D 1.2.1 Prikupljeni podaci za reprezentativne genomske i farmakogenomske probleme

Prikupljeni podaci za testiranje i razvoj računalnih procesa i algoritama na projektu AIGEN dostupni su na GitHub repozitoriju projekta: https://github.com/HRZZ-AIGEN

1. Podaci o vezanju molekula na proteine (DTI - drug-target-interaction)

Za modeliranje predviđanja interakcija izmedu lijekova i ciljanih molekula(proteina), postoji značajan broj javno dostupnih baza podataka s informacijama o lijekovima i ciljanim molekulama (proteinima). Baze se u pravilu konstantno ažuriraju novim podacima dobivenim iz novih istraživanja.

Neke od takvih baza podataka su npr.: DrugBank, ChEBI, KEGG LIGAND, KEGG BRITE, BRENDA, SuperTarget.

U našim preliminarnim analizama i testiranju računalnih procesa korišteni su skupovi podatka koji se koriste u više znanstvenih radova o predikciji interakcija lijekova i ciljanih molekula, pa se zato i smatra „zlatnim standardom“ za usporedivanje DTI predikcija:

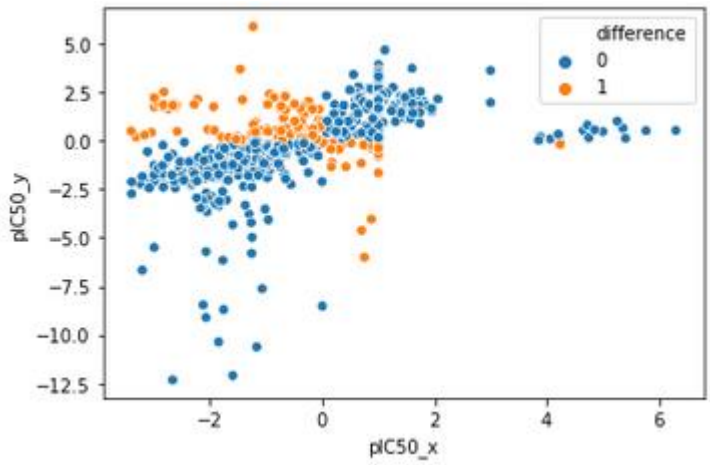
- **Enzyme** - sastoji se od 445 lijekova i 664 ciljanih molekula s 2926 poznatih (eksperimentalno dokazanih) drug-target interakcija
- **Ion channel (IC)** - sastoji se od 210 lijekova i 204 ciljanih molekula s 1476 poznatih (eksperimentalno dokazanih) drug-target interakcija
- **G protein-coupled receptor (GCPR)** - sastoji se od 223 lijekova i 95 ciljanih molekula s 635 poznatih (eksperimentalno dokazanih) drug-target interakcija
- **nuclear receptor (NR) -** sastoji se od 54 lijekova i 26 ciljanih molekula s 90 poznatih (eksperimentalno dokazanih) drug-target interakcija
- **Davis** skup podataka se sastoji od 68 lijekova i 442 ciljane molekule sa izmjerenih cca 30056 interakcija (Kd vrijednosti)
- **KIBA** skup podataka se sastoji od 229 lijekova i 2111 ciljanih molekula sa izmjerenih 118254 interakcija (Kd vrijednosti)

2. Podaci o staničnim linijama tumora i aktivnosti spojeva na staničnim linijama

Farmakogenomske baze podataka koje su korištene za problem predviđanja na staničnim linijama tumora uključuju: NCI-60 [1] farmakogenomska baza podataka koja sadrži otprilike 20 000 lijekova koji su zadovoljili kontrolu kvalitete na 60 staničnih linija; PharmacoDB [2, 3] baza podataka koja sadrži 700 lijekova čiji je odziv izmjeren na više od 1 000 staničnih linija. Biološke značajke staničnih linija su dobivene iz baze podataka DepMap [4].

Metodologija pripreme podataka:
PharmacoDB-1.1.1 verzija (https://zenodo.org/record/1143645) je obrađena tako da su SMILES-i (reprezentacije kemijske strukture) lijekova koji nedostaju, preuzeti sa PubChem baze podataka (https://pubchem.ncbi.nlm.nih.gov/) pomoću CID identifikacijskog broja i ručnim traženjem u relevantnim bazama podataka koje su sadržane u PharmacoDB-u, npr. GDSC-a (https://www.cancerrxgene.org/), pomoću trivijalnih imena lijekova. DepMap 20Q2 verzija (https://doi.org/10.6084/m9.figshare.12280541.v4) je preuzeta sa web portala (https://depmap.org/portal/download/) te je preko trivijalnih imena staničnih linija nađeno podudaranje

