Univerzitet u Beogradu
Matematički fakultet

Master Rad

Predviđanje funkcija proteina metodama binarne klasifikacije

Autor:
Anja Bukurov

Mentor:
dr Jovana Kovačević

Članovi komsije:

dr Jovana Kovačević
prof. dr Gordana Pavlović-Lažetić
dr Mladen Nikolić

Beograd, 2019

Sadržaj

Glava 1

Uvod

Biće dodato kasnije...

Glava 2

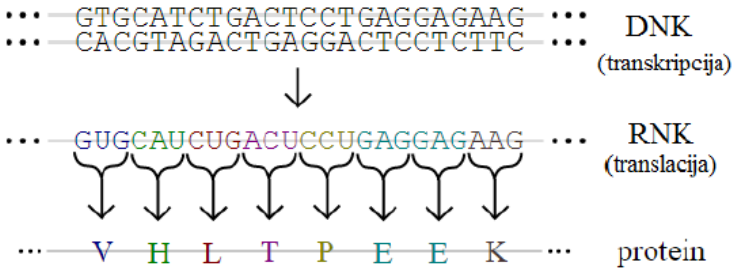
Proteini

Sva živa bića sastoje se iz ćelija. U ćelijama se neprestano odvijaju različiti procesi u kojima učestvuju nukleinske kiseline (dezoksiribonukleinka kiselina - DNK i ribonukleinska kiselina - RNK) i proteini. Unutar molekula DNK šifrovan je genetski materijal koji sadrži uputstva za sintezu proteina.

Proteini su makromolekuli koji igraju mnoge kritične uloge u organizmu. Sačinjavaju više od 50% suvog dela ćelije i važni su za njenu izgradnju i funkcionisanje. Kontrakcija mišića, strukturna podrška, ubrzavanje i usporavanje hemijskih reakcija, odbrana od virusa i bakterija samo su neke od mnogobrojnih uloga koje proteini obavljaju [1, 2].

2.1 Sinteza proteina

DNK sadrži informacije koje su neophodne ćeliji za izgradnju veoma važnog tipa molekula - proteina. Proteini se sintetišu prilikom genske ekspresije i to u dva koraka: transkripcija i translacija (slika 2.1). Prvi korak je dekodiranje genske poruke, prilikom čega se od DNK sekvence dobija RNK sekvenca. U sastav obe nukleinske kiseline ulazi 4 nukleotida i oni su prikazane u tabeli 2.1. S obzirom da su tri nukleotidne baze iste, proces transkripcije sastoji se iz zamene svakog molekula T molekulom U.



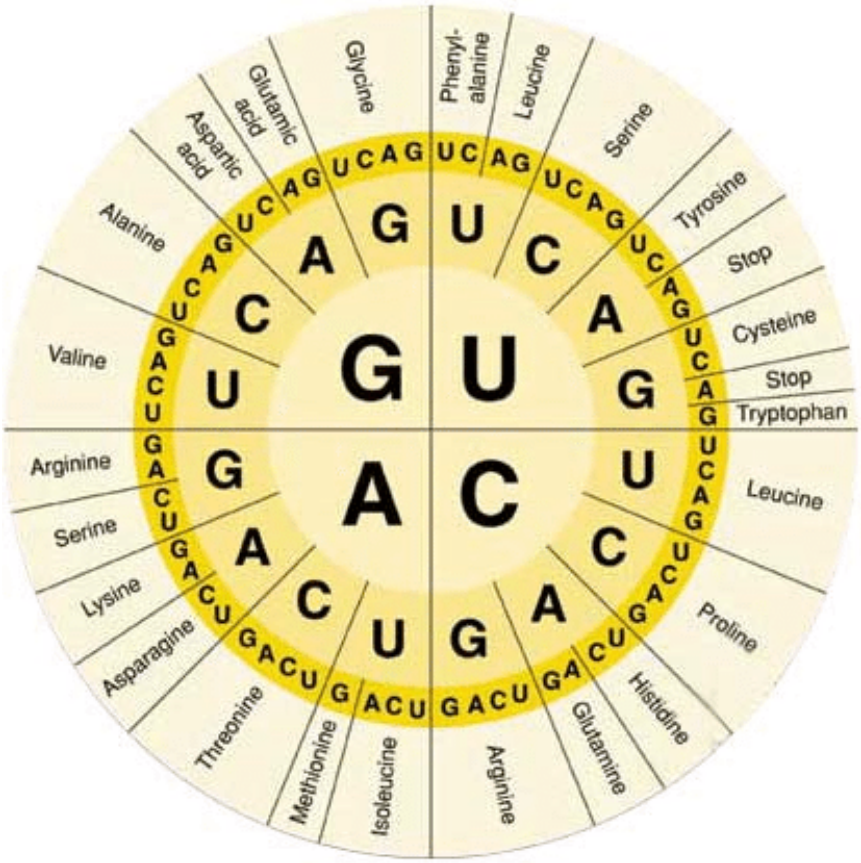
Slika 2.1: Prikaz procesa sinteze proteina [2].

