UVOD

Smanjenje troškova sekvenciranja zajedno sa konstantnim unapređivanjem tehnologije dovodi do dupliranja broja proizvedenih podataka svakih sedam meseci u prethodnih 10 godina. Samo u 2018. godini sekvencirano je više od 1000 genoma, dok u onlajn bazama podataka do danas postoji oko 15 000 sekvenciranih genoma (Mukherjee et al, 2017). Ovakve količine podataka doprinose poznavanju velikog broja proteinskih sekvenci, međutim za većinu njih postoji malo ili nimalo podataka o proteinskim funkcijama. Prema podacima iz 2007., samo 20% proteinskih sekvenci vrste *Homo sapiens* je eksperimentalno karakterizovano (Lee et al, 2007). Problemi eksperimentalnog određivanja funkcije proteina kao što su veliki vremenski i novčani izdaci doprineli su razvijanju velikog broja sofisticiranih kompjuterskih metoda na čijem se unapređivanju svakodnevno radi.

**Gene Ontology**

Velika količina podataka o funkcijama proteina je uskladišten u bazi podataka Gene Ontology (GO). Cilj GO je da pruži dinamičan i kontorlisan rečnik, koji može biti primenjen na sve vrste i koji će bitistalno ažuriran novim saznanjima, koja se svakodnevno akumuliraju i menjaju. GO sadrži 600 000 eksperimentalno potvrđanih anotacija, koje kada se uzmu u obzir iste funkcije proteina u različitim organizmima, čine 6 miliona funkcionalnih anotacija. Pored baze podataka GO sadrži i softver za logičko pretraživanje ontologija, veb pristup anotacijama i analitičke alate za obradu GO baze podataka (Ashburner et al, 2000; http://www.geneontology.org/).

GO anotacija predstavlja iskaz o funkciji određenog gena ili genskog produkta. Svaka GO anotacija se sastoji od gena I termina koji je sa njim povezan određenom relacijom. GO sadrži podatke iz 140 000 publikovanih radova prikazanih u formi GO termina. Termini su organizovani u tri nezavisne ontologije koje zajedno čine logičku strukturu u formi direktnog acikličnog grafaTri glavne i nezavisne ontologije su - biološki proces (*biological process, BPO*), molekulska funkcija (*molecular function, MFO*) i ćelijska komponenta (*cellular component, CCO*). Termini u GO su obeleženi pristupnom šifrom koja je u formi GO:0000000 gde umesto nula stoje brojevi koji jedinstveno određuju konkretan termin.

BPO se odnosi na biološki cilj kojem gen ili genski produkt namenjen. Ovaj cilj se izvršava posredstvom jednog ili više molekulskih funkcija, i često uključuje hemijsku odnosno fizičku transformaciju. Primer BPO termina postavljenog visoko u grafu bioloških funkcija je “transport” (*transport*, GO:0006810), a primeri specifičnijih termina niže u grafu su “transport nukleotida” (*nucleotide transport*, GO:0006862) ili “transport cAMP” (*cAMP transport*, GO:0070730). Grana GO acikličnog grafa koja sadrži ove termine kao i odnose između termina prikazana je na Slici 1.

|  |
| --- |
|  |
| Slika 1 Primer dela acikličnog GO grafa koji sadrži BP termine i odnose između njih. U gorenjem desnom uglu se nalazi legenda koja objašnjava značenje strelica različitih boja, a ujedno i prikazuje sve vrste odnosa koji mogu postojati između dva termina. (https://www.ebi.ac.uk/QuickGO/term/GO:0070730) |

MFO se definiše kao molekulski proces koji može biti izvršen od strane samostalne makromolekulske mašinerije posredstvom fizičkih interakcija ili drugih molekularnih entiteta. U tom smislu, MF predstavlja aktivnost koju određeni genski produkt (ili kompleks) vrši. MF opisuje samo funkciju koju genski produkt ima, bez osvtranja na to gde i kada se ta specifična funkcija vrši, i može biti opisan iz dve perspektive, kao biohemijska aktivnost ili kao komponenta većeg sistema odnosno procesa sa specifičnom ulogom. Primer opštijeg MFO termina je “katalitička aktivnost” (*catalytic activity,* GO:0003824), dok je primer specifičnijeg termina “aktivnost DNK ligaze” (*DNA ligase activity,* GO:0003909) ili “katalitička aktivnost na DNK” (*catalytic activity, acting on DNA,* GO:0140097). Grafička reprezentacija ove grane GO grafa data je na Slici 2.

|  |
| --- |
|  |
| Slika 2 Primer dela acikličnog GO grafa koji sadrži BP i MF termine i odnose između njih. U gorenjem desnom uglu se nalazi legenda koja objašnjava značenje strelica različitih boja (https://www.ebi.ac.uk/QuickGO/term/GO:0003909) |

Termini u CCO predstavljaju lokaciju u odnosnu na ćelijske kompartmente i strukture u kojima se nalazi makromolekulska mašinerija u trenutku vršenja molekulskog procesa. Postoje dva načina na koji termini opisuju lokaciju - (1) u odnosu na celularne strukture (npr. “citoplazmatična strana plazmine membrane” (*cytoplasmic side of plasma membrane,* GO:0009898)) ili kompartmente (npr. “mitohondrija” (*mitochondrion,* GO:0005739)) i (2) kao deo stabilnog makromolekulskog kompleksa (npr. “ribozom” (*ribosome*, GO:0005840)). Grafička reprezentacija povezanosti termina GO:0005840 sa glavnim ontologijama prikazana je na Slici 3.

|  |
| --- |
|  |
| Slika 3 Primer dela acikličnog GO grafa koji sadrži CC i BPtermine i odnose između njih. U gorenjem desnom uglu se nalazi legenda koja objašnjava značenje strelica različitih boja (https://www.ebi.ac.uk/QuickGO/term/GO:0005840) |

Pored dodeljivanja termina, svaka GO anotacija sadrži i informaciju o svojoj pouzdanosti, odnosno izvoru koji dokazuje njenu tačnost. Ovakvi dokazi (*GO evidence codes*) su takođe grupisani hierarhijski i obeleženi kodovima. Postoji nekoliko glavnih nivoa pouzdanosti anotacija - (1) eksperimentalno podržane anotacije, (2) filogenetički izvendene anotacije, (3) kompjuterski izvedene anotacije i (4) ekstenzije anotacija.

