanja, statistike i slaganje proteina

Budući da se i statistika i konačni proteinski proizvod generiranju iz poravnanja, potrebno je poravnanja, tako reći, dovesti u red. Postoji nekoliko slučajeva u kojima određena poravnanja treba zanemariti. Rješenja ovih problema implementirana su u skripti **AlignmentPostprocessing**. Rezultati postanalize bilježe se u za to odabrani zapisnik.

Prvi problem na koji se često nailazi je da je *n*-ti ekson pronađen prije *m*-tog eksona u genu, a . Očigledno jedan od ova dva eksona treba odbaciti, pa se postavlja pitanje koji od ta dva proglasiti nevaljalim. Moguće je gledati isključivo kvalitetu poravnanja ta dva eksona, ali taj se pristup već nakon kraćeg razmišljanja pokaže lošim. Razlog tomu je što se oko tih eksona nalaze ostali eksoni u određenom redoslijedu, i odbacivanjem jednog možemo implicitno cijeli podskup eksona proglasiti nevaljalim. Dakle, potrebno je na neki način pronaći najveći podskup ispravno poredanih eksona. Pri tome treba uzeti u obzir veličinu skupa, ali i kvalitetu poravnanja.

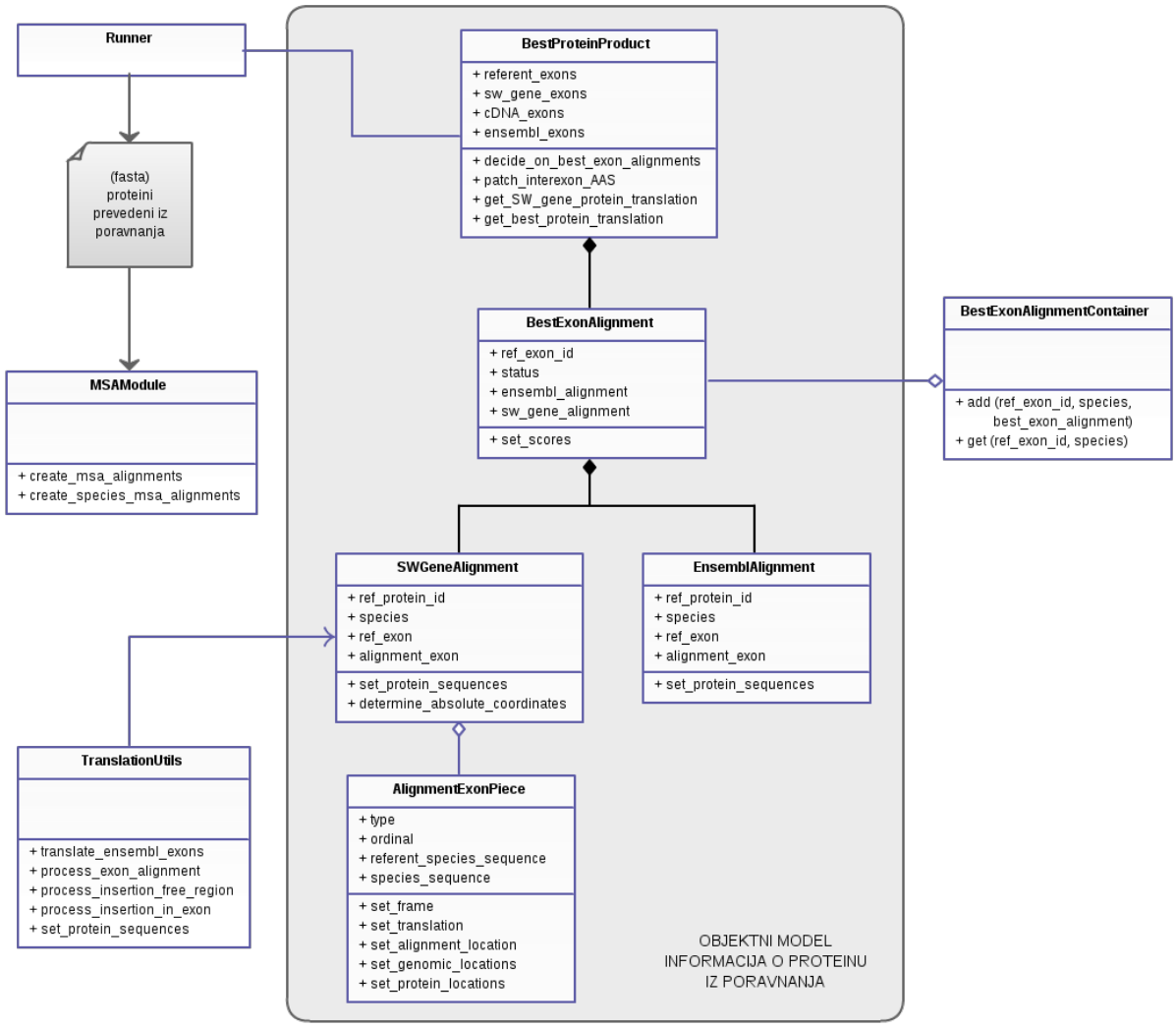
Na drugi se problem nailazi pri ocjeni BLASTn poravnanja kod generiranja statistike. Ovaj problem je prvi put uočen na način da se u statistikama se pokazalo da BLAST nekad ima postotak pokrivenosti eksona veći od 100%. Razlog tomu je što BLAST kao izlaz vraća sva poravnanja koja koja se ocjenjenom kvalitetom nalaze iznad nekog praga. Može se dogoditi, i često se događa, da se ova poravnanja preklapaju. Zbog toga je potrebno pronaći najveći skup poravnanja koja se ne preklapaju, a pokrivaju najveći dio referentnog eksona. Također, BLAST vraća poravnanja na obe niti DNK. Budući da je točna nit i lokacija gena unaprijed poznata, moguće je bez većih problema odbaciti poravnanja na suprotnoj niti. Valja napomenuti da su ponekad ova poravnanja bila zabrinjavajuće dobra.

Nakon što su uklonjena preklapajuća poravnanja, kao i ona u pogrešnom redoslijedu, moguće je pristupiti generiranju statistike i sklapanju proteina. Za svaki se protein generira po jedna *stats.csv* datoteka koja se nalazi u vršnom direktoriju proteina. U toj se datoteci može vidjeti koliki je postotak pokrivenosti pojedinih eksona postignut s različitim algoritmima poravnanja. Metode za generiranje statistika nalaze se u skripti **AlignmentStatistics**.