između 1 478 staničnih linija u DepMap i PharmacoDB bazama. Svaka pojedina karakterizacija (ekspresija, mutacije, RNAi, miRNA, …), staničnih linija koja se može pronaći u DepMap-u je rađena na različitim podskupovima staničnih linija. Nakon sparivanja staničnih linija za koje postoje sve karakterizacije u DepMap-u sa odzivima na lijekove iz PharmacoDB-a dobiveno je 409 staničnih linija koje sadržavaju sve dostupne podatke. Zatim su preuzeti podaci odziva iz baze podataka NCI-60 (https://discover.nci.nih.gov/cellminer/). Ponovno je korištena baza podataka PubChem kako bi se iz strukturalnih identifikacijskih brojeva dobili CID identifikacijski brojevi te kako bi naposljetku svi lijekovi imali jedinstvene, ujednačene identifikacijske brojeve. Iz dostupnih CID-ova sa PubChema su preuzeti kanonski SMILES-i. Vrijednosti odziva u NCI-60 bazi podataka su pretvorene u mikro molarne koncentracije kako bi bile ujednačene preko ovih različitih baza podataka. Na kraju je sparivanjem između svih navedenih baza podataka dobiveno 553 126 odziva preko 409 staničnih linija koje sadržavaju sve molekularne karakterizacije iz DepMap-a. U prvom izvještaju smo ranije pokazali odstupanje u odzivu između istih parova lijekova i staničnih linija u različitim bazama podataka. Odstupanje nakon kategoriziranja odziva na osjetljive i neosjetljive lijekove je dano u slici 1.



Slika 1. pIC50_x vrijednosti su odzivi iz NCI-60 baze, dok su pIC50_y iste kombinacije lijekova i staničnih linija u PharmacoDB bazi. Granična vrijednost za klasifikaciju u problemu dviju klasa je u ovom slučaju proizvoljno odabrana i predstavlja koncentraciju 1 mikro mola. Podaci sa oznakom 0 su jednako klasificirani podaci u obje baze podataka, dok su oni sa oznakom 1 podaci koji su različito klasificirani.

Reference:

[1] Smirnov, Petr, et al. "PharmacoDB: an integrative database for mining in vitro anticancer drug screening studies." Nucleic Acids Research (2017).

3. Podaci i reprezentacije sekvenci za problem predviđanja genskih/proteinskih funkcija

Podaci vezani za ovaj problem sačinjeni su iz 2 dijela: proteinske sekvence – odnosno njihove informativne reprezentacije s aspekta predviđanja funkcije, te postojeće anotacije proteina bazirane na eksperimentalnim odnosno računalnim anotacijskim protokolima.

Osnova podataka za predviđanje genske funkcije korsteći COG/NOG koncepte jest 5 etabliranih reprezentacija genomskih podataka (ref Ext. Compl), uz novije reprezentacije bazirane na funkcionalnom profilu susjedstva gena (NFP (Gene functional neighbprhoods)),

1. Filetički profili koji predstavljaju prisutnost (0/1) gena koji pripada određenom klasteru gena (tzv COG/NOG familije gena) u određenom genomu. U bazi se nalaze profili gena/COG ova preko 2071 genoma.
2. Biofizičke karakteristike i svojstva proteinskih sekvenci uključuje 1170 svojstava koji reprezentiraju svojstva aminokiselinske sekvence, motiva sekvence i drugih statistika vezanih uz sekvencu određenog proteina. Svojstva su bazirana na ProFET softtware-u (Ofer and Linial, 2015). (https://github.com/ddofer/ProFET)
3. Evolucijski konzervirana genska susjedstva, izražena preko profila udaljenosti gena od najfrekventijih genskih familija (COG) preko 100 genoma. Dimenzija ovog profila je 5891.
4. Empirijska kernel-mapa koja sadržava sličnost sekvence prema genima 6 reprezentativnih genoma
5. Efikasnost translacijskog profila pojedinog proteina (ref ; to je mjera tzv kodonske preferencije i povezanosti sa ekspresijom gena). 04). Ovaj profil definiran je preko efikasnosti translacije proteina u 2071 genoma, uz dodatnih 5891 svojstava koja predstavljaju genske/proteinske ko-ekspresije.
6. Funkcijska susjedstva opisuju okruženje gena (COG/NOG) kao prosječni udio gena s određenom genskom funkcijom u okruženju.

   Dodatni podaci i reprezentacije vezane uz problem određivanje genskih funkcija koji se planiraju ispitati uz postojeće reprezentacije/podatke (1-6) uključuju druge pristupe bazirane na sekvenci gena/proteina kao npr.: One-hot-encoding reprezentacija trigrama aminokiselina, PAAC - pseudo amino acid composition, CT - conjoint triad ili distribuirane reprezentacije sekvenci (prot2vec).