**Metode za predviđanje funkcije proteina**

Metode zasnovane na homologiji

Najčešće i najpristupačnije metode za predikciju proteinske funkcije su zasnovane na homologiji proteina, i osnovna ideja koja se nalazi iza medote je da proteini sa sličnom sekvencom i strukturom imaju sličnu funkciju (Lee et al, 2007). Dva proteina se smatraju homolozima ukoliko potiču od iste predačke sekvence, i mogu biti otrolozi i paralozi, u zavisnosti od načina na koji su nastali. Ortolozi nastaju procesom specijacije i zbog toga se nalaze u različitim vrstama i tamo vrše istu ili sličnu funkciju, dok su paralozi nastali procesom duplikacije gena u okviru vrste i zbog toga imaju sličnu strukturu ali različitu funkciju. Alternativno, ukoliko dođe do procesa duplikacije gena pre specijacije, paralozi se takođe mogu naći i u okviru različitih vrsta (Fitch, 2000). Predikcija funkcije proteina na osnovu homologije se dakle zasniva na razlikovanju ortologa i paraloga što je često otežano različitim prirodnim procesima kao što su gubljenje gena, funkcionalno preklapanje, horizontalni transfer gena i slično (Lowenstein et al, 2009). Funkcionalna anotacija proteina je umnogome olakšana standardizacijom i digitalizacijom, za šta su zaslužni onlajn servisi kao što je GO konzorcijum. Metode za predikciju funckije proteina zasnovane na homologiji se mogu podeliti u dve velike grupe, na metode koje se zasnivaju na analizi strukture, i na metode koje se zasnivaju na analizi sekvence proteina.

Pri određivanju funkcije proteina metode zasnovane na analizi sekvence kao prvi korak podrazumevaju pretraživanje homologa nepoznate proteinske sekvence u nekim od onlajn baza podataka. Najekstenzivnije baze podataka pružaju Nacionalni centar za informacionu biotehnologiju (*National Center for Biotechnology Information,* NCBI*)* i Evropska laboratorija za molekularnu biologiju - Evropski bioinformatički institut (*European Molecular Biology Laboratory - European Bioinformatics Institute*, EMBL-EBI). Pretraživanjem proteinske (ili genske) sekvence u nekoj od ovih baza daje korisniku sistematičan pristup velikoj količini pouzdanih informacija, uključujući informacije o proteinskim familijama i domenima, funkcionalnim anotacijama i drugo. Ukoliko konkretna proteinska odnosno genska sekvenca ne postoji u NCBI ili EMBL-EBI bazama, može se pretražiti uz pomoć BLAST alata (*Basic Local Alignment Search Tool*) koji takođe nude NCBI i EMBL-EBI. BLAST pretraga kao rezultat daje listu sekvenci koje imaju određeni stepen homologije sa pretraživanom sekvencom. Funkcionalne anotacije ovih homologa je moguće iskoristiti za anotaciju preraživane proteinske sekvence, u zavisnosti od stepena homologije između njih. Određivanje sigurnog merila u kome je moguće preuzeti anotacije homologog proteina je ono što različite metode za predikciju funkcije proteina zasnovane na homologiji pokušavaju da utvrde. U rešavanju ovog problema postoji nekoliko do sada razvejnih pristupa, a oni uključuju grupisanje proteina u odnosu na evolutivnu bliskost, pripadanju određenim familijma proteina, klasterovanje sekvenci, pozitivnu diskriminaciju funkcionalnih aminokiselinskih rezidua i slično, ili kombinaciju ovih pristupa. (Lowenstein et al, 2009)

PANNZER – alat za funkcionalnu anotaciju proteina zasnovan na homologijama

PANNZER (Protein ANNotation with Z-scoRe) je online alat za predviđanje proteinskih GO anotacija i slobodnih tekstualnih opisa (Koskinen et al, 2015; Toronen et al, 2018). Razvijen je za vreme trajanja prvog CAFA (Critical Assessment of protein Function Annotation challenge) eksperimenta (Radivojac et al, 2013), a kasnije je nadograđen u online servis koji pruža brz, javno dostupan veb server za funkcionalnu anotaciju koji može da analizira desetine hiljada sekvenci za relativno kratak vremenski period. Ulazni podaci koje PANNZER zahteva su proteinske sekvence u FASTA formatu, dok izazne informacije koje alat pruža predstavljaju GO termini i slobodni tekstualni opisi. Postoje dva načina pregledanja rezultata upita koje daje PANNZER, oflajn, nakon preuzimanja, i onlajn uz pomoć web aplikacije koja pored predviđenih GO termina i slobodnih opisa daje i verovatnoće sa kojima su oni predviđeni. Uz pomoć web aplikacije takođe je moguće pronaći i protein koji je homolog upitnoj proteinskoj sekvenci i pratiti trag na osnovu kog je napravljena predikcija.

PANNZER2 radi kao klasifikator tipa *weighted k-nearest neighbour* koji je zasnovan na podudaranju homologih sekvenci i enričment statistici. Predviđanje anotacija za ulaznu proteinsku sekvencu se dešava u tri uzastopne etape. Prva etapa podrazumeva pretraživanje sekvenci homologih upitnoj. Ranije verzije PANNZER-a su za pretragu homologa koristile vremenski zahtevnu BLAST pretragu, koja je kasnije zamenjena SANSparallel pretraživačem homologa (Somervuo i Holm, 2012). SANSparallel je veb server koji pruža mogućnost brze pretrage homologa proteinske sekvence u UniProt bazi podataka. Kraj prve etape predstavlja formiranje liste koja sadrži proteine slične upitnom (*sequence neighborhood*, SN). Druga etapa predstavlja filtriranje SN liste, a kriterijumi koji proteini moraju da zadovoljavaju da bi prošli filtriranje je da im je minimum 40% sekvence identično sa upitnom, i da imaju minimum 60% pokrivenosti poravnanja (eng. *alignment coverage*). Nakon filtriranja, SN lista se sastoji od proteina sličnih ulaznom koji zadovoljavaju kriterijume i svih GO termina sa kojima su ti proteini asocirani. Treća etapa podrazumeva odabir proteina iz filtrirane SN liste od kojih će se GO anotacije “prepisati” upitnoj sekvenci. Ovaj korak se izvodi uz pomoć skoring funkcija, koje predstavljaju računski postupak koji u obzir uzima više faktora i na kraju daje određeni skor koji pokazuje verovatnoću da je transfer anotacija tačan. PANNZER2 nudi korisniku opciju odabira između više skoring funkcija, uključujući i skoring funkcije iz drugih alata - ARGOT (Falda et al, 2012), PANNZER (Koskinen et al, 2015), BLAST2GO (Conesa et al, 2005), kao i neke od standardnih funkcija - hipergeometrijsko poboljšanje (eng. *hypergeometric enhancement*) i najinformativniji pogodak (eng. *best informative hit*). Krajlnji rezultat predstavlja lista GO anotacija sa kojima se asocira upitna proteinska sekvenca, i u slučaju onlajn pregledanja rezultata verovatnoća da je predikcija tačna.

