D1.1.2 Usporedba računalnih procesa na odabranim biološkim problemima

Cilj ovog, početnog istraživanja je detaljno upoznati i analizirati metode, algoritme i računalne procese koji predstavljaju stanje tehnike i znanosti vezano uz biološke probleme koji su predmet istraživanja AIGEN projekta.

Računalni procesi i algoritmi, zajedno s podacima skupljenim za razvoj i testiranje dostupni su na GitHub repozitoriju projekta AIGEN: https://github.com/HRZZ-AIGEN

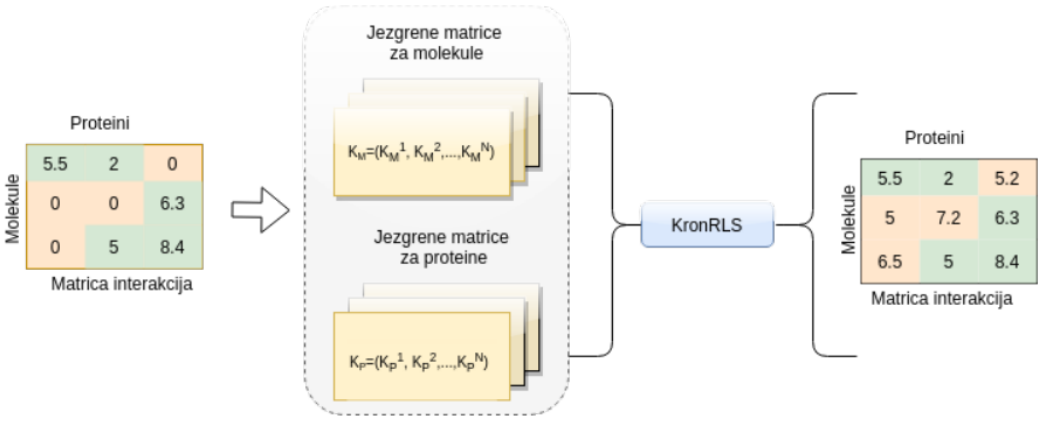
1. Računalni procesi za predviđanje interakcije (afiniteta vezanja) lijekova i proteina

Proučavani su i testirani napredni računalni procesi za modeliranje interakcija lijekova i proteina bazirani na dvjema različitim paradigmama strojnog učenja:

(i) učenju iz više jezgri (Multiple Kernel Leraning) metode, te na
(ii) dubokim, konvolucijskim neuronskim mrežama (CNN)

Ovi su računalni pristupi testirani na nekoliko skupova podataka dostupnih u literaturi, kako bi se ocijenila njihova prediktivna moć.

Učenje iz više jezgri (engl. multiple kernel learning) odnosi se na metode strojnog učenja koje koriste unaprijed definirane jezgrene matrice (engl. kernels) kako bi naučili optimalnu linearnu ili nelinearnu kombinaciju jezgri i postigli optimalnu prediktivnu moć (Slika 1). Jedan od razloga korištenja ovog pristupa je to što omogućava kombinaciju podataka iz različitih izvora, odnosno korištenje jezgrenih matrica sličnosti dobivenih na temelju različitih molekulskih deskriptora za treniranje modela i predvi-đanje afiniteta vezanja.



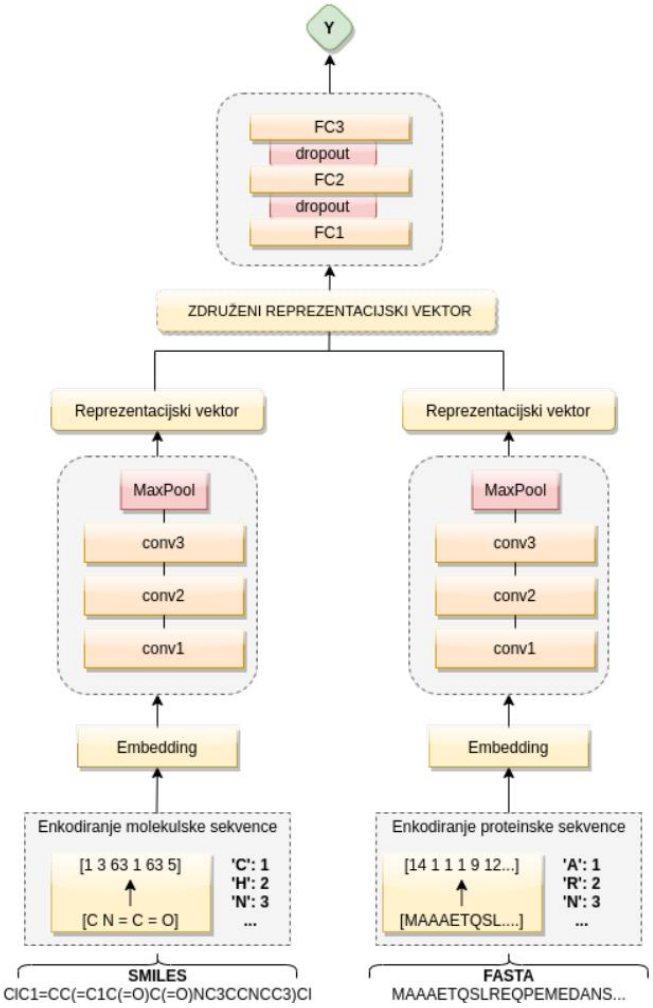
Slika 1. Kron-RLS arhitektura [1]

1.1. Konvolucijske neuronske mreže za predviđanje interakcija

Duboko učenje je podskup strojnog učenja u sklopu umjetne inteligencije, koji se sastoji od mreža sposobnih da provode nenadzirano učenje iz nestrukturiranih i neoznačenih podataka. Jedna od popularnih metoda dubokog učenja su konvolucijske neuronske mreže (eng. Convolutional Neural Networks, CNN's) koje su pretežno najuspješnije kod analize prirodnih slika. Konvolucijski slojevi