Sl. 12: Modul za naknadnu analizu poravnanja i generiranje statistike. generate_structure nudi metode za popunjavanje svih spremnika s biološkim podacima i podacima poravnanja

Budući da je cilj projekta ponuditi bolju informaciju o proteinskom slijedu, zadnji i ujedno najvažniji korak aplikacije je slaganje proteina iz poravnanja. Za ovo se koristi poravnanje eksona referentne vrste na proširenu regiju gena ciljne vrste uporabom Smith-Waterman algoritma. Za prevođenje poravnanja u protein stvoren je prikladan objektni model koji iz kog je lako dohvatiti najbolje moguće poravnanje za pojedini ekson te njegov prijevod u protein, kao i prijevod čitave cDNK u protein.



Sl. 13: Modul za prijevod poravnanja u protein. Objektni model sadrži sve informacije potrebne za rekonstrukciju čitavog proteinskog slijeda iz poravnanja, kao i dijelova proteina koji odgovaraju pojedinim poravnanjima

Vršna klasa modula za prevođenje poravnanja u protein je **BestProteinProduct**. Za svaki od eksona iz referentnog proteina postoji po jedna instanca klase **BestExonAlignment**. Za svaki

se ekson razmatraju dva poravnanja. Prvo je poravnanje ljudskog eksona na regiju gena vrste, a drugo poravnanja ljudskog eksona na cDNK vrste nastalu spajanjem eksona prijavljenim na Ensembl-u. Za svaku se ekson razmatra bolje poravnanje. Zbog tog klasa BestExonAlignment ima statusnu varijablu mogućih vrijednosti *ensembl, sw_gene* i *both* koja govori koji od dva pristupa nudi bolje poravnanje za pojedini ekson referentne vrste. Kvaliteta se ocjenjuje rezultatom SW poravnanja. U svakoj se instanci BestExonAlignment-a nalaze instance dviju klasa koje predstavljaju SW poravnanje eksona na gen i SW poravnanje eksona na cDNK vrste – **SWGeneAlignment** i **EnsemblAlignment**. EnsemblAlignment sadrži identifikatore eksona vrste koji su sadržani u poravnanju te u koji se komad proteina ta regija prevodi.

SWGeneAlignment se sastoji od manjih dijelova poravnanja. Ukratko, svaka će praznina u poravnanju, bilo u referentnoj, bilo u ciljnoj vrsti, biti zaseban komad poravnanja. Taj je komad poravnanja oblikovan klasom **AlignmentExonPiece**. Ovisno o tome nalazi li se u komadu praznina, te ovisno u kojem se slijedu praznina nalazi, tip komada poravnanja može biti *coding, insertion, frameshift* ili *deletion*. Coding znači da ni u jednom slijedu nema praznina. U insertion tipu se praznine nalaze u eksonu referentne vrste (umetanje u proteinski slijed referentne vrste – insertion). Frameshift tip je insertion odgovarajuće duljine (1, 2, 4 ili 5). Ovakva kratka umetanja uglavnom su rezultat pogrešaka u sekvenciranju. Deletion tip podrazumijeva praznine u eksonu ciljne vrste te znači da taj dio proteina nije bilo moguće rekonstruirati u ciljnoj vrsti. Ovaj tip će se u proteinu prevesti u X znakove.

BestProteinProduct klasa nudi metodu za rekonstrukciju proteina iz SW poravnanja na gen (get_SW_gene_protein_translation). Proteinski sljedovi dobiveni na ovaj način se spremaju u fasta datoteke u direktorij proteina `sequence/assembled_protein`. Osim samog slijeda, za svaki od proteina generira se poravnanje triju proteina: protein predložak iz referentne vrste, protein nastao kao rezultat prevođenjem poravnanja i protein ciljne vrste kakav je prijavljen na Ensembl-u. Ova poravnanja mogu se naći u direktoriju `alignment/mafft`.

6. Rezultati

6.1. Ispitni skup i predviđanje ortologije

Kao što je opisano u poglavlju 4.3, prvi korak je definiranje referentnih vrsta za svaku vrstu, nakon čega slijedi definiranje ispitnog skupa i potraga za ortolozima. Trenutno je kao referentna vrsta za sve ostale vrste definiran čovjek zbog neusporedivo bolje kvalitete dostupnih anotacija. Dakle, ispitni skup će biti sačinjen isključivo od ljudskih proteina. Zbog uske veze između formiranja ispitnog skupa i predviđanja ortologije, oboje će biti opisano u ovom poglavlju.

U čovjekovom je genomu trenutno identificirano i obilježeno približno 30,000 gena. Na Ensembl-u jedostupno 100,354 poznatih proteina (za koje su obilježeni geni) te 47,254 proteina koji su predviđeni pomoću Genscan alata, ali za koje ne postoji dostupna anotacija.

6.1.1. Uklanjanje problematičnih proteina iz ispitnog skupa

Korištenjem ovog pristupa za predviđanje ortologa valja biti oprezan. Dvije su stvari koje mogu poći po zlu.

Prvi problem na koji se može naići su paralozi. Kao što je već opisano, paralozi su geni nastali duplikacijom u genomu. Sama riječ duplikacija daje naznaku što bi moglo poći po zlu. U povratnom koraku RBH algoritma mogli bismo pogrešno identificirati paralog kao najbolji rezultat, opet zbog prirode korištenog alata. Iz tog su razloga svi proteini s poznatim paralozima izbačeni iz ispitnog skupa, što je inicijalni broj od 97,080 proteina smanjilo na 29,663.

Drugo, valja uzeti u obzir da jedan gen može imati više proteinskih proizvoda. Zbog prirode RBH algoritma, moglo bi se dogoditi da povratni korak (2a, 2b, slika 15) da pogrešnu informaciju o ortologiji – ili bi bio prijavljen lažan ortolog (u slučaju da je protein ciljne vrste zaista ortolog drugog proteinskog produkta istog gena) ili bi ispravna pretpostavka o ortogologiji izostala. Uzrok tome je, među ostalim, heuristička priroda BLASTp alata i male razlike među rezultatima poravnanja koji ne moraju nužno ilustrirati razlike u kvaliteti. Zbog

toga su iz ispitnog skupa uklonjeni proteini ako geni koji ih kodiraju mogu imati više proteinskih proizvoda, što je ispitni skup smanjilo na 1,789 proteina.

Zadnjem se problemu može doskočiti na način da se za svaki gen uzme proteinski proizvod koji pokriva najveći broj eksona, ili skup proteina koji zajedno pokrivaju čitav skup eksona. U ovom se radu, ipak, nije radilo s genima s više proteinskih proizvoda.

6.1.2. Proteini za koje se ne mogu pronaći ortolozi

Od 1,789 proteina koji sačinjavaju ispitni skup, za njih 329 nije bilo moguće pronaći ortologe niti u jednoj vrsti. Ručnim pregledom funkcija ovih proteina ustvrđeni su mogući razlozi RBH promašaja.