Metode koje nisu zasnovane na homolgiji

U genomima različitih vrsta postoje proteinske sekvence koje nije moguće anotirati uz pomoć metoda zasnovanih na homologiji. Iz tog razloga su razvijane metode koje nisu zasnovane na homologijama, i koje koriste druge karakteristike proteina kao što je dužina, aminokiselinski sastav, izoelektirčna tačka, subcelularna lokalizacija, posttranslacione modifikacije i slično (Lee et al 2007).

FRENKI - alat za funkcionalnu anotaciju proteina zasnovan na spektralnoj analizi sekvenci

FRENKI je razvijen od strane Centra za multidisciplinarna istraživanja i inženjering Instituta za nuklearne nauke Vinča (Gemović et al, 2017). Rad FRENKI-ja je zasnovan na ISM metodi (information spectrum method, metod informacionog spektra) i enričment analizi. ISM je virtuelna spektroskopska metoda koja se primjenjuje za istraživanje protein-protein interakcija kao i za analizu veze strukture i funkcije proteina i aminokiselina. ISM spektar nekog proteina se računa tako što se svakoj aminokiselini u proteinskoj sekvenci se dodeljuje brojčana vrednost koja je jednaka njenoj EIIP (electron ion interaction potential) vrednosti, odnosno potencijalu za ostvarivanje elektron-jon interakcije. EIIP se računa uz pomoć AQVN (average quasivalence number) odnosno prosečnog kvazivalentnog broja, koji opisuje prosečnu valentnost atomskih komponenata u odnosu na ceo molekul. Ovim se dobija numerička sekvenca na koju je moguće primeniti Furijeovu transformaciju, čime se dobija ISM spekar. Enričment program koji koristi FRENKI je DINGO (Davidović et al, 2018), koji predstavlja adaptaciju Cytoscape enričment programa (Shannon et al, 2013).

FRENKI računa predikcije GO termina za upitnu protiensku sekvencu u nekoliko koraka - (1) ISM analiza ulaznog proteina. U dobijenom ISM spektru se identifikuje dominantan pik, odnosno pik sa najvećom amplitudom. Ovaj pik odgovara funkcionalnom domenu proteina. (2) ISM analiza kompletnog humanog proteoma. (3) Selekcija proteina iz humanog proteoma koji u svojim ISM spektrima među prva dva dominantna pika imaju dominantan pik iz ulazne sekvence. (4) DINGO enričment analiza selektovanih proteina. Selektovana lista proteina prolazi kroz enričment analizu koja kao rezultat daje GO termine koji su zajednički za ovu grupu proteina. Finalno, sve GO termine koje DINGO da kao rezultat, postaju predikovani termini za upitnu proteinsku sekvencu.

šššššššššššššššššššššššššema

**CAFA**

CAFA (Critical Assessment of protein Function Annotation challenge) je globalna inicijativa za evaluaciju i unapređivanje kompjuterskih metoda za funkcionalnu evaluaciju proteina. CAFA se organizuje u vidu “izazova” gde se istraživačkim timovima dostave neanotirane proteinske sekvence za koje je potrebno predvideti GO anotacije, u cilju objektivne evaluacije različitih metoda i unapređenja istih. CAFA je do sada organizovana dva puta, CAFA1 u periodu 2010-2011 (Radivojac et al, 2013), i CAFA2 u periodu 2013-2014 (Jiang et al, 2016), dok je CAFA3 trenutno u toku. U CAFA1 i CAFA2 izazovima evaluirano je skoro 200 različitih predikcionih algoritama, a učestvovalo je više desetina istraživačkih grupa. Oba CAFA izazova su funckionisala po istom principu, a razlikovali su uglavnom se po načinu evaluacije. CAFA izazovi obuhvataju nekoliko faza. U početnom trenutku izazova timovima se dodeljuje identičan set trenutno u GO konzorcijumu neanotiranih proteina. Timovi imaju određeni vremenski period da pošalju svoje predikcije zajedno sa verovatnoćama tačnosti predikcija, nakon čega sledi period akumulacije anotacija. Nakon isteka određenog roka od nekoliko meseci u toku kog su se anotacije akumulirale, CAFA tim formira evaluacioni set proteina koji obuhvata one proteine koji su za period akumulacije nakupili dovoljan broj ekperimentalno dokazanih anotacija. Poslednji korak predstavlja evaluaciju, gde CAFA tim testira koliko se predikcije učesnika izazova poklapaju sa eksperimentalno predviđenim podacima. U oba izazova svi učesnički alati su osim međusobno poređeni sa dva osnovna algoritma - BLAST koji “prepisuje” sve GO anotacije proteinu sa njegovih homologa iz BLAST pretrage i kao verovatnoću tačnosti daje procenat poklapanja upitne sekvence i homologa, i NAIVE koja takođe prepisuje anotacije homologa ali kao verovatnoću tačnosti daje učestalost konkretnog termina u kompletnoj GO bazi. Glavna mera koja je korišćena u CAFA izazovima je F skor, jedinstvena mera na osnovu koje se može proceniti odnos preciznosti i obuhvatnosti predikcionog alata.

PANNZER i FRENKI su učestvovali u CAFA izazovima i odabrani su u cilju poređenja različitih algoritama u uspešnosti predikcije lokalizacije proteina u mitohondrijama. Mere korišćene za poređenje alata u ovom radu su takođe preuzete iz CAFA izazova.