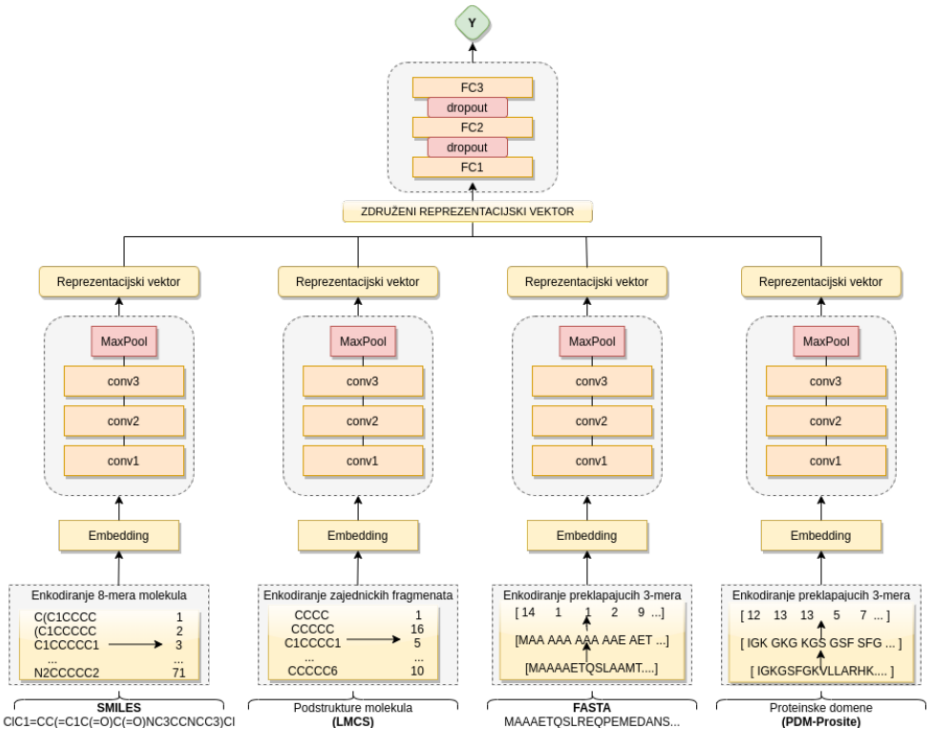
provode preslikavanje ulaznih podataka u prostor manjih dimenzija ovisno o definiranom prozoru i pomaku preko ulazne matrice. Izlazni sloj neuronske mreže predstavlja matricu značajki koje opisuju ulazne podatke. Broj skrivenih slojeva, između ulaznog i izlaznog sloja, je proizvoljan i tipično se adaptira ili optimira naspram problemu koji se rješava.

U našem preliminarnom istraživanju komparirali smo tri arhitekture dubokih neuralnih mreža za predikciju vezanja molekula na proteine: DeepDTA, WideDTA i GraphDTA.



Slika 2. DeepDTA arhitektura

DeepDTA model [2] Predložena prediktivna arhitektura sastoji se od dva zasebna konvolucijska bloka od kojih jedan uči reprezentaciju za male molekule iz SMILES (eng. Simplified Molecular Input Line-Entry System) niza, a drugi iz proteinskih sekvenci u FASTA formatu. Za svaki ulazni blok definiran je embedding sloj dimenzija (N,d), gdje je N veličina vokabulara, a d je duljina vektora čija vrijednost je proizvoljna.



Slika 3. WideDTA arhitektura

Pripremljene enkodirane sekvence za svaki od izvora podataka ulaze u višeslojni konvolucijski blok koji se sastoji od jednog ulaznog, skrivenog i izlaznog sloja. Kako bi se postiglo smanjenje dimenzionalnosti, na samom kraju konvolucijskog bloka implementiran je sloj koji provodi sažimanje maksimalnom vrijednošću (engl. MaxPooling), prilikom kojeg se unutar prozora (engl. kernel) dalje propagira maksimalna vrijednost.

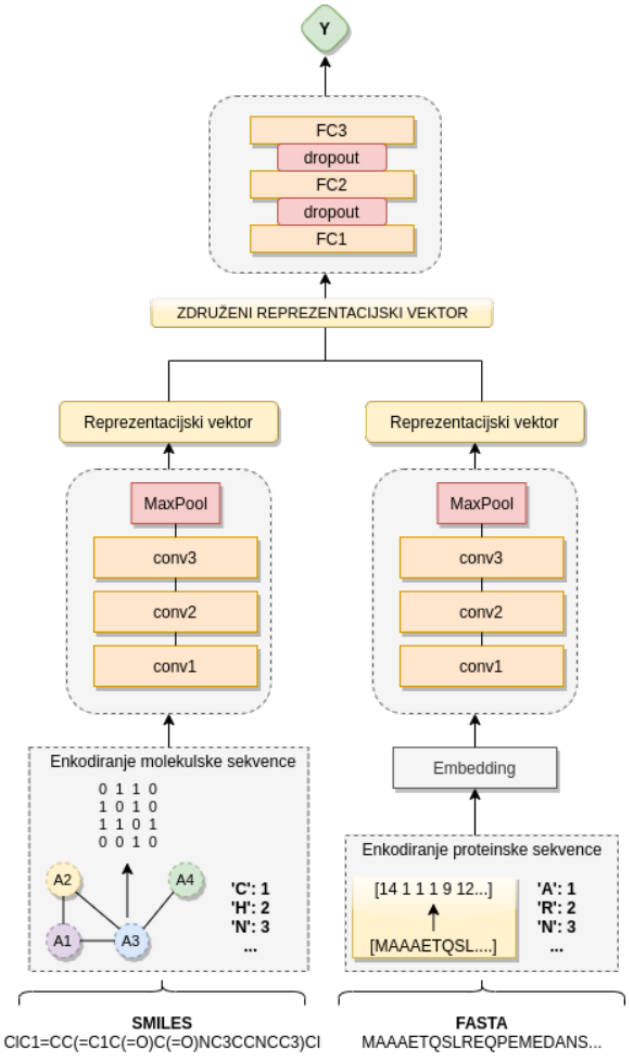
Rezultirajući vektor na izlazu iz konvolucijskog bloka nazivamo reprezentacijskim vektorom. Kod DeepDTA pristupa, na kraju oba CNN bloka nalaze se po jedan vektor značajki koji predstavlja ulazne podatke te se oni združuju u jedan dugački vektor koji ulazi u potpuno povezani sloj neuronske mreže (engl. Fully-connected layer).