3.1. Anotacija klastera gena baziranih na COG/NOG paradigmi

Koristeći eggNOG bazu mapiraju su geni pojedinih prokariotskih organizama u COG/NOG grupe (Clusters of Orthologous Genes), te se potom postojeće anotacije gena potom vežu uz COG-ove i time određuju anotacije COG/NOG grupa, koristeći određeni „threshold“ na minimalnu frekvenciju pojavljivanja određene anotacije/funkcije uz gene određenog COG/NOG-a

Za anotaciju genskim funkcijama može se koristiti alat eggNOG-mapper kojim se dostupan kao servis i samostalni software-ski alat (https://github.com/eggnogdb/eggnog-mapper)

Anotacija gena:

Dostupne su s UniProtKB/SwissProt baze podataka. UniProt dodatno označuje svaku fukcionalnu anotaciju s jednom od 21 različitih evidencijskih oznaka, kojim se indicira porijeklo određene anotacije. Tipično se u razvoju prediktivnih modela koriste anotacije sa oznakom ljudskog nadgledanja (human curation), ili eksperimentalne oznake koje se smatraju visoko-pouzdanim.

3.3. Protokol pripreme podataka za GFP

1. Proteinske sekvence

Proteinske aminokiselinske sekvence u FASTA-formatu nalaze se u UniProt (http://www.uniprot.org/downloads) bazi podataka, s koje se programski skupljaju, zavisno o problemu koji se rješava.

2. GO (Gene Ontology) anotacije sekvenci

Funkcionalne anotacije proteinskih sekvence dostupne su u SwissProt bazi, GOA bazi (http://www.ebi.ac.uk/GOA) i u GO bazi (hhtp://geneontology.org/page/download-annotations)

Tipično se za učenje modela koristi skup anotacija dobiven ekperimentalnim metodama: 'EXP', 'IDA', 'IPI', 'IMP', 'IGI', 'IEP','TAS' or 'IC'. Uobičajeno je kombiniranje svih eksperimentalno dobivenih anotacija za formiranje trening seta.

3.4. CAFA izazovi kao „benchmark" podaci i temelj za usporedbu računalnih procesa za problem GFP

Dobro pripremljene i anotirane baze za učenje i testiranje modela za GFP nalaze se u sklopu CAFA – Bio Function Prediction portala [2] (CAFA2, CAFA3, CAFA4 podaci o anotiranim genima), koje predstavljaju vrlo dobru podlogu za razvoj metodologije jer daju na raspolaganje dobro definirane proteinske sekvence sa jasno određenim genskim funkcijama u formi training / test skupa podataka, što omogućuje reproducibilne rezultate i mogućnost direktne usporedbe vlastite metodologije sa state-of-the-art rješenjima i pristupima predviđanju genskih funkcija. Pored toga CAFA izazovi jasno definiraju metrike koje služe za usporedbu predviđanja i ocjenu kvalitete prediktvnog modela.

Reference:

[1] V. Vidulin, T. Šmuc, and F. Supek, "Extensive complementarity between gene function prediction methods," *Bioinformatics*, vol. 32, no. 23, pp. 3645–3653, Dec. 2016, doi: 10.1093/bioinformatics/btw532.

[2] "CAFA | Bio Function Prediction." https://biofunctionprediction.org/cafa/.

4. Podaci o fenotipskim karakteristikama prokariotskih organizama

Za testiranje računalnih procesa za predviđanje fenotipova sakupljeni su podaci iz [1], a koristit će se i podaci, anotacije i alati iz [2].

Za testiranja metoda interakcija i povezivanja genotipa i fenotipa korišteni su podaci iz [3] i [4].

Za daljnji razvoj metodologije kao i baze podataka u kojima su sadržani najnoviji podaci i anotacije prokariotskih organizama: (i) PATRIC (Pathosystems Resource Integration Center) – baza podataka i skup alata za analizu bakterijskih infektivnih bolesti; (ii) CARD - (The Comprehensive Antibiotic Resistance Database) baza podataka i skup alata za analizu sekvenci uključujući alate za sravnjivanje sekvenci i identifikaciju gena odgovornih za rezistenciju koji su bazirani na homologiji i SNP modelima.

[1] M. Brbić, M. Piškorec, V. Vidulin, A. Kriško, T. Šmuc, and F. Supek, "The landscape of microbial phenotypic traits and associated genes," *Nucleic Acids Res*, vol. 44, no. 21, pp. 10074–10090, Dec. 2016, doi: 10.1093/nar/gkw964.

[2] J. S. Madin *et al.*, "A synthesis of bacterial and archaeal phenotypic trait data," *Scientific Data*, vol. 7, no. 1, Art. no. 1, Jun. 2020, doi: 10.1038/s41597-020-0497-4.

[3] A. Drouin *et al.*, "Predictive computational phenotyping and biomarker discovery using reference-free genome comparisons," *BMC Genomics*, vol. 17, no. 1, p. 754, Sep. 2016, doi: 10.1186/s12864-016-2889-6.

[4] A. Drouin, G. Letarte, F. Raymond, M. Marchand, J. Corbeil, and F. Laviolette, "Interpretable genotype-to-phenotype classifiers with performance guarantees," *Scientific Reports*, vol. 9, no. 1, Art. no. 1, Mar. 2019, doi: 10.1038/s41598-019-40561-2.