

Predviđanje funkcije proteina metodama binarne klasifikacije

Anja Bukurov

Matematički fakultet

26.06.2019.

Sadržaj

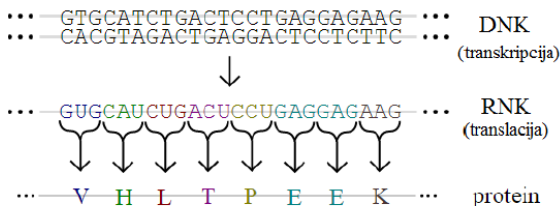
- 1 Motivacija
- 2 Uvodni pojmovi
 - Proteini
 - Binarna klasifikacija
 - Evaluacija modela
- 3 Implementacija
 - Predstavljanje proteina i funkcija
 - Obučavanje modela
 - Objedinjavanje modela
 - Evaluacija prediktora
- 4 Rezultati

Motivacija

- Svake godine sekvencira se veliki broj novih genoma
- Eksperimentalno određivanje funkcije je skup i spor proces
- Ulaže se veliki trud u razvoj računarskih metoda koji mogu da predvide funkciju proteina
- Mnogi postojeći pristupi koriste informacije o sekvenci proteina na neki način

Sinteza proteina

- Makromolekuli koji igraju mnoge kritične uloge u organizmu
- Sačinjavaju više od 50% suvog dela ćelije i važni su za njenu izgradnju i funkcionisanje
- Sinteza proteina sastoji se iz dva koraka: transkripcije i translacije



Struktura proteina

- Proteini su izgrađeni od 20 standardnih aminokiselina
- Aminokiseline su jedinjenja koja sadrže jednu karboksilnu grupu, jednu amino grupu i bočni R-lanac
- Dve aminokiseline se vezuju peptidnom vezom koja se formira između ugljenika iz karboksilne i azota iz amino grupe
- Redosled aminokiselina jedinstven je za svaki protein i čini njegovu primarnu strukturu

Binarna klasifikacija

- Zadatak dodeljivanja objekta jednoj od dve predefinisane kategorije
- Svaki klasifikator koristi algoritam za učenje kako bi odredio model koji najbolje odgovara vezi između skupa atributa i kategorija ulaznih podataka
- Model bi trebalo da odgovara ulaznim podacima i da tačno predviđa klasu slogova koje ranije nije video

Metod potpornih vektora

- Tehnika zasnovana na pronalasku razdvajajuće hiperravni
- Sve instance iste klase treba da se nađu sa iste strane hiperravni
- Takvih ravni množe biti mnogo, ali nisu sve podjednako dobre
- Traži se ona koja maksimizuje rastojanje između instanci dve klase

Logistička regresija

- Statistički zasnovana metoda
- Zadatak je pronaći hiperravan koja deli podatke tako da sa jedne strane budu instance iste klase
- Računa se verovatnoća da instanca pripada jednoj od klasa
- Verovatnoća je veća što je instanca dalja od hiperravni sa odgovarajuće strane

Slučajne šume

- Metod asambla dizajniran za stabla odlučivanja
- Čvorovi stabla sadrže pitanja, a grane su odgovori na njih
- Listovi sadrže oznake klasa
- Klasifikacija se vrši glasanjem - svako od stabla klasifikuje instancu, a prebrojavanjem se odlučuje koja je klasa

Evaluacija modela

- Tačnost
- Preciznost
- Odziv
- f_1
- Površina ispod ROC krive

Predstavljanje proteina

- Aminokiseline su u računar predstavljene kao jedan karakter
- Sekvenca je predstavljena kao niska aminokiselina
- Protein je predstavljen kao niz dimenzije 20^3 gde svaki element predstavlja broj pojavljivanja odgovarajućeg trigrama
- Trigram se preslikava u broj po formuli

$$trigram_broj = ak_1 * 20^2 + ak_2 * 20 + ak_3$$

pri čemu svaka aminokiselina ima dodeljen broj iz intervala $[0, 19]$

Predstavlanje funkcija

- Sistem za predstavljanje funkcije proteina koji se trenutno najviše koristi je *Gene Ontology*
- Funkcije proteina podeljene su na tri ontologije: biološki procesi (BPO), molekulske funkcije (MFO) i ćelijske komponente (CCO).
- Ontologija je predstavljena kao usmereni aciklički graf gde su čvorovima pridruženi nazivi funkcija, a grane koje ih povezuju definišu relaciju *is_a*
- Svaki čvor ima specifičniju funkciju od roditeljskog čvora
- U korenu svake ontologije nalazi se funkcija sa nazivom te ontologije, a u listovima su najspeifičnije fukcije