Potpuno povezani sloj najčešće je zadnji sloj u neuronskoj mrezi odgovoran za predviđanje klasa ili kontinuiranih vrijednosti na temelju vektora naučenih značajki. Kako bi se izbjeglo pretreniravanje mreže, između svakog potpuno povezanog sloja nalazi se sloj izbacivanja (engl. dropout) s ciljem nasumičnog gašenja pojedinih neurona.

WideDTA [3] predstavlja proširenje DeepDTA arhitekture s dodatnim ulaznim podacima uključujući informacije o domenama i motivima kod ciljanih proteina, te informacije o zajedničkim podstrukturama između ulaznih molekula. WideDTA model zasnovan je na četiri zasebna ulazna CNN bloka, od kojih svaki funkcionira kao zaseban model za učenje reprezentacija iz tekstualnih podataka.

Korištena su četiri izvora podataka koji uključuju: **(i) SMILES** (engl. Simplified Molecular-Input Line-Entry System). Molekulske sekvence prikupljene s PubChem baze podataka su kanonizirane i rastavljene na preklapajuće oktamere. Skup 8-mera jedne molekule predstavljen je numeričkim vektorom, gdje svaki broj odgovara jedinstvenom oktameru iz cjelog skupa podataka, **(ii) LMCS** (engl.

Ligand maximum common substructure). Za razliku od preklapajućih oktamera, molekule su predstavljene numeričkim vektorom gdje brojevi predstavljaju prisutnost fragmenata u molekuli koji su zajednički svim molekula u danom skupu podataka, **(iii) Proteinske sekvence**. Skup proteinskih sekvenci prikupljen je sa UniProt baze podataka. Sekvence su rastavljene na skup „riječi" koje su konstruirane kao skup preklapajućih 3-grama (3-mera).**(iv) PDM** podaci o konzerviranim dijelovima sekvenci proteina, motivima i domenama, bazirano na višestrukom poravnanju proteinskih sekvenci na PROSITE web alatu.



Slika 4 GraphDTA arhitektura

GraphDTA [4] je nadgradnja DeepDTA arhitekture kod koje je zadržan konvolucijski blok za učenje reprezentacije iz proteinskih sekvenci kao i kod DeepDTA modela, gdje mreža sama uči reprezentaciju iz ulazne sekvence, ali je uveden poseba graf-konvolucijski blok kojim se reprezentira informacija o prostornom rasporedu atoma u molekuli, u kojem je svaka molekule predstavljena kao graf gdje su atomi čvorovi, a veze između atoma bridovi grafa. Karakteristike svakog čvora u grafu opisane su sa simbolom odgovarajućeg atoma, brojem vezanih susjednih atoma uključujući broj vodika i ukupan broj vodikikovih atoma te radi li se o aromatskom atomu ili ne. Svaka molekula je opisana s navedenim karakteristikama prikazanim kao višedimenzionalni binarni vektor.

Testovi na dva skupa interakcija malih molekula sa ciljanim proteinima (familija kinaza), prikazani su u Tablicama 1 i 2. Tablica 3 prikazuje osnovne karakteristike skupova podataka *Davis* i *KIBA*.

Tablica 1. Usporedba 4 pristupa modeliranju interakcija na Davis skupu podataka.

Model	Skup	Spojevi	Proteini	MSE	CI
KronRLS	Davis	Pubchem-Sim	SW[1]	0.379	0.871
DeepDTA	Davis	1D	1D	0.261	0.878
WideDTA	Davis	1D+LMCS	1D+PDM	0.262	0.886
GraphDTA	Davis	Graph	1D	0.88	0.254

Tablica 2. Usporedba 4 pristupa modeliranju interakcija na KIBA skupu podataka.

Model	Skup	Spojevi	Proteini	MSE	CI
KronRLS	KIBA	Pubchem-Sim	SW[1]	0.411	0.782
DeepDTA	KIBA	1D	1D	0.194	0.863
WideDTA	KIBA	1D+LMCS	1D+PDM	0.179	0.875
GraphDTA	KIBA	Graph	1D	0.866	0.179

Tablica 3. Skupovi interakcija malih molekula s proteinima, Davis i KIBA, korišteni u testiranju.

Skup	Br.spojeva	Br.proteina	Br.interakcija
Davis (pKd)	68	442	30 056
KIBA	229	2111	118254

Reference:

[1] A. Cichonska, B. Ravikumar, E. Parri, S. Timonen, T. Pahikkala, A. Airola, K. Wennerberg, J. Rousu, and T. Aittokallio, "Computational-experimental approach to drug-target interaction mapping: a case study on kinase inhibitors," PLoS computational biology, vol. 13, no. 8, p. e1005678, 2017.

[2] H. Öztürk, A. Özgür, and E. Ozkirimli, "DeepDTA: deep drug–target binding affinity prediction," Bioinformatics, vol. 34, no. 17, pp. i821–i829, 2018.