DNK	adeinin (A)	guanin (G)	citozin (C)	timin (T)
RNK	adeinin (A)	guanin (G)	citozin (C)	uracil (U)

Tabela 2.1: Prikaz nukleinskih kiselina sa nukeotidima koji ih grade.

Sledeći korak, proces translacije, jeste grupisanje aminokiselina kako bi se dobio protein. Genetski kod se čita u grupama od 3 nukleotida koje nazivamo *kodoni*. Svaki

kodon odgovara tačno jednoj aminokiselini ili služi da označi kraj sekvence (stop kodon). Na primer, kodon **GUA** kodira aminokiselinu valin, dok kodon **UAG** označava kraj sekvence. Na slici 2.2 je dat šematski prikaz svih kodona i odgovarajućih aminokiselina. Jedan po jedan, kodoni se prevode u odgovarajuće aminokiseline čime se dobija sekvenca aminokiselina koja čini protein [3, 4].



Slika 2.2: Prikaz kodona i odgovarajućih aminokiselina.

2.2 Aminokiseline

Aminokiseline su organska jedinjenja koja se sastoje od karboksilne grupe ($COOH$), aminogrupe (NH_2) i bočnog lanca (R-grupa) koji je vezan za α-ugljenikov atom i karakterističan je za svaku aminokiselinu. Postoji 20 standardnih aminokiselina i one su prikazane u tabeli 2.2. Pod standardnim aminokiselinama podrazumevaju se one aminokiseline za koje postoji najmanje jedan specifičan kodon u genetskom kodu [5, 6, 7].

Aminokiseline možemo podeliti u nekoliko grupa prema osobinama bočnog lanca [5, 6, 8]:

1. **aminokiseline sa nepolarnim bočnim lancem**
   Bočni lanac ovih aminokiselina ne može da otpušta niti da vezuje protone, kao ni da učestvuje u vodoničnim ili jonskim vezama. Zbog svoje nepolarnosti, one su hidrofobne i obično popunjavaju praznine u unutrašnjosti proteina čime doprinose oblikovanju njegove strukture. U ovu grupu ubrajamo 7 standardnih aminokiselina: alanin, valin, leucin, izoleucin, metionin, fenilalanin, triptofan.

Aminokiselina	Oznaka	Simbol	Aminokiselina	Oznaka	Simbol
Alanin	ALA	A	Arginin	ARG	R
Asparagin	ASN	N	Asparaginska kiselina	ASP	D
Cistein	CYS	C	Glutamin	GLN	Q
Glutaminska kiselina	GLU	E	Glicin	GLY	G
Histidin	HIS	H	Izoleucin	ILE	I
Leucin	LEU	L	Lisin	LYS	K
Metionin	MET	M	Fenilalanin	PHE	F
Prolin	PRO	P	Serin	SER	S
Treonin	THR	T	Triptofan	TRP	W
Tirosin	TYR	Y	Valin	VAL	V

Tabela 2.2: Prikaz standardnih aminokiselina sa oznakama i simbolima

2. **aminokiseline sa nenaelektrisanim polarnim bočnim lancem**
   R-grupa aminokiselina iz ove grupe može da gradi vodonične veze sa molekulima vode što ih čini rastvorljivijim u odnosu na aminokiseline iz prethodne grupe. Zbog polarnosti, ove aminokiseline se obično nalaze na spoljašosti proteina. Ova grupa obuhvata 6 standardnih aminokiselina i to: serin, treonin, tirozin, asparagin, glutamin i cistein.