**Mitohondrijalni proteini**

Mitohondrije su dinamične organele esencijalne za mnoge ćelijske procese, a između ostalog rast, diferencijaciju i ćelijsku smrt. Učestvuju u velikom broju metaboličkih puteva i predstavljaju centar za homeostazu jona i apopotozu (Friedman, 2014). Disfunkcija mitohondrija može doprineti razvoju preko 50 različitih bolesti čije manifestacije variraju od neonatalne smrti pa do neurodenerativnih oboljenja u starosti. Takođe se pretpostavlja da je malfunkcija mitohondrija povezana sa razvojem dijabetesa tipa II i karcinoma (Pagliarini et al, 2008). Mitohondrijalni genom je sekvenciran 1981. (Anderson et al) i od tada je poznato da on sadrži sekvence za sintezu 13 proteina. Ipak, pretpostavlja se da se mitohondrijalni proteom sastoji od oko 1500 proteina čije se DNK sekvence nalaze u nukleusu, a kompletan sastav mitohondrijalnog proteoma i dalje nije poznat (Calvo et al, 2015). Određivanju sastava mitohondrijalnog proteoma pristupa se na različite načine, uključujući eksperimentalne metode kao što je masena spektroskopija i mikroskopija, ali i kompjuterski zasnovane metode, kao što je mašinsko učenje i kompjuterska funkcionalna anotacija (Pagliarini et al, 2008). S obzirom na veliki broj bolesti koji je povezan sa mitohondrijalnom disfunkcijom, utvrđivanje sastava mitohondrijalnog genoma i unapređivanje metoda za njihovo predviđanje je od velikog značaja za medicinu i nauku.

**Predviđanje lokalizacije proteina u mitohondrijama**

Fali deo koji opisuje alate za predikciju mito lokalizacije i pasus o IMPIju, za koji čekam literatutu da mi pošalji z te laboratorije

**Cilj** ovog diplomskog rada je bio uporediti uspešnost dva različita alata za predviđanje lokalizacije proteina u mitohondrijama. Alati koji su upoređivani bili su FRENKI, čiji se rad zasniva na analizi proteinskih sekvenci korišćenjem matematičkog modelovanja fizičko-hemijskih osobina aminokiselina (zasnovan na ISM) i PANNZER, čije se predviđanje zasniva na evolutivnoj analizi homologih proteina. Za evaluaciju alata korišćen je R programski jezik.

MATERIJALI I METODE

Poređena je uspešnost softvera za funkcionalnu anotaciju u predviđanju lokalizacije humanih proteina u mitohondrijima. Eksperiment je obuhvatao predviđanje funkcija za kompletan ljudski proteom od strane FRENKI-ja i PANNZER-a, čišćenje i pripremu predikcija za statističku analizu, i samu statističku analizu, odnosno poređenje broja tačnih i netačnih predikcija. Evaluacija je rađena naspram referentnog seta mitohondrijalnih proteina.

**Referentni set mitohondrijalnih proteina**

Referentni set je sastavljen uz pomoć IMPI-ja, odnosno proteina koje kodiraju geni sa IMPI liste. IMPI lista gena je dostupna na vebsajtu laboratorije u kojoj je razvijen (<http://www.mrc-mbu.cam.ac.uk/impi>), a odgovarajući sekvence proteina, kao njihovi identifikacioni brojevi (ID) su preuzeti pomoću UniProt alata (<http://www.uniprot.org/uploadlist>). Korišćeni su samo proteini statusa “ručno recenzirani” (*reviewed*), i ovakva lista proteina se sastojala od 1547 proteina. Da bi evaluacija odabranih predikcionih alata bila validna, iz test seta proteina su izbačeni oni proteini čije su funkcionalne anotacije od ranije poznate i koji su korišćeni za kreiranje PANNZER i FRENKI alata zbog toga su iz referentnog seta mitohondrijalnih proteina izbačeni svi oni proteini koji su u GO bazi podataka trenutno anotirani kao mitohondrijalni, ili se nalaze u CAFA1 trening setu. Proteini anotirani kao mitohondrijalni u GO su izbačeni zbog toga što je PANNZER metoda zasnovana na homologijama, i pretraživanje proteina uz pomoć BLAST pretrage bi automatski pokazalo njihovu lokalizaciju u mitohondrijama, što umanjuje validnost evaluacije. Takođe, proteini iz CAFA1 trening seta su izbačeni zbog toga što predstavlja trening set na kome je obučavan FRENKI. Finalna referentna lista sa proteinima sa pouzdanim dokazima za lokalizaciju u mitohondrijama, odnosno referentni set mitohondrijalnih proteina, sastojao se od 653 proteina.

U cilju dobijanja predikcija, kao upitne sekvence u oba softvera je postavljan ceo ljudski proteom. Ljudski proteom je u trenutku skidanja fajla sa UniProta-a (10. maj 2018.) imao 20 373 proteina.

Predikcije dobijene od FRENKI-ja i PANNZER-a su bile u obliku liste proteina zajedno sa svim GO treminima sa kojima su oni asocirani. S obzirom da se za protein vezuju ne samo predikovani termini, već i svi njihovi roditeljski termini, pre početka analize predikcije je bilo potrebno propagirati do najstarijih GO termina. Propagacija termina obuhvata proces u kome se listi predikcija dodaju svi roditeljski termini svakaog konkretnog termina, u cilju dobijanja liste svih mogućih termina sa kojima je protein asociran. Takođe, liste proteina su očišćene od potencijalnih smetnji, kao što su duplikati ili NA vrednosti. Ovako dobijene liste su filtrirane, i iz njih su izvučeni samo proteini koji su asocirani sa terminom GO:0005739. Termin GO:0005739 - mitochondrion je najstariji CC termin koji obuhvata sve termine vezane za lokalizaciju u mitohondrijama. Na slici 4 je prikazan njegov položaj u GO stablu kao i položaj svih nižih termina koje obuhvata. Iz ovakao filtriranih lista predikcija izbačeni su svi proteini koji su u tom trenutku u GO bazi već anotirani kao GO:0005739 ili se nalaze u CAFA trening setu. Kako su ti proteini eliminisani i iz referentnog seta mitohondrijalnih proteina, njihovo pojavljivanje u setu predikcija bi predstavljalo lažne pozitivne vrednosti, te zbog toga su izbačeni. Nakon

|  |
| --- |
|  |
| Slika 4. Prikaz GO termina “mitochondrion” (GO:0005739) i njegovih roditeljskih i potomačkih temrina. (http://amigo.geneontology.org/amigo/term/GO:0005739) |