- **129** proteina pripada skupu proteina imunuloškog sustava. Oni nisu dio genoma te nisu podložni standardnim selekcijskim i mutacijskim mehanizmima primjenjivima na ostatak proteina.
- **57** su neokarakterizirani proteini kojima funcija još nije utvrđena
- **1** je pseudogen
- za **131** ne postoji nikakav opis na Ensembl-u, pretpostavka je da im je funkcija nepoznata
- za **5** ih se pretpostavlja da su nastali nakon separacije čovjeka i gorile
- za **6** uzrok nije poznat

Ovime je ispitni skup smanjen na 1,460 proteina iz čovjeka za koje je pronađen najmanje jedan ortolog.

6.1.3. Statistike broja pronađenih ortologa

Za svaki od proteina se generira datoteka u kojoj se nalazi popis vrsta za koje je uspješno pronađen ortolog. Ime datoteke je *ID_proteina.descr*. Za svaku od vrsta zapisuje se ID proteina i gena (informacija koja nije dostupna za proteine klase *pep:genscan*, te se za njih ne zapisuje), te lokacija gena.

Za proteine klase *pep:known* i *pep:novel* taj zapis izgleda ovako:

```
Ailuropoda_melanoleuca          ENSAMEP00000013619          ENSAMEG00000012932
ENSAMET00000014187      scaffold:ailMel1:GL192391.1:2632274:2648660:1
```

Za proteine klase *pep:genscan* taj zapis izgleda ovako:

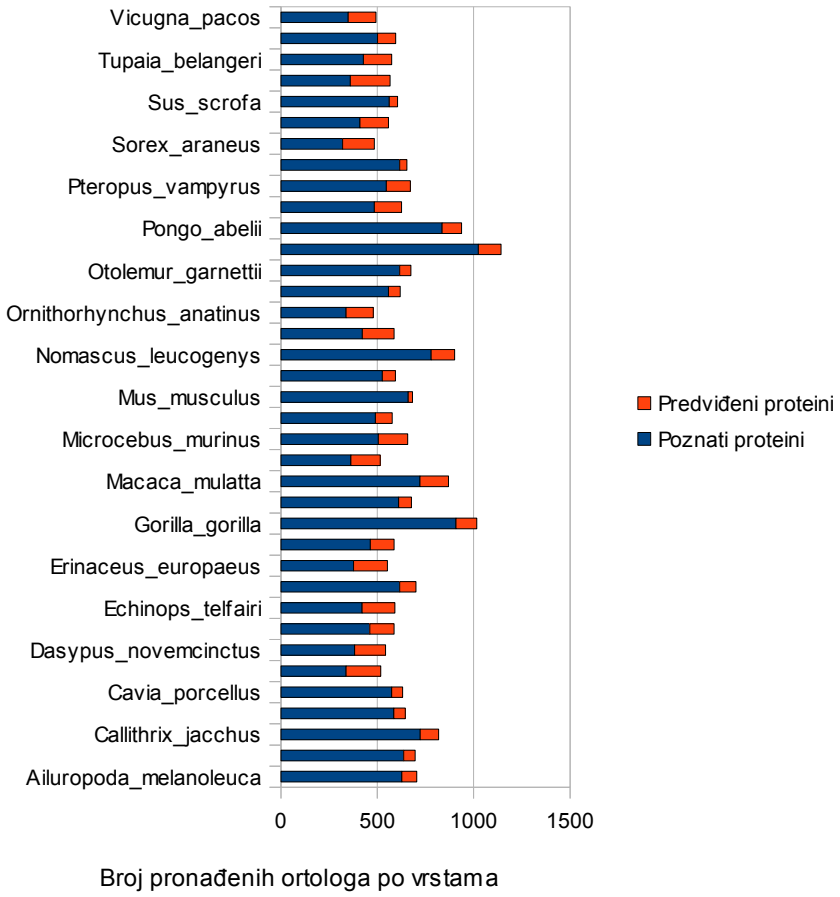
```
Erinaceus_europaeus      GENSCAN_GS00000149474
      scaffold:HEDGEHOG:scaffold_347015:52585:53518:-1
```

Detaljan opis svih informacija može se pronaći u Dodatku A.

Sl. 14: Za svaki protein pronađen je određen broj ortologa. Na ovom se histogramu može vidjeti raspodjela broja proteina u odnosu na broj pronađenih ortologa. Prva dva stupca (0 i 1 ortolog) su ona 329 proteina za koje nije pronađen nijedan ortolog iz razloga što stupac 1 odgovara proteinima u kojima je ortolog pronađen samo u čovjeku.

Na slici 14 može se vidjeti raspodjela broja proteina u odnosu na broj pronađenih ortologa. Prvi stupac, s nula ortologa, indicira da je došlo do greške[3] jer ortolog nije pronađen ni za referentnu vrstu – čovjeka. Drugi stupac, s jednim ortologom odgovara proteinima za koje je pronađen samo jedan ortolog, onaj iz čovjeka. Ova prva dva stupca odgovaraju proteinima opisanima u poglavlju 6.1.2.

[3]Jedna od mogućih grešaka je postojanje proteina različitih haplotipova koji su u slijedu gotovo isti, ali izvorni protein ima nešto manji BLAST rezultat, čime ne biva prijavljen kao ortolog



Sl. 15: Prikaz pronađenog broja ortologa za svaku od vrsta. Plavom je bojom označen broj ortologa pronađen među proteinima klase pep:known i pep:novel, a crvenom među proteinima klase pep:genscan

Na slici 15 može se vidjeti histogram koji predstavlja broj pronađenih ortologa po vrstama. Plava boja označava proteine klase *pep:known* i *pep:known*, dok crvena boja predstavlja pretpostavljene proteine klase *pep:genscan*. Može se vidjeti da je najveći broj ortologa pronađen za čimpanzu (*Pan Troglodytes*), gorilu *(Gorilla Gorilla),* orangutana *(Pongo Abelii)*, gibona (N*omascus Leucogenys*), makaka *(Macaca Mulatta)*, što je i za očekivati budući da su najsrodniji čovjeku. Ipak, značajan je broj ortologa pronađen i za udaljenije vrste, poput ljenjivca i čudnovatog kljunaša, što je u komparativnoj analizi proteina korisnije od iznimno srodnih vrsta.

6.2. Rekonstrukcija proteina iz poravnanja

Proteini su rekonstruirani za sve proteine iz ispitnog skupa i sve vrste za koje je bilo moguće pretpostaviti ortologe. Također, za sve su proteine generirana poravnanja među tri proteina: protein referentne vrste, proteina ciljne vrste nastalog prevođenjem poravnanja i proteina ciljne vrste s Ensembl-a. Na osnovi ovih poravnanja generirane su statistike koje bi trebale pokazati u koliko slučajeva i za koliko je bolja pojedina metoda pronalaska proteina. Primjer ovakvog poravnanja može se vidjeti na slici 16.