3. **aminokiseline sa naelektrisanim polarnim bočnim lancem**
   U ovu grupu spadaju veoma hidrofilne aminokiseline zbog čega se one nalaze na površini proteina. Dodatno ih možemo podeliti na kisele i bazne aminokiseline. Kisele imaju jednu karboksilnu grupu više i imaju negativno naelektrisanje, dok su bazne aminokiseline pozitivno naelektrisane. Asapraginska i glutaminska kiselina su kisele aminokiseline, a lizin, histidini i arginin spadaju u bazne aminokiseline.

4. **konformaciono važne aminkiseline**
   Preostale dve standardne aminokiseline, glicin i prolin se po svojoj strukuturi razlikuju od ostalih. Glicin nema bočni lanac i može da se prilagođava konformacijama koje su nedostupne drugim aminokiselinama. Prolin sadrži jedan heterociklički prsten i u svojoj strukuturi sadrži sekundarnu amino grupu.

Bilo koje dve aminokiseline mogu izgraditi veći molekul, dipeptid, formiranjem peptidne veze između njih. Peptidna veza se ostvaruje između atoma ugljenika iz karboksilne grupe i atoma azota iz amino grupe. Peptidne veze omogućavaju stvaranje lanaca aminokiselina, tzv. polipeptida. Peptidna veza nastaje reakcijom dve aminokiseline pri čemu se spajaju karboksilna grupa jedne sa amino grupom druge aminokiseline uz izdvajanje vode. Prilikom tog vezivanja pojavljuje se niz koji se zove kičma polipeptidnog lanca koji čine ugljenikov atom karboksilne grupe, atom azota aminogrupe i α-ugljenikov atom. To je osnovni niz i isti je za sve proteine, a oni se međusobno razlikuju po bočnim lancima aminokiselina [6, 8].

Peptide možemo podeliti prema broju aminokiselina koje sadrže i to na oligopeptide i polipeptide. Oligopeptidi su sačinjeni od najviše 10 aminokiselina, dok polipeptidi sadrže do 100 aminokiselina. Jedinjenja sa više od 100 aminokiselina u lancu spadaju u proteine [6].

2.3 Struktura proteina

U sastav proteina ulazi 20 standardnih aminokiselina. Sekvenca aminokiselina, koja se formira peptidnim vezama, specifična je za svaki protein. Ona je primarni izvor informacija o proteinu i njegovoj funkciji. Složenost proteinske strukture najbolje se analizira kroz četiri nivoa: primarna, sekundarna, tercijerna i kvaterna struktura [5].

Primarna strktura Jedinstveni redosled aminokiselina koje su povezane peptidnom vezom kako bi formirale protein čini primarnu strukturu proteina. Proteini koji imaju slične sekvence često imaju i slične osobine i funkcije. Zbog toga je poređenje sekvenci prvi korak u izučavanju proteina. Razumevanje primarne sekvence je bitno zbog mnogih genetskih bolesti koje za posledicu imaju proteine sa neispravnim sekvencama što vodi do pogrešnog savijanja i nefunkcionalnog proteina [5, 6, 8].

Sekundarna struktura Polipeptidni lanac ne zauzima bilo kakav oblik u prostoru već ima opšti raspored aminokiselina koje se u lancu nalaze jedna blizu druge. Taj raspored označava sekundarnu strukturu proteina i podrazumeva savijanje ili uvijanje polipeptidnog lanca. Lanac može da uzme oblik α-heliksa (engl. α-helix), β-traka (engl. β-sheet) ili β-okreta (engl. β-turn). α-heliks je periodična struktura u kojoj se kičma proteina spiralno uvrće, a bočni lanci aminokiselina izviruju izvan nje. β-traka formiraju se kao parovi lanaca aminokiselina koji se uzdužno vezuju vodoničnim vezama. β-okret menja pravac polipeptidnog lanca čime mu pomaže dobije kompaktan, loptast oblik [5, 8, 9].