[3] H. Öztürk, E. Ozkirimli, and A. Özgür, "WideDTA: prediction of drug-target binding affinity," arXiv preprint arXiv:1902.04166, 2019.

[4] T. Nguyen, H. Le, and S. Venkatesh, "GraphDTA: prediction of drug–target binding affinity using graph convolutional networks," BioRxiv, p. 684662, 2019.

2. Računalni procesi za problem predikcije aktivnosti spojeva na staničnim linijama tumora (Duboki modeli u farmakogenomici)

Duboki modeli koje smo ispitali na problem predikcije aktivnosti spojeva na staničnim linijama su graf neuronske mreže. Grafovi su vrsta podataka koja modelira objekte, u ovom slučaju čvorove, i njihove odnose, tj. veze između čvorova. Graf neuronske mreže, isto kao konvolucijske neuronske mreže, mogu prepoznati uzorke u više-dimenzionalnim lokaliziranim prostornim značajkama i stvoriti njihove generalizacije [1]. S obzirom da je iz molekularne graf teorije poznato da su grafovi dobre reprezentacije strukturalnih formula kemijskih spojeva, nije iznenađujuće da graf neuronske mreže postižu sve bolje rezultate u području predviđanja svojstava kemijskih spojeva [2].

Korišteni modeli:

a. **GraphDRP** (Graph convolutional networks for drug response prediction [3]):
Rad iz literature koji je prvi upotrijebio graf neuronske mreže za predviđanjima na GDSC skupu podataka. U radu su prezentirani modeli sa različitim graf konvolucijskim operatorima. Modeli se sastoje od dva dijela od kojih jedan modelira kemijske spojeve, dok se drugim model uči na reprezentacijama staničnih linija. Dio za modeliranje kemijskih spojeva se sastoji od 4 graf konvolucijska sloja. Graf konvolucijski slojevi koji su prezentirani su u modelima različiti, odnosno GCN (graf konvolucijski operator), GIN (graph isomorphism network), GAT (graph attention network), GCN i GAT kombinirano. Svi modeli za regularizaciju između slojeva koriste Batch normalization. Drugi dio, za modeliranje staničnih linija, je koristi jedan obične jednodimenzionalne konvolucijske filtere na one-hot kodiranim binarnim (je li gen mutiran) mutacijama. Vektori na izlazu iz graf konvolucijskog dijela molekula i generalizacija jednodimenzionalnih konvolucijskih filtera mutacija se spajaju te ulaze u statičku unaprijednu mrežu koja je regularizirana dropout-om.

b. **ChemProp:**
ChemProp [4] je arhitektura korištena u dobrom poznatom radu u području [5] u kojemu su uz pomoću ChemProp arhitekture autori uspjeli otkriti novi antibiotik, koji je u istom radu i eksperimentalno potvrđen. ChemProp koristi standardni postupak "prosljeđivanja poruka (message passing)" koji može generalizirati nekoliko graf neuronskih mreža i graf konvolucijskih operatora [1]. Model se sastoji od dvije faze, prosljeđivanja i očitavanja. Fazu prosljeđivanja poruka, odnosno same poruke, koje mogu biti različito definirano, model koristi za ažuriranje funkcija skrivenih stanja. Faza očitavanja se sastoji od računanja vektora značajki cijelog grafa koristeći funkciju očitavanja. Sa različitim postavkama ovih funkcija (ažuriranja, prosljeđivanja i očitavanja) moguće je napraviti model koji generalizira različite graf neuronske mreže i operatore. Također u model je moguće uključiti i numeričke deskriptore molekula koji poboljšavaju prediktivnu moć modela. U našim eksperimentima ovaj model je dao najbolje rezultate u usporedbi s ostalim literaturnim modelima

c. **DimeNet:**
Dimenet [6] je model razvijen za predviđanje kvantno-mehaničkih svojstava molekula. Model postiže trenutno najbolje rezultate u većini problema u domeni predviđanja kvantno mehaničkih svojstava molekula. Arhitektura modela je vrlo kompleksna. U modelu je integrirano usmjereno prosljeđivanje poruka sa sferičnim Fourier-Bessel reprezentacijama. Sastoji se od jednog bloka za kodiranje, jednog bloka međudjelovanja sa prosljeđivanjem poruka. Svaki blok prosljeđuje naučene embeddinge u izlazni blok gdje su oni transformirani radijalnim skupovima (radial basis) te su zbrojeni preko atoma. Na kraju su izlazi iz svih slojeva zbrojeni kako bi se dobilo krajnje predviđanje.

d. PaccMann:

PaccMann model [7] kao ulazne značajke modela koristi ekspresije gena (16 000 gena) te 512 bitne otiske ili SMILES za molekularne značajke. Iz ulaznih 16 000 gena širenjem kroz mrežu se dobiva manji podskup informativnijih gena tako što se genima koji su poznati kao mete lijekova dodjeljuju veće težine nego ostalima. Dobiveni podskup gena i molekularne reprezentacije zatim ulaze u encoder koji koristi mehanizam "pažnje" (attention). Reprezentacije dobivene iz encodera ulaze u duboku statičku unaprijednu mrežu koja zatim računa vrijednost osjetljivosti danog lijeka. Ova arhitektura zbog attention encodera omogućava identificiranja gena i dijelova struktura ili atoma u molekuli koji su odgovorni za osjetljivost određene stanične linije na lijek.

e. Ostali modeli:

Korišteni su i modeli koji koriste GIN (Graph isomorphism network) operatore praćene sa dubokom statičkom unaprijednom mrežom. Također su korišteni algoritmi XGBoost i Random Forest za stvaranje „baseline" modela koji koriste „standardnu" metodologiju strojnog učenja, uz korištenje numeričkih deskriptora molekula i vektora ekspresije gena kao ulaznih značajki.