Kao mera efikasnosti predviđanja ćelijske lokalizacije u mitohondrijama korišćena je F mera, preuzeta iz CAFA izazova kao jedan od načina evaluacije softvera za anotaciju (Radivojac et al, 2013). Računanju F mere prethodi procena preciznosti i odziva metode koju alat koristi. Preciznost pokazuje koji broj od ukupnih predikcija je tačan i računa se kao količnik broja tačnih predikcija (*true positives*) i broja ukupnih predikcija (*true positives + false positives*). U ovom slučaju preciznost predstavlja količnik broja proteina kojima je alat tačno predvideo mitohondrijalnu lokalizaciju i broja ukupnih proteina sa predviđenom mitohondrijalnom lokalizacijom (1.1). Odziv govori o broju pogođenih predikcija (*true positives*) od strane alata u odnosu na ukupan broj proteina iz referentne liste (*true positives + false negatives*) (1.2). F mera se koristi za poređenje različitih alata jer u jednom broju daje informaciju i o preciznosti i o odzivu, a predstavlja harmonijsku sredinu ova dva broja (1.3).

|  |  |
| --- | --- |
| (1.1) |  |
| (1.2) |  |
| (1.3) |  |

Sva poređenja i obrada podataka rađeni su u R programskom jeziku koristeći RStudio interfejs. Finalne analize bez pripreme fajlova su objedinjene u jednu skriptu dostupnu u prilogu na kraju rada zajedno sa skriptom u kojoj se nalazi sav kod za pravljenje grafika upotrebljenih u radu. Kompletan kod koji je pisan pri izradi diplomskog rada, zajedno sa literaturom, graficima i beleškama otvoren je za sve na GitHub-u (https://github.com/aleksicmil/diplomski).

REZULTATI

Poređena je uspešnost predikovanja ćelijske lokalizacije proteina u mitohondrijama za dva alata zasnovana na različitim metodama. Mere kojima je ocenjivana njihova uspešnost su preciznost metode, odziv, i F mera.

Nakon čišćenja seta predikcija i ekstrahovanja predikcija za GO:0005739, set predikovanih mitohondrijalnih proteina FRENKI-ja sastojao se od 2747 anotacija. Broj tačnih predikcija (*true positives*) bio je 99, što znači da su ostalih 2648 proteina iz seta FRENKI predikcija činili lažne pozitivne predikcije (*false positives)*. Odnos broja tačnih predikcija i ukupnog broja predikcija predstavlja preciznost, koja za ovaj alat iznosi 0,036. Ostalih 554 proteina iz evaluacionog seta koje FRENKI nije predvideo kao mitohondrijalne čine lažne negativne predikcije (*false negatives*). Odziv ovog alata, tj. odnos broja tačnih predikcija i broja proteina u evaluacionom setu je 0,151. Ovakvi rezultati pokazuju da je FRENKI tačno predvideo lokalizaciju za oko 15% proteina iz referentnog seta mitohondrijalih proteina, i da oko 3% ukupnog seta predikcija čine tačne predikcije. F mera za FRENKI-ja izosila je 0,058. Ovi podaci su prikazani grafički Venovim dijagramima na slici 5.

|  |
| --- |
|  |
| Slika 5. Grafički prikazani rezultati evaluacije FRENKI-ja. Imena setova prikazana su iznad skupova, kao i broj proteina u svakom skupu. Veličina skupova je proporcionalna broju proteina u setovima. |

Set PANNZER-ovih predikcija za ćelijsku lokalizaciju proteina iz ljudksog proteoma se nakon čišćenja sastojao od 951 anotacije za GO:0005739. Evaluacija je rađena uz pomoć referentnog seta od 653 mitohondrijalnih proteina. Broj tačnih predikcija (*true positives*) za ovaj alat predstavljao je 343, što znači da je odziv alata 0,52. To znači da je PANNZER tačno pogodio gotovo polovinu evaluacionog seta, kao što se može videti sa slike 6, gde su grafički prikazani rezultati evaluacije. Broj predikcija koje su netačne (*false positives*) iznosio je 608, što ukazuje na preciznost alata od 0,36. Dakle, nešto više od trećine seta predikcija PANNZER-a je tačno. F mera za ovaj alat iznosi 0,427.

|  |
| --- |
|  |
| Slika 6. Grafički prikazani rezultati evaluacije PANNZER-a. Imena setova prikazana su iznad skupova, kao i broj proteina u svakom skupu. Veličina skupova je proporcionalna broju proteina u setovima. |

Pored preklapanja sa referentnim setom, utvrđena su i preklapanja između predickionih setova međusobno. Grafički prikaz ovog poređenja je na Slici 7. Interesantno je primetiti da od 343 tačne predikcije koje je PANNZER napravio samo 55 deli sa FRENKI-jevim predikcijama. Takođe, između 2648 pogrešnih predikcija FRENKI-ja i 608 pogrešnih predikcija PANNZER-a, zajedničke su samo 82 predikcije. Ovo ukazuje na to da se ova dva alata ponašaju komplementarno prilikom predviđanja lokalizacije proteina u mitohondrijama.

|  |
| --- |
|  |
| Slika 7. Preklapanja između predikcionih setova međusobno, i referentnog seta. Ispod imena skupa napisana je i broj proteina u njemu. Veličina skupova je proporcionalna broju proteina u svakom setu. |

DISKUSIJA

Kako se broj sekvenci dostupnih u javnim bazama podataka svakodnevno povećava, javlja se sve veća potreba za razvojem metoda za kompjutersku funkcionalnu anotaciju (Clark i Radivojac, 2013). Takođe, kompletno određivanje mitohondrijanog proteoma je od izuzetne važnosti za nauku i medicinu, a u velikoj meri se oslanja na *in silico* metode (Pagliarini et al, 2008). U ovom radu testirana je uspešnost predickije ćelijske lokalizacije proteina u mitohondrijama dva alata zasnovanih na suštinski različitim metodama. Pokazao je da je PANNZER, alat zasnovan na homologijama, uspešniji u predikovanju lokalizacije proteina u mitohondrijama u odnosu na ISM zasnova FRENKI. Takođe, ispitano je i preklapanje u predikcijama dva izabrana alata. Dobijeni rezultati ukazuju na malo prekapanje između setova tačnih i netačnih predikcija, što govori o komplementarnosti ova dva alata. Analize ovog tipa imaju za cilj da doprinesu razvitku alata kako za funkcionalnu anotaciju generalno, tako i za specijalizovanu anotaciju proteina za određene delove ćelije.