Tercijarna struktura Prostorna struktura čitavog molekula proteina predstavlja ternarnu strukturu. Hidrofobni bočni lanci nepolarnih aminokiselina teže da budu unutar molekula proteina zaštićeni od vode, dok se kisele i bazne aminokiseline obično nalaze na površini proteina pošto su hidrofilne. α-heliksi i β-listovi služe da obezbede maksimalan broj vodoničnih veza u unutrašnjosti molekula, čime sprečavaju da se molekuli vode vežu za hidrofilne grupe i time naruše integritet proteina [5, 6].

Kvaternarna struktura Mnogi proteini su formirani grupisanjem više savijenih polipeptidnih lanaca. Pojedinačnu komponentu nazivamo podjedinica. One mogu biti međusobno različite ili potpuno iste. Raspored ovih podjednica predstavlja kvaternarnu strukturu. U kvaternarnu strukturi podjedinice se međusobno drže zajedno nekovalentnim interakcijama i kovalentnim vezama [2, 5, 9].

2.4 Uloga proteina

Proteini su najbrojniji i funkcionalno najrazličitiji molekuli u živom svetu. Svaki od njh ima veoma važnu ulogu u organizmu. Na primer:

- Enzimi su proteini koji olakšavaju hemijske reakcije. Učestvuju u skoro svim reakcijama u ćelijama i pomažu u izgradnji novih molekula.
- Antitela su proteini koje proizvodi imuni sistem da bi pomogli u odstranjivanju stranih supstanci i kako bi se borile protiv infekcija. Oni se vezuju za nepoznate čestice, poput bakterija i virusa čime brane telo
- Kontrakcijski proteini učestvuju u kontrakcijama mišića i kretanju.

- Strukturni proteini su vlaknasti i obezbeđuju strukturu i podršku ćelijama. Učestvuju u izdgradnji kose, noktiju, kože, kostiju, itd.
- Transportni proteini prenose molekule kroz telo.
- Hormonski proteini prenose signale kako bi upravljali biološkim procesima među ćelijama, tkivima i organima.
- Skladišni proteini čuvaju aminokiseline za kasniju upotrebu [10, 11, 12].

Glava 3

Podaci i metode binarne klasifikacije

U ovom poglavlju biće opisani korišćeni podaci i način njihovog predstavljanja u računaru. Zatim će ukratko biti opisane metode binarne klasifikacije koje su korišćene za predviđanje funkcija proteina.

3.1 Podaci

Podaci o proteinima mogu se pronaći u biomedicinskim bazama podataka, a neke od njih prikazane su u tabeli 3.1.

Baza podataka	URL	Opis
UniProtKB	uniprot.org	Proteinske sekvence i funkcije proteina
PFAM	pfam.xfam.org	Proteinske familije
PDB	wwpdb.org	Eksperimentalno utvrđene strukture
ModBase	modbase.compbio.ucsf.edu	Strukture utvrđene predviđanjem
I2D	ophid.utoronto.ca	Interakcije između proteina
GEO	www.ncbi.nlm.nih.gov/geo	Podaci o genskoj ekspresiji
PRIDE	www.ebi.ac.uk/pride	Podaci dobijeni masenom sprektometrijom

Tabela 3.1: Prikaz nekih javno dostupni biomedicinskih baza podataka [1, 2].

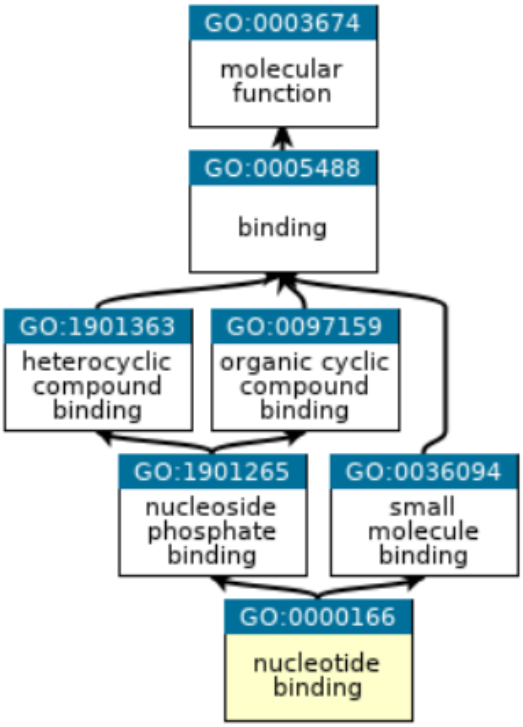
3.1.1 Predstavljanje proteina

Kao što je već rečeno, proteini su izgrađeni od 20 različitih aminokiselina, a svaka aminokiselina ima jedinstveni simbol (tabela 2.2). Najjednostavniji način za predstavljanje proteina u računaru jeste kao niska karaktera nad azbukom $\Sigma = \{A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y\}$. Nad ovako predstavljenim proteinima mogu su koristiti algoritmi za rad sa tekstom kao što je poravnanje sekvenci. [1]