Sl. 16: Gore: Isječak poravnanja proteina iz čovjeka (gornji protein), proteina iz cjevovoda za analizu (srednji protein) i proteina iz Ensembl-a (donji protein) za ljenjivca, dolje: poravnanje tri opisana proteina za čudnovatog kljunaša.

Na gornjem dijelu slike 16 može se vidjeti primjer u kom je protein iz cjevovoda pokrio puno veći dio proteina iz srodne vrste (u ovom slučaju čovjeka). Na donjem dijelu slike može se vidjeti primjer u kom je samo dio proteina rekonstruiran iz poravnanja. Na slici 17 može se vidjeti primjer poravnanja više proteina s Ensembl-a (gore) i rekonstruiranih iz poravnanja (dolje).

Sl. 17: Gore: poravnanje proteina rekonstruiranih iz poravnanja. Dolje: poravnanje proteina preuzetih s Ensembl-a.Proteini iz poravnanja se većinom podudaraju u regijama koje su već prijavljene na Ensembl-u, dok se u vrstama gdje su dijelovi proteina nedostajali može često vidjeti rekonstrukcija većeg dijela proteina.

6.2.1. Statistika na razini proteina

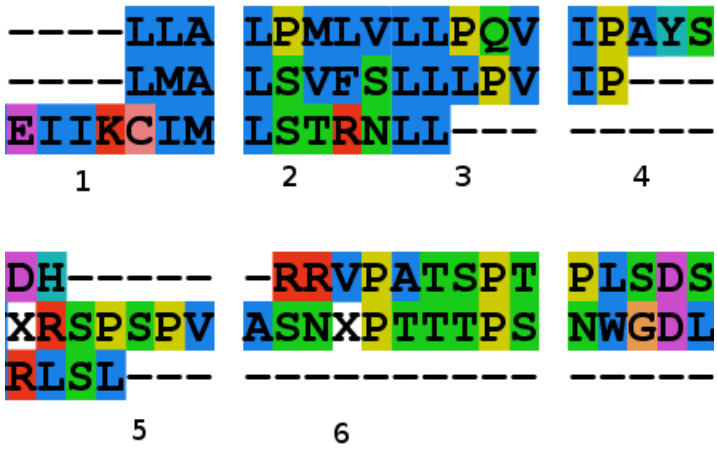
Proteini s Ensembl-a iz ciljne vrste i proteini koji su rezultati opisanog cjevovoda mogu se razlikovati od referentnog proteina na više načina:

- mogu ne pokrivati dio proteina (praznine u poravnanju)
- mogu imati "višak", kao što je vidljivo na donjem dijelu slike 16 za slučaj proteina s Ensembl-a
- mogu odgovarati duljinom, ali imati značajno[4] različit slijed.

Potrebno je naći način za evuluaciju uspješnosti rekonstrukcije proteina. Kako je ovo izrazito složeno zbog svih neodgovorenih bioloških pitanja[5], pri generiranju prve statistike uzeta je izrazito jednostavna mjera. U poravnanju proteina pobrojeni su slučajevi u kojima proteini iz ciljne vrste imaju brisanje ili umetanje u odnosu na referentni protein, te u kojima podudaranje postoji. Podudaranje u ovom slučaju ne govori ništa o stvarnoj sličnosti aminokiselina, već samo da na pojedinom mjestu u proteinu ne postoji ni umetanje ni brisanje. Prikaz ovih slučajeva može se naći na slici 18.

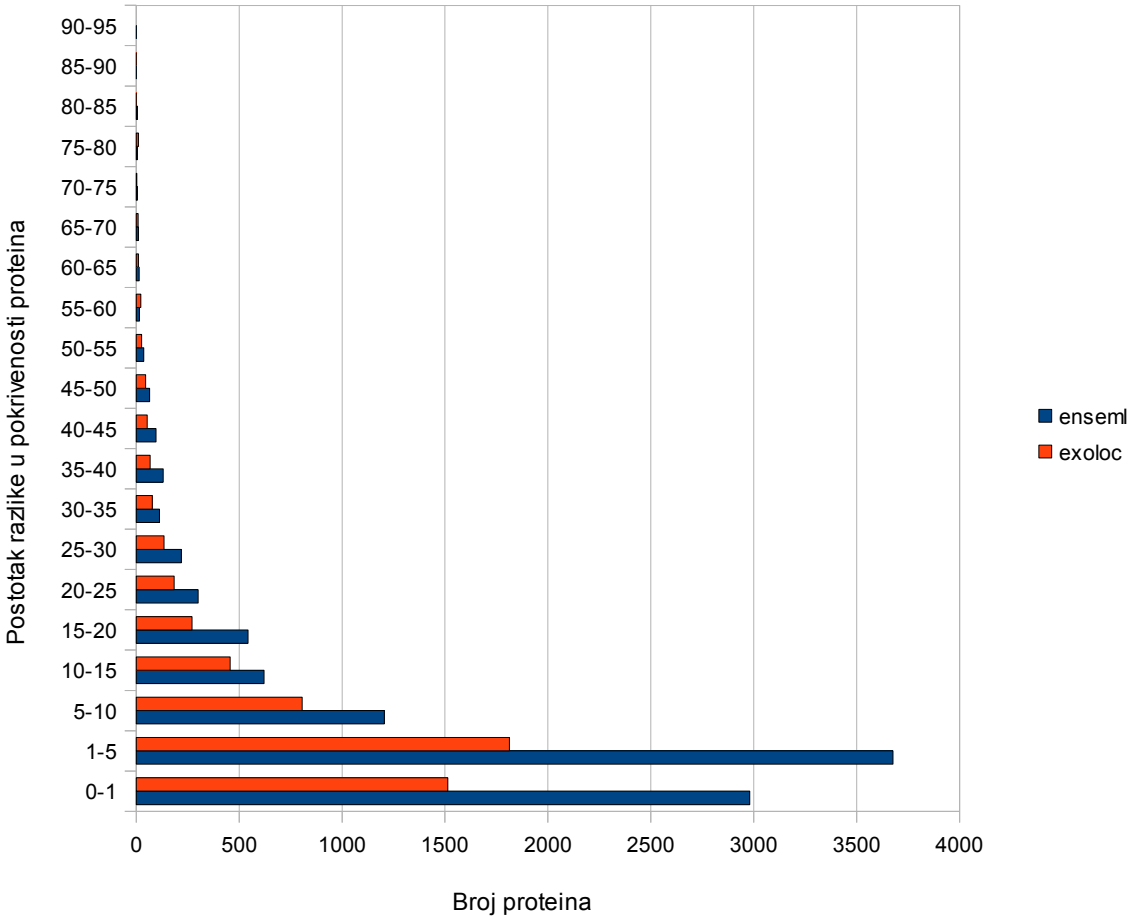
[4]Odrediti kada je slijed značajno različit nije nipošto jednostavno. Značajna razlika bi implicirala da evolucijom tog dijela proteina dolazi do izmjene u njegovoj funkciji u stanici. Da bi se znala funkcija proteina, potrebno je puno više informacija od samog proteinskog slijeda, iz zbog toga ovaj izraz valja uzeti s rezervom.

