Plan upravljanja istraživačkim podacima

|  |
| --- |
| Opće informacije |
|  | Ime i prezime predlagatelja  | Zoran Štefanić |
|  | Matična organizacija | Institut Ruđer Bošković |
|  | Naziv projekta | Alosterički komunikacijski putevi u oligomernim enzimima |
|  | Upravitelj podacima | Zoran Štefanić (zoran.stefanic@irb.hr) |
| 1. | Prikupljanje podataka i dokumentacija |
|  | Koje ćete podatke prikupljati, obrađivati, stvarati ili se ponovno njima koristiti? (navedite format, vrstu i opseg podataka) | Podaci koji će biti stvoreni tijekom projekta spadaju u nekoliko vrsta. * Jedno će biti rezultati računalnih simulacija molekulske dinamike (MD) u vidu trajektorija u kojima su zapisani položaji i brzine svih atoma tijekom simulacije. Glavna dva programa za MD simulacije su Gromacs i Amber i oni generiraju datoteke s ekstenzijama .xtc i .nc. Datoteke za simulacije heksamera PNP-a u trajanju od 1 mikro sekunde (400000 koraka) su veličine oko 500GB! Ukupno je u dosadašnjem dijelu projekta generirano oko 25 TB trajektorija. (<https://alokomp.irb.hr/md/trajectories/>)
* Druga vrsta podataka su podaci koji se prikupljaju na rendgenskom difraktometru na kristalima proteina. Jedno mjerenje sastoji se od nekoliko stotina ili čak tisuća ekspozicija za razne položaje kristala. Takve ekspozicije su datoteke od po nekoliko MB i imaju nastavak .rodhypix. Ukuono jedan set podataka može biti od nekoliko stotina MB pa do nekoliko GB.
* Podaci koji će proizlaziti iz obrade gore navedenih podataka će biti spremljeni u vidu baze podataka koja je u osnovi web stranice projekta: <https://alokomp.irb.hr/>. Baza podataka je PostgreSQL relacijska baza podataka koja sadrži razne vrste podataka: podatke derivirane iz baze podataka Protein Data Bank (PDB) u kojoj su deponirane sve do sada poznate 3D strukture proteina (ovdje su uzete samo strukture PNP-ova, njih 224, https://alokomp.irb.hr/pdbase/); podatke generirane međusobnim preklapanjem svih ovih struktura <https://alokomp.irb.hr/alignments/list/>, njih ukupno 176715; podatke generirane međusobnim sravnjivanjem sekvenci PNP proteina <https://alokomp.irb.hr/msa/1/>; podatke o ovisnostima Ramachandranovih kuteva generiranih iz trajektorija itd…
* Sav razvijeni kod je podvrgnut kontroli verzija programom Git i deponiran na stranicu <https://github.com/zoranstefanic/alokompweb>
 |
|  | Kako će se podaci prikupljati, obrađivati ili stvarati? (ukratko navedite metodologiju i procese osiguranja kvalitete te načine organiziranja podataka) | Trajektorije se izračunavaju programima za MD simulacije Gromacs i Amber i provode se ili na računalu koje je specijalizirano za tu namjenu (Supermicro radna stanica, opremljena ultra brzom GPU jedinicom i procesorom sa 20 jezgri, spremišnim prostorom od oko 70 TB.) ili na superačunalu Isabella na SRCU. Podaci su organizirani u direktorije i svaka trajektorija se mora izvoditi u najmanje 3 replike (oznake R01-R03). Trajanje jedne simulacije je otprilike 10 dana računa!  |
|  | Koju ćete dokumentaciju i metapodatke ustupiti osim podataka? (navedite koje su informacije potrebne korisnicima kako bi mogli čitati i interpretirati podatke u budućnosti te koji će se standardi koristiti pri tumačenju podataka) | Podaci o trajektorijama su smješteni na web stranicu projekta (https://alokomp.irb.hr/md/trajectories) . Tu se prikupljaju metapodaci o vremenima nastanka trajektorije, datoteci, direktoriju u kojem se nalazi, trajanju, broju atoma koji sadrži i sl. Također se za neke trajektorije nalaze poveznice za analize koje su na njima napravljene (analize vremenske ovisnosti Ramachandranovih kuteva, koje su prikazane kao gif slike, grafovi i sl.). Same trajektorije nije moguće niti praktično dohvatiti preko web stranice. |
| 2. | Pravna i sigurnosna pitanja |
|  | Jeste li ograničeni sporazumom o povjerljivosti? Imate li potrebna dopuštenja za prikupljanje, obradu, čuvanje i dijeljenje podataka? Jesu li osobe čiji se podaci pohranjuju informirani o tome i jesu li dali privolu? Kojim ćete se metodama koristiti u svrhu zaštite osjetljivih podataka (GDPR - posebne kategorije osobnih podataka)? | Podaci generirani tijekom trajanja projekta do sada nisu bili zaštićeni nikakvom formom prijave na sustav, npr. lozinkom. Ipak to je planirano u 3. fazi projekta, da bi se zaštitile ideje koje se planiraju publicirati i sama unutarnje struktura baze podataka. To će se izvesti da će svatko tko treba imati pristup (npr. članovi projekta) dobije pristup dijelovima sustava koji su zaštićeni. Nakon objavljivanja rezultata planira se prezentacija rezultata u vidu web aplikacije sa općenito otvorenim pristupom.  |