Postoji više radova koji se bave evaluacijom alata za funkcionalnu anotaiciju (You et al, 2017; Koskinen et al 2013, Toronen et al 2015, Gong et al, 2016), ali je najekstenzivnija i najinformativnija evaluacija rađena u okviru CAFA izazova. Zbog obima i značajnosti CAFA izazova, važno je rezultate ovog rada prodiskutovati u svetlu CAFA-inih. Rezultati CAFA2 izazova pokazuju da se uspešnost anotacije alata značajno razlikuje u sve tri ontologije. Razlozi za to mogu biti različiti, uključujući i (1) osobine same ontologije, kao što je veličina, dubina ili faktor grananja; (2) apstraktnost termina, BP ontologije se generalno smatraju aprstraktnijim od MF i CC termina; (3) zastupljenost konkretne ontologije u trening setu, odnosno dubina anotacija iz konkretne ontologije (Schnoes et al, 2013). Rezultati pokazuju da alati za anotaciju imaju najviše uspeha u predviđanju MF ontologije, i da u toj kategoriji premašuju uspeh BLAST i Naive metoda. Ono što je interesantno je da u kategoriji predviđanja CC ontologija, F mera većine alata jedva prelazi F meru postignutu Naive metodom. Autori objašnjavaju da je to verovatno posledica velike zastupljenosti termina koji su površno u GO stablu (na primer GO:0043226 “organelle”, drugi nivo; ili GO:0044424 “intracellular part”, treći nivo) koje Naive lako pogađa i tako povećava svoju F meru (Jiang et al, 2013). Na Slici 8 prikazane su F mere prvih 10 najuspešnijih alata iz CAFA2 izazova u predviđanju CC ontologija. Može se zapaziti da F mera PANNZER-a od 0,427 se nalazi relativno blizu F mera ovih metoda, što ukazuje na dobru uspešnost ovog alata. Pri poređenju ovih F mera, važno je imati u vidu da se F mere prikazane na Slici 8 računaju na drugačiji način u odnosu na to kako je račnunato u ovom radu. F mera iz CAFA-e se računa kao srednja vrednost za sve termine, bez obzira na njihovu dubinu. Ukoliko se neki termin nalazi plitko u GO stablu, njega je mnogo lakše tačno predvideti u odnosu na konkretan termin sa dna GO stabla. S obzirom da je GO:0005739 relativno konkretna ontologija (5. nivo dubine), a F mera računata u CAFA izazovu je aprokismirana na pojmove sve dubine, ovu razliku u računanju F mere treba uzeti u obzir pri njihovom poređenju. Za sada neobjavljeni rezultati CAFA-e 3 pokazuju da metode najbolje u predikciji CCO imaju F meru od oko 0,62 što ukazuje na znatan napredak metoda između dva CAFA izazova. U CAFA-i 3 PANNZER takođe spada među najuspešnije metode, i manja F mera dobijena u ovom radu ukazuje na to da je PANNZER generalno bolji u predikovanju CCO nego konkreno lokalizacije u mitohondrijama. Isto se može zapaziti i za FRENKI, čiji F skor za CCO u CAFA-i 3 iznosi 0,36. Ovo ukazuje na veliku potrebu razvinja softvera za anotaciju proteina koji su usko specijalizovani.

|  |
| --- |
|  |
| Slika 8. F mere deset najuspešnijih alata u CAFA2 izazovu u predviđanju CC anotacija. Svi proteini iz seta na kojem je rađena evaluacija na početku izazova nisu imali nikakve anotacije iz bilo koje ontologije, i pripadali su proteomima više različitih vrsta. C vrednosti prikazane na slici odnose se na pokriće evaluacionog seta koji je svaki od alata imao. (Preuzeto od Jiang et al 2013) |

FRENKI je alat za funkcionalnu anotaciju proteina koji je zasnovan na ISM metodi, što je važno napomenuti s obzirom da se, za razliku od PANNZER-a, prilikom pravljenja predikcija ne oslanja na evolutivni pristup I korišćenje homologih proteina. Ovakve specifikacije čine ovaj alat pogodnim za korišćenje u situacijama kada ulazni protein nije dobro okarakterisan ili nema dovoljno homologih sekvenci na osnovu kojih bi se mogle izvesti valjane evolutivno-zasnovane predikcije. Niža F mera je delimično posledica velikog broja lažno pozitivnih predikcija, što bi se moglo smanjiti povećanjem trešhold vrednosti prilikom enričment faze anotacije, ili zadržavanjem na terminima koji nisu toliko duboko u GO stablu, koji bi predstavljali manje precizne ali pouzdanije predikcije. Takođe, interesanto je primetiti da je F mera FRENKI-ja dosta niža u ovom istraživanju (0,058) u odnosu na F meru u CAFA izazovu (0,36). Ovo ukazuje na to da je FRENKI značajno lošiji u predikovanju lokalizacije proteina u mitohondrijama u odnosu na predikciju CCO uopšte. To može biti posledice specifičnosti sekvenci mitohondrijalnih proteina. Naime, najveći deo proteina iz mitohondrijalnog proteoma je kodirano u nuklearnom DNK, što znači da postoji određeni sistem za raspoznavanje specifično mitohondrijalnih proteina. Mitohondrijalne target sekvence su N-terminalni delovi sekvenci koji služe za obeležavanje proteina čija je lokalizacija u mitohondrijama, i koji se nakon dopremanja proteina na određeno mesto odsecaju (Clarlos et al, 1996). Kako u ISM spektru pik najveće amplitude odgovara najčešće funkcionalnom domenu, a ne target sekvenci, moguće je da FRENKI na ovaj način previđa neki deo mitohondrijalnih proteina. Ovome u prilog ide i informacija da postoje proteini koji imaju istu ili sličnu funkciju (a time i slične ISM spektre), a različite lokalizacije, kao što je na primer topoizomeraza 1 (Pommier et al 2016).