Tablica Metode i podaci korišteni u preliminarnim testiranjima

Modeli			Metapodaci		
	Model	Optimizacija	Skup podataka	N(staničnih linija)	N(lijekova)
1	DimeNet + conv block	-	GDSC1000	1074	223
2	Dimenet + conv block	-	GDSC1000	1074	223
3	Dimenet	-	NCI60	1 (MCF7)	11263
4	Chemprop	-	NCI60	1 (MCF7)	12914
5	Random Forest	-	NCI60	1 (MCF7)	12914
6	XGBoost	-	NCI60	1 (MCF7)	12914
7	XGBoost	Bayesian opt	NCI60	1 (MCF7)	12914

Detaljni prikaz svih testiranja metoda i dobivenih rezultata dan je na https://docs.google.com/spreadsheets/d/1T2pOWAYeI7-dorvT1yf7CsjbNgTaBteDvWV1iH1R5is/edit#gid=0

Reference:

[1] J. Zhou, G. Cui, Z. Zhang, et al., Graph Neural Networks: A review of Methods and Applications

[2] K. Liu, X. Sun, et al., Chemi-Net: A molecular graph convolutional network for accurate drug property prediction

[3] Tuan Nguyen, et al., Graph convolutional networks for drug response prediction

[4] K. Yang, et al., Analyzing Learned Molecular Representations for Property Prediction

[5] J.M. Stokes, et al., A Deep Learning Approach to Antibiotic Discovery

[6] J. Klicpera, et al., Directional Message Passing for Molecular Graphs

[7] A. Oskooei, et al., PaccMann: Prediction of anticancer compound sensitivity with multi-modal attention-based neural networks

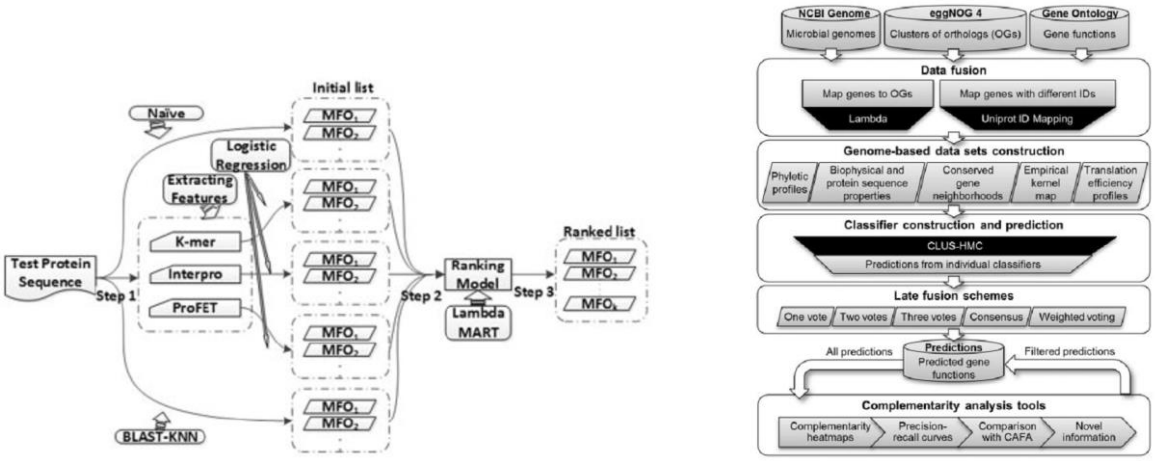
3. Računalni procesi za problem predviđanja funkcije gena

Značajnim povećanjem dostupnih podataka o genomima i sekvenciranjem velikog broja genoma metode strojnog učenja počele su se intenzivno koristiti za predviđanja funkcija novo-otkrivenih gena, a u posljednje vrijeme pojavile su se i metode bazirane na dubokom učenju za predviđanje genskih funkcija.

Tokom prve godine istraženi su state-of-the-art modeli strojnog učenja i metoda predviđanja funkcija gena (engl. GFP - gene function prediction). S obzirom da su geni/proteini definirani primarno sekvencom aminokiselina pristupi za učenje iz sekvenci poput konvolucijskih neuronske mreže su logični osnovni pristup dubokog učenja.

U proučavanju i razvoju metoda za strojno učenje u primjeni na konkretni problem važno je postaviti osnovna ili najjednostavnija rješenja problema kako bi imali referentno rješenje prema kome se sva rješenja i poboljšanja mogu uspoređivati. Kod GFP najosnovniji pristup je naivno anotiranje svih neanotiranih gena funkcijama prema frekvenciji pojavljivanja određenih funkcija u samom trening skupu. Drugi, točniji pristup je preko sličnosti sekvenci [1] gdje se novoj sekvenci gena pripisuju genske funkcije prema funkcijama najsličnijeg(ih) anotiranih gena.

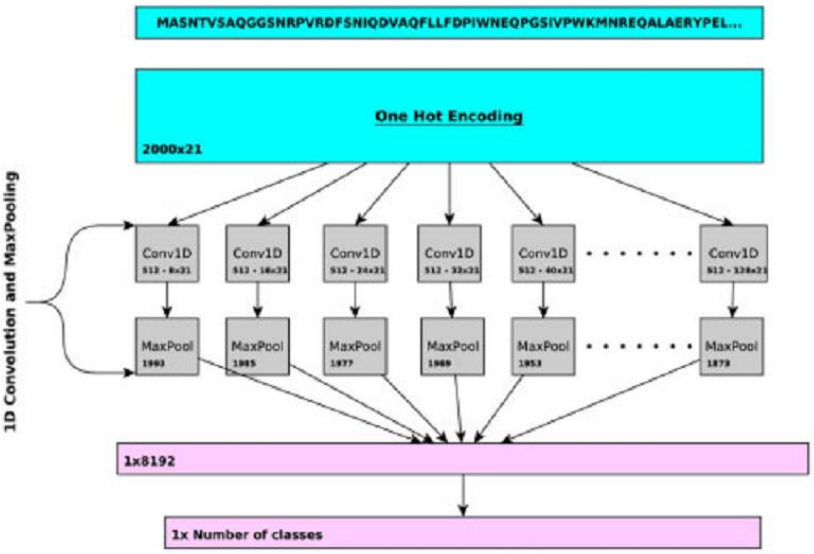
Konkurentniji odnosno točniji pristupi GFP kao GOLabeler [2] kao i vlastiti pristup [3], tipično koriste različite reprezentacije i informacije o proteinskoj sekvenci (motivi i anotacije vezane uz motive u proteinskoj sekvenci, različita fizikalno-kemijska svojstva, sličnosti s drugim proteinima, lokaciju i pojavnost gena), koje se koriste za ravoj individualnih modela čije se predikcije potom kombiniraju po principu utežnjenog usrednjavanja [3], ili rangiranjem [2] kako bi ukupni, ansambl model, točnije predviđao genske funkcije.