|  | Kako će se regulirati pristup podacima i njihova sigurnost? Koji su potencijalni rizici koje treba uzeti u obzir? Kako ćete osigurati sigurnost pohrane osjetljivih podataka?  | Pristup podacima biti će reguliran računom na web poslužitelju <https://alokomp.irb.hr/account/login/>. Za sada račun imaju svi sudionici projekta, ali pristup je potpuno otvoren svim stranicama, što će se u najskorije vrijeme regulirati i biti će dostupan samo nekim dijelovima za opću javnost, a ostali dijelovi web stranice biti će zaštićeni. |
|  | Kako ćete upravljati zaštitom autorskih prava i intelektualnog vlasništva? Tko će biti vlasnik podataka? Koje će se licencije primjenjivati na podatke? Koja će se ograničenja primjenjivati na ponovnu uporabu osobnih podataka? | Po završetku projekta i objavljivanja rezultata planira se potpuna dostupnost svih podataka iako to neće biti lako izvedivo jer se planira da će to biti i do 100TB. Svakako će biti važno sačuvati te podatke barem nekoliko godina nakon završetka projekta tako da bi se mogli nezavisno provjeriti rezultati istraživanja ili eventualno neki drugi načini korištenja generiranih podataka. |
| 3. | Pohrana i čuvanje podataka |
|  | Kako će podaci biti pohranjeni i kako će biti napravljena sigurnosna kopija podataka (*backup*) tijekom istraživanja? Koji su kapaciteti čuvanja podataka kojim raspolažete? Kojim se procedurama koristite za sigurnosnu kopiju (*backup*)? | Jedan dio podataka (npr. baza podataka web stranice <https://alokomp.irb.hr/>) se redovito sigurnosno kopira preko mreže na drugo računalo unutar laboratorija (upotrebom automatskog programa cron i rsync). To je nemoguće s ostalim podacima jer naprosto nedostaje prostora. No jedan dio trajektorija se čuva i na SRCU, a ostale postoje duplikati na diskovima (ukupno 6 diskova kapaciteta oko 60 TB).  |
|  | Koji je vaš plan čuvanja podataka? U kojim će se formatima čuvati? | Bilo bi idealno koristiti neku vrstu sažimanja podataka (tar, zip i sl.) međutim trajektorije su već same po sebi komprimirane pa se ne mogu dobro dalje komprimirati. Stoga će biti sačuvane kao originalne datoteke. |
| 4. | Dijeljenje i ponovna uporaba podataka |
|  | Kako i gdje će se podaci dijeliti? Na kojem repozitoriju planirate dijeliti podatke? Kako će potencijalni korisnici doznati za podatke? | Centralno mjesto depozicije svih podataka je zamišljeno na web stranici projekta (<https://alokomp.irb.hr/>), uz gore navedeni sustav regulacije pristupa podacima. Programerski kod koji proizađe iz projekta također će biti dostupan u otvorenom Git repozitoriju na platformi Github. |
|  | Ako postoje podaci koji se ne smiju dijeliti (prijavitelji vezani zakonskim, etičkim, autorskim pravila, povjerljivošću i sl.), pojasnite razloge ograničenja. | Ne postoje takvi podaci. |
|  | Potvrdite da ćete se koristiti digitalnim repozitorijem koji je u skladu s načelima *FAIR-a*. | U skladu sa svime gore navedenim podaci će biti u digitalnom obliku i dostupni u skladu s načelima FAIR-a. |
|  | Potvrdite da ćete se koristiti digitalnim repozitorijem koji održava neprofitna organizacija (ako ne, objasnite zašto ne možete dijeliti podatke na digitalnom repozitoriju koji nije komercijalan). | Stranica <https://alokomp.irb.hr/> je u sklopu matične institucije IRB, te shodno tome u web domena, te je samim time održava neprofitna institucija. Platforma Github je profitna ali omogućava i otvorene repozitorije, i predstavlja de facto standard za dijeljenje koda u programerskoj zajednici, pa tako i otvorenog koda. |

Ref:

[1] Celjak, D., Dorotić Malič, I., Matijević, M., Poljak, Lj., Posavec K. i Turk, I.: „Istraživački podaci - što s njima?“ [Istraživački podaci - što s njima? : priručnik o upravljanju istraživačkim podacima | Digitalni repozitorij Srca (unizg.hr)](https://repozitorij.srce.unizg.hr/islandora/object/srce%3A327)