Za razliku od FRENKI-ja, PANNZER pri donošenju predikcija koristi anotacije homologa upitne sekvence, i prilikom “prepisivanja” anotacija u obzir uzima više različitih faktora. Podrazumevana skoring funkcija u PANNZER2 koja je zadužena za odlučivanje sa kog će se homologa i koje anotacije “prepisivati” na upitnu sekvencu analizira različite podatke - procenat identičnosti sekvenci, poravnanje sekvenci, taksonomsku udaljenost (u NCBI taksonomskom stablu kao aproksimaciju evolutivne udaljenosti), učestalost GO termina u GO bazi, učestalnost pojavljivanježa GO termina među ostalim homolozima za tu upitnu sekvencu itd. Prilikom izrade PANNZER alata korišćena je metoda mašinskog učenja u cilju treniranja softvera da svaki od navedenih parametara rangira po važnosti i tako doprinese maksimizaciji broja tačnih predikcija (Koskinen et al, 2015). Jedan od zaključaka CAFA2 izazova jeste da najupešnije metode za funkcionalnu anotaciju kombinuju više različitih tipova podataka kao i podatke od više zasebno potencijalno nepouzdanih sekvenci prilikom donošenja predikcija (Jian et al, 2016). Imajući ovo u vidu, visoka F mera koju je postigao PANNZER u ovom eksperimentu nije iznenađujuća, kao ni činjenica da se F mera PANNZER-a iz ovog rada nalazi među 10 najuspešnijih metoda iz CAFA2 izazova (Slika 8). Ipak, najveća F mera u CAFA3 izazovu za predikovanje CCO je oko 0,63, a kako je PANNZER jedna od najuspešnijih metoda u tom izazovu, može se zapaziti da ima slabiju mogućnost predikcije mitohondrijalnih proteina nego predikcije lokalizacije proteina generalno.

Interesantno je primetiti da je preklapanje kako između tačnih tako i između netačnih predikcija između alata veoma malo. Kao što se može videti sa slike 7, samo 82 proteina su zajednička za lažno pozitivne predikcije, što čini oko 13% PANNZER lažnih pozitivnih predickija i 3% FRENKI lažnih pozitivnih predikcija. Ovo je verovatno posledica toga što ova dva alata za anotaciju koriste u potpunosti različite pristupe. Takođe, važno je napomenuti da gotovo polovina tačnih predikcija FRENKI-ja nije zajednička sa PANNZER-om. Ovo ukazuje na to da ISM kao metoda može doprineti proširenju seta tačnih predikcija. Kako je u ranijim CAFA izazovima i potvrđeno (Radivojac et al 2013; Jiang et al, 2016), pravilna implementacija više metoda u okviru jednog alata dovodi do ostvarivanja najboljih rezultata. U usko specifičnim istraživanjima ovog tipa moguće je dodatno ispitati proteine koji se nalaze u ovim poljima preklapanja. Određene strukturne sličnosti ili poklapanje u sekvenci, pripadanje specifičnim familijama proteina, postojanje konkretnog strukturnog domena ili slično, mogu ukazati na eventualne tendencije alata ka specifičnim grupama proteina. Ovakve analize mogu dodatno poboljšati kvalitet anotacija, i ukazati na smer u kojem bi se alat mogao dalje razvijati.

Funkcionalna anotacija proteina i dalje ostaje jedan od većih izazova današnje nauke. Iako su metode za anotaciju proteina zasnovane na homologiji najzastupljenije i generalno najpristupčanije (Lee et al, 2007), metode zasnovane na drugim principima se takođe razvijaju. Rezultati ovog rada pokazuju da metode kao što je ISM, iako pojedinačno ne pokazuju visoke performanse, imaju dosta potencijala za razvijanje i dopunjavanje dosada razvijenih alata. Kao što su i dosadašnji CAFA izazovi pokazali, kombinacija različitih pristupa objedinjenih metodama mašinskog učenja daju najuspešnije rezultate (Jiang et al, 2016). Određivanje kompletnog sastava mitohondrijalnog proteoma je samo jedan od izazova koji bi se mogao rešiti unapređivanjem računarskih metoda za funkcionalnu anotaciju. Iz tog razloga, povećanje broja istraživanje koja su usko specifična, kao što je ovo, ali i istraživanja velikog obima kao što je CAFA izazov su od presudnog značaja za dalji razvoj ove oblasti medicine i nauke.

ZAKLJUČAK

Nagli porast broja sekvenciranih genoma doveo je do identifikacije velikog broja proteinskih sekvenci kojima se ne zna funkcija, a trenutni intenzitet kojim se oni ekpreimentalno anotiraju nije dovoljan. Ovim se razvija velika potreba za unapređivanjem *in silico* metoda za funkcionalnu anotaciju proteina. Kompjutersko određivanje funcije proteina je jeftin i brz način za dobijanje velikog broja podataka koji se mogu koristiti kao smernice za dalja istraživanja. Veliki izazov u nauci danas takođe predstavlja i određivanje mitohondrijalnog proteoma, što bi znatno unapredilo trenutna saznanja o mitohondrijalnim i drugim bolestima. Kompjuterska funkcionalna anotacija ima veliki udeo u rešavanju ovog problema. U ovom radu poređena je uspešnost predikcije ćelijske lokalizacije u mitohondrijama dva alata koja se zasnivaju na suštinski različitim metodama. Alati koji su poređeni bili su PANNZER, zasnovan na analizi homologih proteina, i FRENKI, zasnovan na metodi analize informacionog spektra (ISM). Referentni set mitohondrijalnih proteina sastavljen je uz pomoć IMPI liste gena koji kodiraju mitohondrijalne proteine. Kao mera za evaluaciju perfomansi predikcionih alata korišćena je F mera, koja zavisi od broja tačnih i netačnih predikcija. Pokazano je da je F mera PANNZER-a viša od F mere FRENKI-ja, za šta je verovatno zaslužno kombinovanje više ulaznih infomracija vezanih za upitnu sekvencu, kao i primena mašinskog učenja pri prioritizovanju tih informacija koja je korišćena pri pravljenju PANZZER algoritma. Slični rezultati dobijeni su i u CAFA2 izazovu, gde je poređeno više desetina različitih alata za funkcionalnu anotaciju. Niža F mera FRENKI-ja može služiti kao smernica za dalje modifikacije ovog alata, a neki od predloga su smanjenje cut-off vrednosti u enričment fazi rada alata, ili zadržavanje na opštijim terminima koji imaju manju preciznost ali daju veću pouzdanost alata. Takođe su ispitana i preklapanja između predikcija oba alata, i interesanto je primetiti da se kako tačne tako i pogrešne predikcije veoma malo preklapaju. Ovo govori o komplementarnosti ova dva alata i njihovoj mogućoj integraciji, uz određene adaptacije i izmene. Dalja istraživanja koja bi se nastavila na ovaj rad mogla bi da obuhvataju detaljnije analize svake od grupa proteina koji se nalaze u skupovima tačnih i pogrešnih predikcija ova dva alata, u cilju davanja smernica za njihovo dalje unapređivanje. Takođe, uključivanje drugih alata iz CAFA izazova bi dalje razjasnilo probleme i potrebe određivanja mitohondrijalnog proteoma. Kompjuterska funkcionalna anotacija je relativno novo polje u biološkim naukama, ali ono pokazuje veliki potencijal. Veći broj uže specijalizovanih istraživanja kao što je ovo, ali i velikih projekata kao što je CAFA, doprineli bi daljem razvoju i unapređenju ovog polja.