Slika 1 GOLabeler [2] proces fuzije različitih reprezentacija i svojstava sekvence (lijevo) i vlastiti pristup [3] baziran na kombinaciji različitih svojstava sekvence gena i njegove pojavnosti (desno).

U analizi metoda dubokog učenja u primjeni na GFP analizirali smo i testirali dva novija pristupa DeepGOPlus [4] i DEEPRed[5]. DeepGOPlus metoda bazirana je na kombinaciji dva pristupa/predikcije: jedan je baziran na učenju iz sekvence konvolucijskom dubokom mrežom dok je drugi – klasični pristup preko sličnosti sekvenci proteina. S druge strane DeepRED pristup se bazira na

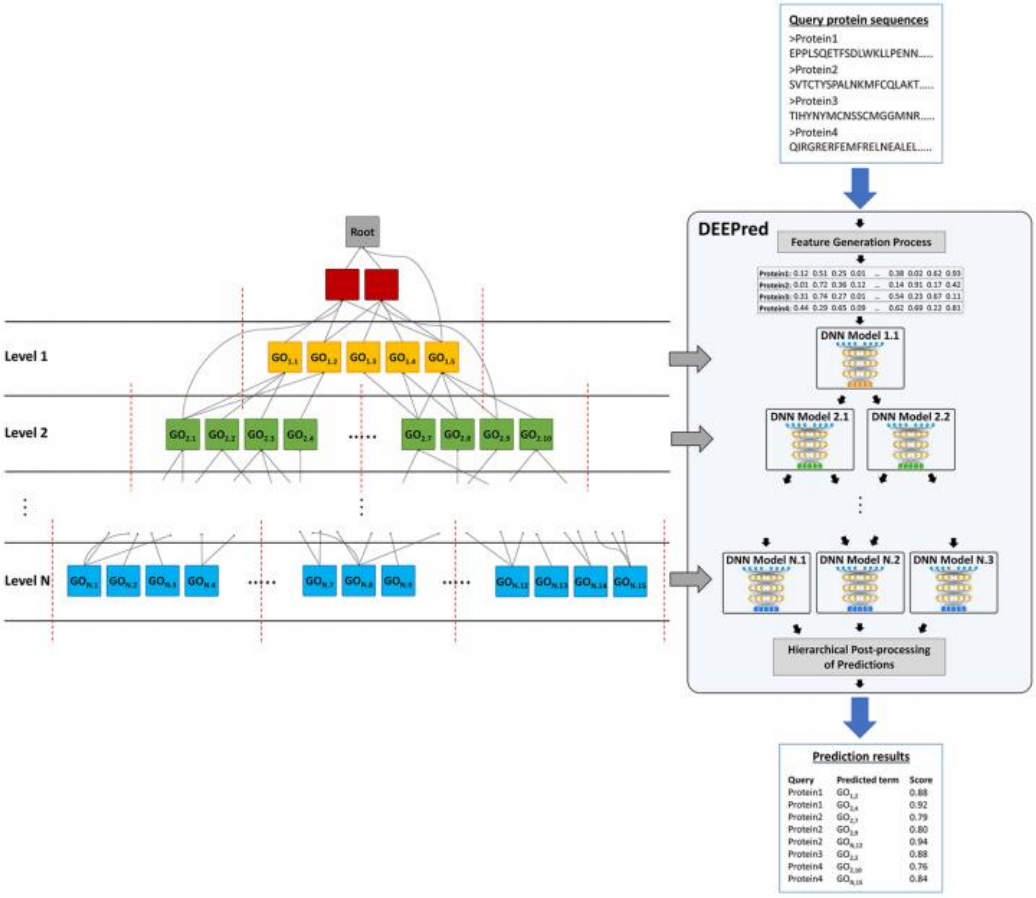
tri različita pristupa opisu sekvence proteina, preko profila podsekvenci proteina, tzv. Conjoint triada(trimera) i pseudo-aminokiselinskog sastava



Slika 2 DeepGOPlus [4] arhitektura za treniranje GFP modela

Ulazni podaci su stoga prikazani preko tzv. trigrama (tri uzastopne aminokiseline u sekvenci - ukupno 20^3 trigrama što je 8000 parametara). DeepGOPlus koristi tzv. one-hot-encoding reprezentaciju (8000 * 21) na ulazu iza kojeg se nalazi nekoliko konvolucijskih slojeva s različitim duljinama filtera. U DeepGOPlus klasifikacija se vrši u završnom „flat" klasifikacijskom sloju koji tipično ima i do nekoliko 1000 klasa, a koji broj zavisi o dostupnosti anotacija po GO kategoriji.