[5]Jesu li proteini zaista ortolozi? Je li praznina zapravo biološki opravdana i taj dio proteina ne postoji u ciljnoj vrsti? Je li "višak" biološki opravdan i zaista se nalazi u proteinu ciljne vrste?



Sl. 18: Objašnjenje statistike, referentni protein je gornji. Slučaj 1: višak u donjem proteinu. Slučaj 2: podudaranje sva tri proteina. Slučaj 3: praznina u donjem proteinu, podudaranje referentnog i srednjeg proteina. Slučaj 4: praznina u oba proteina ciljne vrste. Slučaj 5: višak u srednjem proteinu, Slučaj 6: x se interpretira kao praznina, praznina u oba proteina ciljne vrste

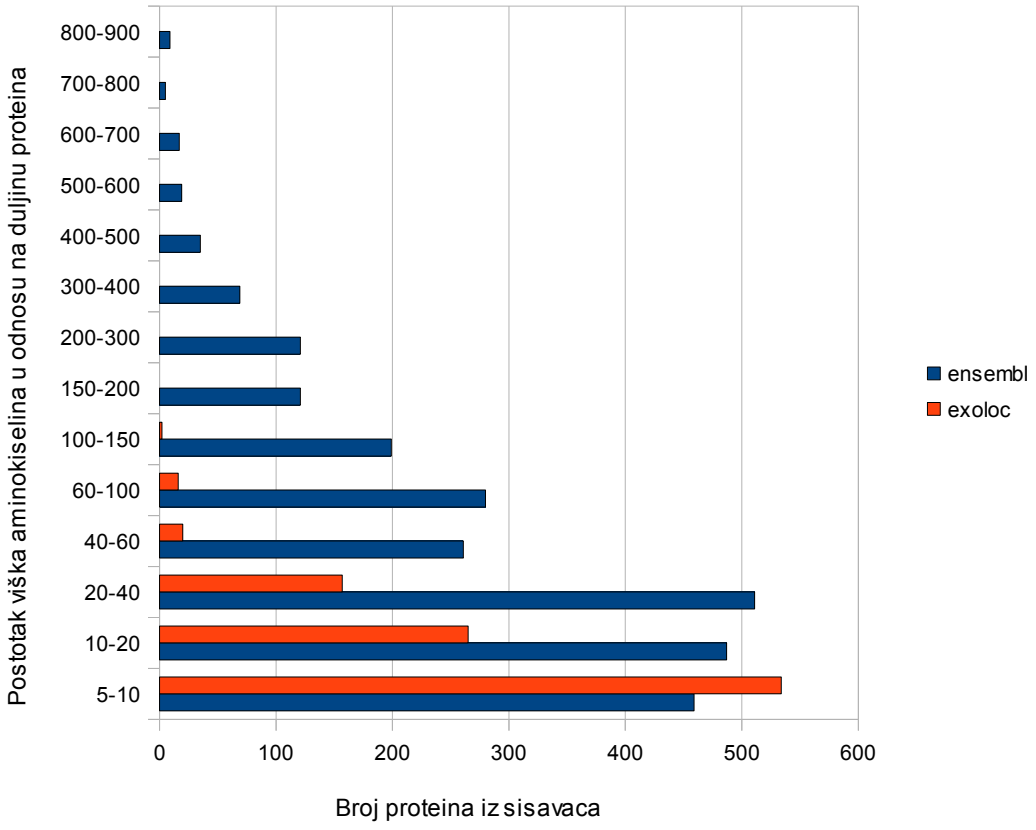
Razlog iz kog bi se mogao pronaći samo manji dio proteina je što se ti proteini, iako ortolozi, u različitim vrstama doista razlikuju. Drugi razlog je taj što ti proteini nisu zaista ortolozi (istinitost ove tvrdnje nažalost nije moguće ustvrditi). Ono što statistika može reći postoje il dijelovi proteina koji su evolucijski dovoljno konzervirani da ih se može pronaći u srodnim vrstama uporabom poravnanja sljedova. S druge strane, može nam reći koliki dio proteina se ne može pronaći na ovaj način, a na proteinu s Ensembl-a je uspješno pronađen. Mogući razlog je manja evolucijska konzerviranost na razini cijelog niza, ali očuvanje bioloških signala početka i kraja gena i / ili eksona, što se može uspješno pronaći uporabom programa poput Genewise-a.



Sl. 19: Broj proteina sa Ensembl-a (plavo) i iz Exolocator-a (crveno) gdje razlika u pokrivenosti proteina referentne vrste odgovara postocima na y osi.

Na slici 19 može se vidjeti raspodjela broja proteina po razlici u postotku pokrivenosti referentnog proteina. Postoci po kojima je broj proteina grupiran su zapravo razlike između postotka pokrivenosti referentnog proteina Ensembl-ovim proteinom i proteinom iz cjevovoda. Dakle, ako je razlika u postocima od 10-15% u korist Ensembl-a, onda će se u toj kućici broj proteina povećati za jedan. Plavom je označen broj proteina za koje odgovarajuća razlika u postocima znači bolju pokrivenost od Ensembl proteina, a crvenom bojom bolju pokrivenost od strane proteina iz cjevovoda. Potrebno je napomenuti da se broj proteina odnosi na proteine ortologe iz sisavaca za 1460 ljudskih proteina ispitnog skupa.

Na slici 20 može se vidjeti raspodjela broja proteina po postotku viška slijeda u odnosu na duljinu referentnog proteina. Najčešći uzrok viška u proteinima iz cjevovoda je taj što nekada jedan ekson iz referentne vrste odgovara dva ili više eksona iz ciljne vrste. Kako Smith-Waterman algoritam vraća samo najbolje poravnanje, on će u većini slučajeva vratiti poravnanje koje se proteže i kroz intronsku regiju u ciljnoj vrsti. Primjer takvog slučaja je poravnanje jedinog eksona (protein je ENSP00000373139) iz čovjeka na regiju gena orangutana. U orangutanu taj protein kodiraju dva eksona udaljena približno 1,300 nukleotidnih baza. Poravnanje uspješno pronađe oba eksona s prazninama u intronskoj regiji koja, zbog načina prevođenja poravnanja u protein, rezultira viškom u proteinu.