3.1.2 Predstavljanje funkcije proteina

Da bi predviđanje funkcija proteina bilo moguće neophodno je da postoji kontrolisan rečnik i dobro definisani odnosi između funkcija. Sistem za predstavljanje funkcije proteina koji se trenutno najviše koristi je *Gene Ontology*. Ovaj sistem deli funkcije proteina na tri ontologije: biološki procesi (BPO), molekulske funkcije (MFO) i ćelijske komponente (CCO).

Svaka ontologija predstavljena je kao usmereni aciklički graf gde su čvorovima pridruženi nazivi funkcija, a grane koje ih povezuju definišu relaciju „is_a". Hijerarhijska organizacija obezbeđuje da svaki čvor ima specifičniju funkciju od roditeljskog čvora. Međutim, hijerarhija nije striktna zbog čega jedan čvor može imati više roditeljskih čvorova. U korenu svake ontologije nalazi se funkcija sa nazivom te ontologije, a u listovima su najspecifičnije fukcije. Na slici 3.1 prikazan je podgraf ontologije molekulskih funkcija. [2, 13]



Slika 3.1: Prikaz podgrafa ontologije molekulskih funkcija za funkciju „nucleotid binding".

3.2 Binarni klasifikatori

Klasifikacija, odnosno, zadatak dodeljivanja objekata jednoj od više predefinisanih kategorija, rasprostranjen je problem koji obuhvata mnoštvo različitih primena. Primeri uključuju otkrivanje spam poruka na osnovu zaglavlja poruke i njenog sadržaja, kategorisanje ćelija kao maligne ili beningne na osnovu rezultata magnetne rezonance, klasifikacija galaksija na osnovu njihovog oblika, itd. Binarna klasifikacija je slučaj klasifikacije u kojoj postoje tačno dve predefinisane kategorije u koje treba razvrstati date objekte. Obično se za jednu kategoriju kaže da je to pozitivna klasa, a za drugu da je negativna.

Svaki klasifikator upotrebljava algoritam za učenje kako bi odredio model koji najbolje odgovara vezi između skupa atributa i klase ulaznih podataka. Model koji algoritam generiše trebalo bi da odgovara ulaznim podacima kao i da tačno predviđa klasu slogova koje ranije nije video.

Bibliografija

[1] Predrag Radivojac. *A (not so) Quick Introduction to Protein Function Prediction*. 2013.

[2] Jovana Kovačević. *Strukturna predikcija funkcije proteina i odnos funckionalnih kategorija i neuređenosti*. 2015.

[3] Rick Ricer Michael A. Lieberman. *Biochemistry, Molecular Biology, and Genetics*. New Science Press Ltd, 2004.

[4] "Anatomy and Phisiology". In: ().

[5] Denise R. Ferrier Richard A. Harvey. *Biochemistry Fifth Edition*. Lippincott Williams & Wilkins, a Wolters Kluwer business, 2011. ISBN: 978-1-60831-412-6.

[6] Vesna Spasojević-Kalimanovska Slavica Spasić Zorana Jelić-Ivanović. *Opšta biohemija*. 2002.

[7] Dubravka Cvorišćec Ivana Čepelak. *Štrausova medicinska biokemija*. Medicinska naklada, 2009.

[8] Marek Kimmel Andrzej Polanski. *Bioinformatics*. Springer-Verlag, 2007. ISBN: 978-3-540-24166-9.

[9] Dagmar Ringe Georgy A Pesko. *Protein Structure and Function*. New Science Press Ltd, 2004.

[10] Regina Bailey. "The Function and Structure of Proteins". In: (2019).

[11] "Role of proteins in the body". In: (2011).

[12] "What are proteins and what do they do?" In: (2019).

[13] *Gene Ontology*.