DeepGOPlus je efikasna arhitektura kojom su na CAFA3 challenge-u postignuti vrlo dobri rezultati. Druga prednost je da za korištenje kao ulazni podatak treba samo sekvencu proteina, kao i da nije ograničena duljinom sekvence. S obzirom na efikasnost, može se koristiti za anotacije novo sekvencioniranih genoma.



Slika 3. DEEPRed arhitektura [5] za treniranje GFP modela, na pojednostavljenom (hipotetičkom) GO grafu.

Na ilustraciji DEEPRed procesa (Slika 2), DNN model 1.1 uključuje GO pojmove: $GO_{1,1}$ do $GO_{1,5}$ na najopćenitijoj, prvoj razini GO. S obzirom da su donje razine popunjene s velikim brojem kategorija, radi traktabilnosti učenja, potrebno je više DNN da bi se modeliralo sve GO kategorije na određenom sloju GO crvene isprekidane crte predstavljaju način grupiranja GO pojmova koji se onda modeliraju zajedno u okviru jedne DNN); U primjeru na slici DNN modeli iste razine *N* označuju se sa N.1, N.2 i N.3 uključuju GO pojmove od 1-5; 6-10; 11-15.

U koraku predviđanja/klasifikacije, kada se na DEEPred model ulazna sekvenca se transformira u skup Značajki i izvršava se kompletna hijerarhija multi-target DNN-a. Nakon toga se procjenjuju predviđanja GO termina iz svakog modela zajedno u posebnom hijerarhijskom postupku naknadne obrade kako bi se predstavio konačni popis predviđanja GO kategorija koje prelaze određenu granicu pouzdanosti.

Očiti nedostatak ove arhitekture je velika složenost i zahtjevi na računalne resurse prilikom učenja modela, ali i u primjeni modela.

Tablica 1. Usporedba jednostavnih (baseline) pristupa (Naive, Blast) i DeepGO i DEEPRed arhitekture na [5].

		Precision		Recall (at F-max)			Smin
	All	No-knowledge	All	No-knowledge	All	No-knowledge	All
Molekularna funkcija							
Naive	0.29	0.49	0.41	0.27	0.23	6.87	6.43
Blast	0.39	0.42	0.36	0.38	0.44	6.99	6.48
DeepGO	0.34	0.58	0.48	0.3	0.27	6.36	6.01
DEEPRed	0.5	1	1	0.32	0.33	5.41	5.03
Stanični dio							
Naive	0.54	0.56	0.58	0.55	0.5	7.61	7.65
Blast	0.45	0.39	0.39	0.56	0.53	9.74	9.94
DeepGO	0.53	0.61	0.58	0.48	0.49	7.68	7.55
DEEPRed	0.35	1	1	0.2	0.22	9.85	9.53
Biološki proces							
Naive	0.3	0.25	0.39	0.26	0.24	24.27	20.85
Blast	0.32	0.22	0.27	0.37	0.38	25.11	21.35
DeepGO	0.34	0.4	0.52	0.21	0.26	23.41	20.19
DEEPred	0.33	1	1	0.19	0.19	22.04	19.69

Plan daljnjih istraživanja u kontekstu problema predviđanja genskih funkcija

Konstruiranje novih-informativnijih reprezentacija: Same sekvence aminokiselina kao značajke modela se mogu reprezentirati na mnogo načina. One-hot reprezentacija kodira svaku podsekvencu fiksne duljine kao vektor, čija je veličina broj jedinstvenih podsekvenci. Distribuirane reprezentacije se temelje na matrici "pojave" koja pokazuje frekvenciju riječi (u ovom slučaju podsekvenci) koje se javljaju u kontekstu druge riječi. Dimenzionalnost matrice pojave se može smanjiti uobičajenim metodama za smanjenje dimenzija kako bi se dobilo na računalnoj učinkovitosti. Još jedna od reprezentacija koja se može koristi je BERT (Bidirectional Encoder Representations from Transformers), koja predstavlja tzv. Distribuiirane reprezentacije (eng. embeddings). U ovom kontekstu, testirati ćemo i složenije n-grame (tipično se u primjenama koriste tri-grami).

Transformer arhitekture - plan nam je istražiti učinkovitost transformer arhitekture u modelima za predviđanje GO oznaka. Transformer arhitekture su inače najpopularnija arhitektura za prevođenje jezika te za obradu prirodnog jezika općenito. Obzirom na hipotezu da veza između motiva u proteinskoj sekvenci koji su međusobno udaljeni može biti informativna za predviđanje određene GO oznake, mehanizam pažnje u transformerima bi mogao otkriti motive koji su udaljeni a doprinose određenim klasama GO funkcija. Nedostatak ovog pristupa je to što su sekvence aminokiselina velikih dimenzija (u kontekstu ulaznih značajki za transformer model te u odnosu na riječi ili rečenice) pa bi ovakav pristup mogao biti izuzetno računalno zahtjevan. Iako se u zadnje vrijeme ulažu veliki napori kako bi se transformeri učinili računalno učinkovitijima, bit će potrebno osmisliti odgovarajuću (jednostavniju) ulaznu reprezentaciju aminokiselina kako bi se transformeri mogli evaluirati na ovom problemu.

Reference

[1] Buchfink,B.et al. (2015) Fast and sensitive protein alignment using Diamond. Nat. Methods,12, 59.

[2] You,R.et al. (2018b) GOLabeler: improving sequence-based large-scale protein function prediction by learning to rank. Bioinformatics, 34, 2465–2473

[3] 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.

[4] M. Kulmanov and R. Hoehndorf, "DeepGOPlus: improved protein function prediction from sequence," *Bioinformatics*, vol. 36, no. 2, pp. 422–429, 15 2020, doi: 10.1093/bioinformatics/btz595.

[5] A. Sureyya Rifaioglu, T. Doğan, M. Jesus Martin, R. Cetin-Atalay, and V. Atalay, "DEEPred: Automated Protein Function Prediction with Multi-task Feed-forward Deep Neural Networks," *Scientific Reports*, vol. 9, no. 1, Art. no. 1, May 2019, doi: 10.1038/s41598-019-43708-3.