Sl. 20: Histogram postotka umetanja u odnosu na referentni protein. Na okomitoj se osi nalaze postoci umetanja koji se računaju kao duljina umetnutog niza / duljina niza referentnog proteina

Ručnim pregledavanjem uočeno je da se većina umetanja u proteinima iz Ensembl-a nalazi na početku ili kraju proteina. Mogući razlog je taj što programi za predviđanje lokacija gena i proteinskih proizvoda imaju najviše problema s početnim i krajnjim eksonima.

6.2.2. Konačni proteinski proizvod

Budući da je cilj aplikacije ponuditi što potpuniju informaciju u proteinu, konačni proteinski proizvod ne sastoji se od prevođenja samog poravnanja, budući da ono ne mora biti izvor najbolje informacije. Za svaki je dio proteina poznat izvor najkvalitetnije informacije, bila ona s Ensembl-a ili rezultat poravnanja. Kako je proces obrade i obilježavanja poravnanja dugotrajan, svi podaci se nakon obrade učitavaju u bazu podataka - uključujući proteine, gene, eksone iz poravnanja te prevedene dijelove poravnanja. Dohvatom potrebnih podataka iz baze (uz nešto dodatne obrade) moguće je napraviti rekonstrukciju proteina.

Da bi rezultat ovog rada bio koristan široj publici, dobiveni su podaci javno dostupni preko mrežnog sučelja na adresi http://exolocator.bii.a-star.edu.sg/. Ovaj dio nije dio aplikacije, već se koristi podacima generiranima u aplikaciji za rekonstrukciju najboljeg proteinskog proizvoda.

6.3. Rasprava

Iz postignutih je rezultata vidljivo da je korištenjem optimalnog poravnanja eksona na gen moguće zaista pronaći veći broj eksona no što je to moguće korištenjem heurističkog poravnanja proteina na gen. Usporedba je vršena sa sofisticiranim i kompleksnim Ensembl cjevovodom za anotaciju genoma u kojem je ovo poravnanje samo jedan od mnogih koraka u pronalasku eksona / intron regija gena. Pokazano je da je čak jednostavnim pristupom, kao što je prevođenje poravnanja u protein, moguće postoji jako dobre rezultate, koji kvalitetom u velikom broju slučajeva odgovaraju proteinima preuzetima s Ensembl-a.

Ono što ovaj pristup ne osigurava je da su pronađeni eksoni zaista okruženi pravim intronskim regijama. Intronske regije sadrže signale potrebne na biološki mehanizam prepozna eksone kao protein-kodirajuće regije. Za ovu svrhu Ensembl koristi GeneWise. Jednaka bi kvaliteta anotacije bila osigurana kada bi se regije eksona, detektirane pomoću SW# alata, proširile određenim brojem nukletida koji bi zahvatio intronske regije, spojile u slijed koji bi se potom predao kao ulaz u program poput Genewise-a.

Drugi od mogućih problema ovog pristupa je preslikavanje eksona iz jedne vrste na eksone druge vrste. Nije rijetkost da tokom evolucije u jednoj vrsti dođe do podjele eksona, a u

drugoj ne. Pokuša li se poravnati jedan ekson na regiju gena koja sadrži taj slijed, ali u dva dijela, između kojih se može nalaziti velik broj intronskih nukleotida, vjerojatnost da će ovo poravnanje biti prijavljeno kao optimalno opada s duljinom intronske regije. Ovome bi se problemu moglo doskočiti korištenjem algoritma dinamičkog programiranja koji osim optimalnog, prijavljuje i određen broj suboptimalnih poravnanja.

Zaključak

Pri anotaciji gena, to jest njemu odgovarajućih ekson – intron regija, često se pribjegava korištenju heurističkih metoda poravnanja. Količina informacija u anotaciji genoma nalaže upotrebu brzih rješenja, što najčešće isključuje mogućnost korištenja optimalnog poravnanja. Zbog dostupnog alata SW# za optimalno poravnanje algoritmom Smith – Waterman, implementiranog za paralelno izvođenje na grafičkim karticama, u ovom je radu bilo moguće iskoristiti prednosti koje takvo poravnanje nudi. Prednost optimalnog poravnanja je posebice uočljiva u vrstama čiji je genom loše kvalitete (brojne greške, praznine i umetanja).

Za proteine u sisavcima sličnih sljedova pokušano je pronaći veći broj eksona od onog trenutno prijavljenog na Ensembl-u. Pristup se sastojao u poravnavanju eksona dobro anotirane vrste (poput čovjeka) na regije gena ortolognih proteina u njemu srodnim vrstama. Pokazalo se da je ovim pristupom moguće ispraviti neke od grešaka koje su rezultat korištenja heurističkog poravnanja u potrazi za eksonima. Za svaki je ekson utvrđen izvor najkvalitetnije informacije (Ensembl ili SW poravnanje). Ova se informacija koristi za naknadnu rekonstrukciju proteinskog slijeda. Cilj je bio popraviti kvalitetu dostupnih proteinskih sljedova, čime bi se značajno olakšala komparativna analiza proteina.

Nedostatak ovog pristupa je nemogućnost provjere postojanja signalnih sljedova koji obilježavaju početak i kraj ekson / intron regija. Ipak, ako bi se ovaj pristup koristio kao preliminarni korak pronalaženja "grubih" regija eksona, bilo bi moguće sačuvati veći dio informacije od one trenutno dostupne na Ensembl-u. Također, moguće unapređenje ovog pristupa bilo bi korištenje algoritma za optimalno lokalno poravnanja koji je u mogućnosti prijaviti ne samo najbolje, već i suboptimalna poravnanja.

Literatura

[1] Curwen, V.; Eyras E.; Andrews, D.: The Ensembl Automatic Gene Annotation System, *Genome Research (14)*, pp. 934-941, USA, 2004

[2] Potter, S.C.; et al.: The Ensembl Analysis Pipeline, *Genome Research (14)*, pp. 942-950, USA, 2004

[3] Haiminen, N.; et al.: Evaluation of Methods for *De Novo* Genome Assembly from High-Throughput Sequencing Reads Reveals Dependencies That Affect the Quality of the Results. Plos ONE 6(9): e24182, USA, 2011.

[4] Trust Sanger Institute.: Havana, s Interneta, http://www.sanger.ac.uk/research/projects/vertebrategenome/havana/, 30. svibnja 2012.