LITERATURA

Anderson, Sharon, et al. "Sequence and organization of the human mitochondrial genome." *Nature* 290.5806 (1981): 457.

Ashburner, Michael, et al. "Gene Ontology: tool for the unification of biology." Nature genetics 25.1 (2000): 25.

Calvo, S. E., Clauser, K. R., & Mootha, V. K. (2015). MitoCarta2. 0: an updated inventory of mammalian mitochondrial proteins. *Nucleic acids research*, *44*(D1), D1251-D1257.

Claros, Manuel G., and Pierre Vincens. "Computational method to predict mitochondrially imported proteins and their targeting sequences." *European Journal of Biochemistry* 241.3 (1996): 779-786.

Conesa, Ana, et al. "Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research." *Bioinformatics* 21.18 (2005): 3674-3676.

Davidović R, Gemović B, Veljković N, Perović V. 2018. DiNGO: stand-alone application for GO and HPO term enrichment. Belgrade BioInformatics Conference 2018 (BelBi 2018), Belgrade, Serbia, June 18-22 2018. In: Biologia Serbica, 2018, 40(1): 70.

Falda, Marco, et al. "Argot2: a large scale function prediction tool relying on semantic similarity of weighted Gene Ontology terms." *BMC bioinformatics* 13.4 (2012): S14.

Fitch, Walter M. "Homology: a personal view on some of the problems." *Trends in genetics* 16.5 (2000): 227-231.

Gemović B, Davidović R, Šumonja N, Veljković N, Perović V. 2017. Contribution of features based on sequence, predicted PPIs and GO similarities to the prediction of gene-HPO associations. ISMB/ECCB 2017, Prague, Czech Republic, July 21-25, 2017.

Gong, Q., Ning, W., & Tian, W. (2016). GoFDR: a sequence alignment based method for predicting protein functions. *Methods*, *93*, 3-14.

Jiang, Yuxiang, et al. "An expanded evaluation of protein function prediction methods shows an improvement in accuracy." *Genome biology* 17.1 (2016): 184.

Koskinen, Patrik, et al. "PANNZER: high-throughput functional annotation of uncharacterized proteins in an error-prone environment." *Bioinformatics* 31.10 (2015): 1544-1552.

Lee, David, Oliver Redfern, and Christine Orengo. "Predicting protein function from sequence and structure." *Nature Reviews Molecular Cell Biology* 8.12 (2007): 995.

Loewenstein, Yaniv, et al. "Protein function annotation by homology-based inference." *Genome biology* 10.2 (2009): 207.

Mukherjee, Supratim, et al. "Genomes OnLine Database (GOLD) v. 6: data updates and feature enhancements." *Nucleic acids research* (2016): gkw992.

Pagliarini, David J., et al. "A mitochondrial protein compendium elucidates complex I disease biology." *Cell* 134.1 (2008): 112-123.

Pommier, Yves, et al. "Roles of eukaryotic topoisomerases in transcription, replication and genomic stability." *Nature Reviews Molecular Cell Biology* 17.11 (2016): 703.

Radivojac, Predrag, et al. "A large-scale evaluation of computational protein function prediction." *Nature methods*10.3 (2013): 221.

Radivojac, Predrag, et al. "A large-scale evaluation of computational protein function prediction." *Nature methods*10.3 (2013): 221.

R Development Core Team (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL [http://www.R-project.org.](http://www.r-project.org./)

RStudio Team (2015). RStudio: Integrated Development for R. RStudio, Inc., Boston, MA URL <http://www.rstudio.com/>.

Schnoes, Alexandra M., et al. "Biases in the experimental annotations of protein function and their effect on our understanding of protein function space." *PLoS computational biology* 9.5 (2013): e1003063.

Shannon, Paul, et al. "Cytoscape: a software environment for integrated models of biomolecular interaction networks." *Genome research* 13.11 (2003): 2498-2504.

Somervuo, Panu, and Liisa Holm. "SANSparallel: interactive homology search against Uniprot." *Nucleic acids research*43.W1 (2015): W24-W29.

The UniProt Consortium, UniProt: the universal protein knowledgebase, *Nucleic Acids Res.* 45: D158-D169 (2017)

Törönen, Petri, Alan Medlar, and Liisa Holm. "PANNZER2: a rapid functional annotation web server." *Nucleic acids research*(2018).

Veljkovic, N., Glisic, S., Prljic, J., Perovic, V., Botta, M., & Veljkovic, V. (2008). Discovery of new therapeutic targets by the informational spectrum method. *Current Protein and Peptide Science*, *9*(5), 493-506.

Veljkovic, Nevena, et al. "Discovery of new therapeutic targets by the informational spectrum method." *Current Protein and Peptide Science* 9.5 (2008): 493-506.

You, Ronghui, et al. "GOLabeler: improving sequence-based large-scale protein function prediction by learning to rank." *Bioinformatics* 1 (2018): 9.