[6] M. Kulmanov, M. A. Khan, and R. Hoehndorf, "DeepGO: predicting protein functions from sequence and interactions using a deep ontology-aware classifier," *Bioinformatics*, vol. 34, no. 4, pp. 660–668, Feb. 2018, doi: 10.1093/bioinformatics/btx624.

4. Računalni procesi za problem predviđanja fenotipa i modeliranje povezanosti genotipa i fenotipa

U osnovi, problem predviđanja fenotipa ima velike sličnosti s problemom određivanja funkcija gena sa pozicije strojnog učenja. U principu se radi o multi-label klasifikacijskom problemu, stoga će se prediktivni procesi za problem predviđanja genskih funkcija prilagođavati različitim varijantama problema predviđanja fenotipa. Pored problema predviđanja ukupnih fenotipskih oznaka, što je bio predmet našeg prethodnog istraživanja [1], fokusirati ćemo se i na fenotipske oznake koje su izuzetno bitne s medicinskog stanovišta (antibiotička rezistencija), odnosno problem pronalaženja veze genotipa i fenotipa.

Druga sličnost između problema predviđanja fenotipa i određivanja funkcija gena je što je u pozadini fenotipa/funkcije genska sekvenca - u prvom slučaju cijelog genoma, a u drugom pojedinačnog gena. No, s obzirom da se kod problema predviđanja fenotipa radi o svojstvu organizma reprezentacija podataka odnosno svojstva kojima opisujemo genom su ponešto različita. Stoga će se metode reprezentacije genoma više podudarati sa reprezetacijama korištenim za stanične linije tumora, prije svega zato što se i kod ovog potonjeg slučaja radi o tome da je potrebno opisati genom.

Računalne metode za otkrivanje povezanosti genotipa i fenotipa (kao i kod problema aktivnosti spojeva na staničnim linijama) vezane su stoga uz efikasnu identifikaciju genomskih biomarkera. To u principu uključuje bilo koje genomske varijacije od supstitucije jednostrukim nukleotidima, tzv. indelima, do većih genomskih preuređenja/rearanžmana. S porastom broja i uz smanjenje troškova sekvenciranja DNA, sada je moguće tražiti takve biomarkere u cijelim genomima i to na velikim skupovima genoma/jedinki. To je i motiv za razvoj računalnih alata i procesa koji se mogu nositi s velikim količinama genomskih podataka i prepoznati suptilne varijacije koje predstavljaju biomarkere fenotipa.

Tipično se genomi uspoređuju na temelju skupa jednonukleotidnih polimorfizama (SNP). Identifikacija SNP-a oslanja se na višestruko sravnjivanje, što je računski skupo i može proizvesti netočne rezultate u prisutnosti velikih genomskih promjena, poput umetanja gena, brisanja, dupliciranja, inverzije ili translokacije.

Novije metode za usporedbu genoma ne oslanjaju se na višestruko sravnavanje, poput metode koja se oslanja na usporedbe temelju k-mera, tj. Kraće sekvence od k-nukleotida. Glavna je prednost ovog pristupa u tome što je robusna naspram genomskih preuređenja i pruža nepristran način uspoređivanja genomskih sekvenci i identificiranja varijacija koje su povezane s fenotipom. Problem je što je ovakav genomski prikaz daleko manje kompaktan od skupa SNP-a i samim tim postavlja znatno veći izazov u smislu potrebnih računalskih resursa.

U ovakvom pristupu cilj je pronaći najsažetiji skup genomskih značajki koji omogućuje točnost predviđanja fenotipa. Uključivanje neinformativnih ili suvišnih značajki dovode do dodatnih troškove

provjere valjanosti rezultata. Stoga se u ovakvom pristupu najčešće koriste računalni procesi zasnovani na strojnom učenju, kako bi se došlo do robustnih modela fenotipa koji su točni i kompaktni(rijetki), tj. Koji su bazirani na minimalno potrebnom skupu genomskih značajki.

Najjednostavnije metode su metode odabira varijabli bazirane na univarijantnim statističkom testiranju, tzv. filter metode. No, te metode ne mogu otkriti interakcije između različitih značajki, a sam odabir je nezvisan o modeliranju što može dovodi do suboptimalnog odabira značajki.

Ugrađene (embedded) metode integriraju odabir značajki u algoritam za učenje modela, tako da odabiru značajke na temelju procjene da se njihovim odabirom generira točniji prediktivni model fenotipa. Neke od metoda strojnog učenja poput varijanti Random Forest algoritma[1,2], ili Redescription Mining algoritma [3] odnosno sličnih metoda[4] mogu otkriti multivarijantne interakcije između značajki. Ove smo metode implementirali i testirali u ovom periodu, i detaljnije će se istraživati u ovom kontekstu, kao i njihova nadgradnja ili povezivanje s metodama dubokog učenja.

Reference

[1] R. D. Shah and N. Meinshausen, "Random Intersection Trees," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 629–654, Jan. 2014.

[2] S. Basu, K. Kumbier, J. B. Brown, and B. Yu, "Iterative random forests to discover predictive and stable high-order interactions," *Proceedings of the National Academy of Sciences*, vol. 115, no. 8, pp. 1943–1948, Feb. 2018, doi: 10.1073/pnas.1711236115.

[3] M. Mihelčić, S. Džeroski, N. Lavrač, and T. Šmuc, "A framework for redescription set construction," *Expert Systems with Applications*, vol. 68, pp. 196–215, Feb. 2017, doi: 10.1016/j.eswa.2016.10.012.

[4] Marchand M, Shawe-Taylor J. The set covering machine. J Mach Learn Res. 2002;3:723–46