[5] Burge, C; Karlin, S.: Prediction of complete gene structures in human genomic DNK. *J. Mol. Biol.* 268: 78-94., USA, 1997.

[6] Burge, C: Modelling dependencies in pre-mRNA splicing signals. *Computational Methods in Molecular Biology*, pp. 127-163, Nizozemska, 1998

[7] Birney, E.; Clamp, M; Durbin, R..: Genewise and Genomewise, *Genome Research (14)*, USA, 2004

[8] Uniprot Consortium : Criteria used to assign the PE level of entries, http://www.uniprot.org/docs/pe_criteria, 28/05/2012

[9] Korpar, M. (2011): *Implementacija Smith Waterman algoritma koristeći grafičke kartice s CUDA arhitekturom*, Bcc thesis. Fakultet eletrotehnike I računarstva, Republika Hrvatska

[10] Genome 10K Community of Scientists: Genome 10K: A Proposal to Obtain Whole-Genome Sequence for 10 000 Vertebrate Species, *Journal of Heredity*, USA, 2009

[11] The 1000 Genomes Project Consortium: A map of human genome variation from population – scale sequencing, *Nature*, USA, 2010

[12] Lindblad-Toh, K.; et al : A high resolution map of human evolutionary constraint using 29 mammals, *Nature*, pp. 476-482, USA, 2011

[13] Altschul, S.; et al : Basic local alignment search tool, *Journal of Molecular Biology* **215** *(3)*, pp. 403-410, USA, 1990

[14] Tatusov, R.L.; et al : The COG Database: an updated version includes eukaryotes, *BMC Bioinformatics* **4:**41, USA, 2003

[15] Altenhoff, A.M.; Dessimoz, C. : Phylogenetic and functional assessment of orthologs inference projects and methods, *PLOS Computational Biology 5(1)*, USA, 2009

[16] Johnson, T. : Reciprocal best hist are not a logically sufficient condition for orthology, arXiv:0706.0117v1 [q-bio.GN], 2008

[17] Haider, S. et al : BioMart Central Portal – unified access to biological data, *Nucleic Acids Research,* Vol. 37, pp. W23-W27, USA, 2009

[18] camacho, C.; Madden, T.; Coulouris, G. et al : BLAST Command Line Applications User Manual, 2008 Jun 23 [Updated 2012 Jan 30]. In: BLAST® Help [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2008-.

[19] Watson, M.; Birney, E. : Frameshifts vs. introns, s Interneta, http://lists.ensembl.org/ensembl-dev/msg00757.html, 29. svibnja 2012.

[20] Tao, T.. : BLAST XML output, s Interneta, ftp://ftp.ncbi.nlm.nih.gov/blast/documents/xml/README.blxml, 30. svibnja 2012.

[21] Vielella, A.J.; Severin, J.; Ureta-Vidal, A. et al : EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates, *Genome Research* **19**(2), pp. 327-335, USA, 2009

[22] Tatusov, R.L.; Koonin, E.V.; Lipman, D.J. : A genomic perspective on protein families, *Science* **278** (5338), pp. 631-637, USA, 1997

[23] Muller, J. et al : eggNOG v2.0: extending the evolutionary genealogy of genes with enhanced non-supervised orthologous groups, species and functional annotations, *Nucleic Acids Research* **38** (Database issue), pp. 190-195, USA, 2010

[24] Ostlund, G. et al : InParanoid 7: new algorithms and tools for eukaryotic orthology analysis, *Nucleic Acids Research* **38** (Database issue), pp. 196-203, USA, 2010

[25] Waterhouse et al : OrthoDB: the hierarchical catalog of eukaryotic orthologs in 2011, *Nucleic Acids Research* **39** (Database issue), pp. 283-288, USA, 2011

[26] Chen, F. et al : OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups , *Nucleic Acids Research* **34** (Database issue), pp. 363-368, USA, 2006

[27] Ranwez, V. et al : OrthoMAM: a database of orthologous genomic markers for placental mammal phylogenetics, *BMC Evol. Biol.* **7**, USA, 2007

[28] Smith, T.F.; Waterman, M.S.: Identification of common molecular subsequences, *Journal of Molecular Biology* **147**, pp. 195-197, USA, 1981

[29] Pavy, N. et al.: Evaluation of gene prediction software using a genomic data set: application to Arabidopsis thaliana sequences, *Bioinformatics* **15**: pp. 887-899, Fracuska, 1999

[30] Boeckmann, B.: "The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003," Nucleic Acids Research, vol. 31, no. 1, pp. 365-370, Jan. 2003.

[31] Pruitt, K.D.; Tatusova, T.; Klimke, W.; Maglott, D.R.: "NCBI Reference Sequences: current status, policy and new initiatives.," Nucleic acids research, vol. 37, no. Database issue, pp. D32-6, Jan. 2009.

[32] Loveland, J.E. Et al: "Community gene annotation in practice.," Database : the journal of biological databases and curation, vol. 2012, p. bas009, Jan. 2012.

[33] Slater, G.S.C.; Birney, E.: "Automated generation of heuristics for biological sequence comparison.," BMC bioinformatics, vol. 6, p. 31, Jan. 2005.

[34] Boeckmann Birney, E.; Durbin, R.: “Dynamite: a flexible code generating language for dynamic programming methods used in sequence comparison.,” International Conference on Intelligent Systems for Molecular Biology, vol. 5, pp. 56-64, Jan. 1997.

Sažetak

Program za automatsku analizu protein kodirajućih gena

Ključne riječi: protein, gen, ekson, anotacija genoma, najbolji uzajamni pogodak, homologija, optimalno poravnanje sljedova, heurističko poravnanje sljedova

Pri anotaciji genoma uobičajena je praksa koristiti heurističke algoritme poravnanja za pronalazak i obilježavanje intron – ekson regija gena pri anotaciji genoma. Zbog prirode heurističkog poravnanja, određen broj eksona nije uspješno identificiran, posebice u vrstama čiji je genom lošije kvalitete. Aplikacija SuperExonRetriever2000 razvijena je s ciljem uporabe optimalnog poravnanja sljedova za pronalak protein kodirajućh regija u genu. Aplikacija nudi mogućnost pronalaska ortolognih proteina u srodnim vrstama, dohvata i pohrane bioloških podataka (protein, gen, eksoni) te generiranja 4 vrste poravnanja eksona. Za svaki se protein može generirati statistika s postocima pokrivenosti pojedinog eksona za svako od četiri vrste poravnanja. Ovo omogućava detaljan pregled slučajeva u kojima se uporabom heurističkih metoda ne može pronaći ekson. Nadalje, iz svakog se poravnanja može rekonstruirati proteinski proizvod.

Aplikacija nudi dohvat svih informacija o biološkim sljedovima i poravnanjima na jednostavan i sistematičan način. Za svaki ekson proteina određen je najbolji ponuđeni slijed (Ensembl ili poravnanje). Sve informacije dobivene učitavanjem i obradom poravnanja preslikane su u bazu podataka. Ova baza podataka iskorištena je za generiranje najboljeg proteinskog proizvoda i pregled dobivenih rješenja je uobličen na mrežnim stranicama s nadom da će ona biti korisna pri komparativnoj analizi proteina.

Summary

Application for automatic analysis of protein coding genes

Key words: protein, gene, exon, genome annotation, reciprocal best hit, homology, optimal sequence alignment, heuristic sequence alignment

It is common practice to use the heuristic sequence alignment methods in the genome annotation for finding the exon-intron gene structures. Due to the nature of the heuristic alignment methods, a certain number of exons seems to be missing, especially in the species with poor genome sequencing and assembly quality. The SuperExonRetriever2000 application has been developed in order to address this issue and to make use of the optimal alignment methods in the annotation of the protein coding regions. The application offers a systematic way to infer the putative protein orthologous in related species, to acquire and store the necessary biological data and to create four types of sequence alignments. Statistics can be generated for each of the proteins with coverage percentages indicating how successful each of the alignment methods was in locating exons. The optimal alignments can be translated to a protein sequence.

The simple and systematic way of retrieving biological data and alignment information from the application was used to map this data to a MySQL database. Since the goal was to offer the best protein product possible, the best sequences for each of the exons are stored in the database. This database has been made publicly available through a web interface with hopes that such information will be useful for the comparative protein analysis.

Dodatak A: Detaljan opis informacija iz .descr datoteke

Za proteine *pep:known* i *pep:novel* su dostupni ID-evi gena i transkripta. Za proteine *pep:genscan* ova informacija nije dostupna.

Lokacija gena je uvijek u sljedećem formatu:

tip_slijeda:naziv_assembly-ja:ID_slijeda:početna_lokacija_gena:završna_lokacija_gena:lanac

Informacije su redom:

1. Tip slijeda. Slijed može biti kromosom, contig, supercontig, scaffold ili genescaffold. **Contig** je skup DNK segmenata dobivenih sekvenciranjem koji se preklapaju. **Supercontig** je skup contiga koji se preklapaju. **Scaffold** je skup preklapajućih contig-a odijeljenih prazninama poznate duljine. **Genescaffold** je skup scaffolda nastao spajanjem različitih scaffolda na kojima se nalaze dijelovi istoga gena.
2. Naziv assembly-ja. Identifikator projekta sekvenciranja i mapiranja genoma.
3. ID slijeda. Kodno ime slijeda na kom se nalazi gen.
4. Početna i završna lokacija gena.
5. Lanac. Može biti *unaprijedni* lanac (označen brojem 1) ili *unazadni* lanac (označen brojem -1)

Dodatak B: Opis strukture mapa proteina

Svaki od proteina iz ispitnog skupa nalazi se u zasebnoj mapi. Ime mape odgovara identifikatoru proteina. U nastavku je opisana organizacija mape proteina.

PROTEIN_ID(eg.ENSP00000389666)

```
├── alignment
│   ├── blastn          poravnanje ljudskih eksona na gen vrste korištenjem BLASTn-a
│   ├── mafft           poravnanje proteina referentne vrste, proteina vrste s Ensembl-a
                        i proteina rekonstruiranog iz poravnanja
│   ├── SW
│   │   ├── exon        poravnanje ljudskih eksona na cDNK vrste korištenjem SW#-a
│   │   └── gene        poravnanje ljudskih eksona na gen vrste korištenjem SW#-a
│   └── tblastn         poravnanje ljudskih eksona na protein vrste, tBLASTn
│
├── annotation
│   └── genewise        lokacije ekson-intron regija za vrste bez dostupnih eksona
│
├── ENSP00000389666.descr    opisna datoteka s prijavljenim ortolozima
├── .status                  datoteka sa statusima izvođenja svih dijelova
                             aplikacije
└── sequence
    ├── assembled_protein    protein sastavljen iz SW/gene poravnanja
    ├── exon
    │   ├── ensembl          sljedovi DNK eksona preuzeti s Ensembl-a
    │   └── genewise         sljedovi DNK eksona s lokacija prijavljenih Genewise-om
    ├── exon_database        eksoni referentne vrste bez UTR regija
    ├── expanded_gene        proširene regije gena vrsta
    ├── gene                 regije gena vrsta
    └── protein              proteinski sljedovi vrsta
```

Dodatak C: Popis korištenih vrsta

1. Ailuropoda melanoleuca (panda)
2. Bos taurus (cow)
3. Callithrix jacchus (common marmoset)
4. Canis familiaris (dog)
5. Cavia porcellus (guinea pig)
6. Choloepus hoffmanni (two-toed sloth)
7. Dasypus novemcinctus (armadillo)
8. Dipodomys ordii (kangaroo rat)
9. Echinops telfairi (tenrec)
10. Equus caballus (horse)
11. Erinaceus europaeus (european hedgehog)
12. Felis catus (cat)
13. Gorilla gorilla (gorilla)
14. Homo sapiens (human)
15. Loxodonta africana (elephant)
16. Macaca mulatta (macaque)
17. Macropus eugenii (wallaby)
18. Microcebus murinus (mouse lemur)
19. Monodelphis domestica (opossum)
20. Mus musculus (mouse)
21. Myotis lucifugus (brown bat)
22. Nomascus leucogenys (white-cheeked gibbon)

23. Ochotona princeps (pika)
24. Ornithorhynchus anatinus (platypus)
25. Oryctolagus cuniculus (rabbit)
26. Otolemur garnettii (greater galago)
27. Pan troglodytes (chimpanzee)
28. Pongo abelii (orangutan)
29. Procavia capensis (hyrax)
30. Pteropus vampyrus (flying fox)
31. Rattus norvegicus (rat)
32. Sorex araneus (common shrew)
33. Spermophilus tridecemlineatus (squirrel)
34. Sus scrofa (pig)
35. Tarsius syrichta (philippine tarsier)
36. Tupaia belangeri (treeshrew)
37. Tursiops truncatus (dolphin)
38. Vicugna pacos (alpaca)

Podaci o sekvenciranju i sklapanju genoma mogu se naći na stranicama Ensembl-a. Potrebno je samo u izborniku "All genomes" na početnoj stranici odabrati vrstu. Za većinu vrsta dostupni su podaci i o projektu obilježavanja genoma. Primjer takve stranice za miša je http://www.ensembl.org/Mus_musculus/Info/Index (primjer dobro obilježene vrste).