Skupovi za obučavanje i evaluaciju

- Početni skup od 20960 proteina podeljen je na trening i test skup
- Za trening je izdvojeno 20860 proteina koji su korišćeni za obučavanje pojedinačnih binarnih modela
- Na test skupu od 100 proteina upoređeni su konačni prediktori čiji odgovor predstavlja uniju odgovora svih binarnih klasifikatora

Obučavanje modela

- Korišćene su implementacije iz Python biblioteke *sklearn*
- Za svaku funkciju trenirano je više modela sa različitim parametrima i odabran je najbolji po f_1 -meri
- Svaki model je obučen na 75% trening skupa koji je prethodno podeljen na pozitivne i negativne instance
- 25% skupa korišćeno je za validaciju i odabir modela

Parametri

- Metod potpunih vektora: $C \in \{0.001, 0.1, 1, 10\}$ i $kernel \in \{linear, rbf\}$
- Linearna regresija: $C \in \{0.00001, 0.0001, 0.001, 0.1, 1\}$
- Slučajne šume: $n_estimatoris \in \{100, 400, 700, 1000\}$

Objedinjavanje modela

- Za svaku od tri metode napravljen je jedan prediktor
- Prediktor kao ulaz prima jednu proteinsku sekvencu, a izlaz je podgraf ontologije koji predstavlja funkciju zadatog proteina
- Svaki je fomiran na osnovu prethodno obučenih binarnih modela
- Tokom klasifikacije jednog proteina, sekvenca se prosleđuje svakom od modela koji daju odgovor za konkretnu funkciju
- Svi odgovori se spajaju u konačan odgovor prediktora - podgraf ontologije

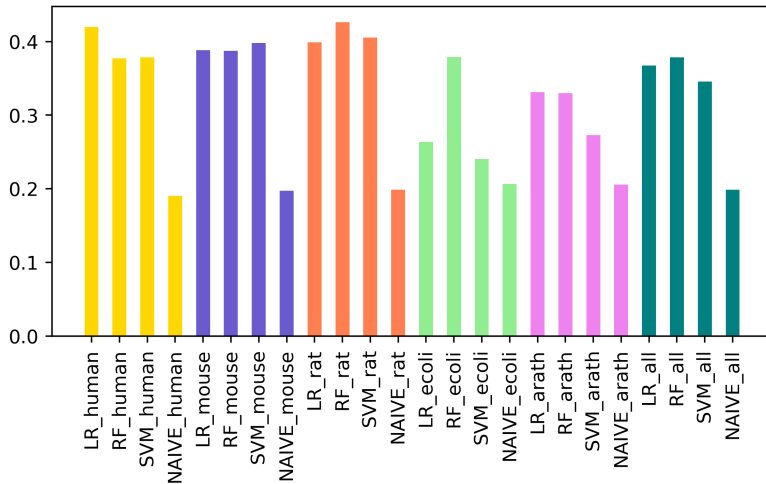
Evaluacija prediktora

- Svi odgovori se spajaju u jedan niz čime se dobija klasa za svaku od funkcija
- Za svaki protein iz test skupa je eksperimentalno određena funkcija na osnovu čega je formiran i niz stvarnih klasa za svaku od funkcija
- Na osnovu ovih nizova - stvarne i predviđene klase, lako se određuju sve mere kvaliteta prediktora

Rezultati

- Prediktori su testirani nad skupom od 100 proteina i međusobno upoređeni
- Implementiran je i naivan klasifikator kao osnovni metod za poređenje
- Vrednosti za svaku meru kvaliteta su računate na nivou svih organizama, i na pojedinačnim organizmima (čovek, miš, pacov, ešerihija i arabidopsis)

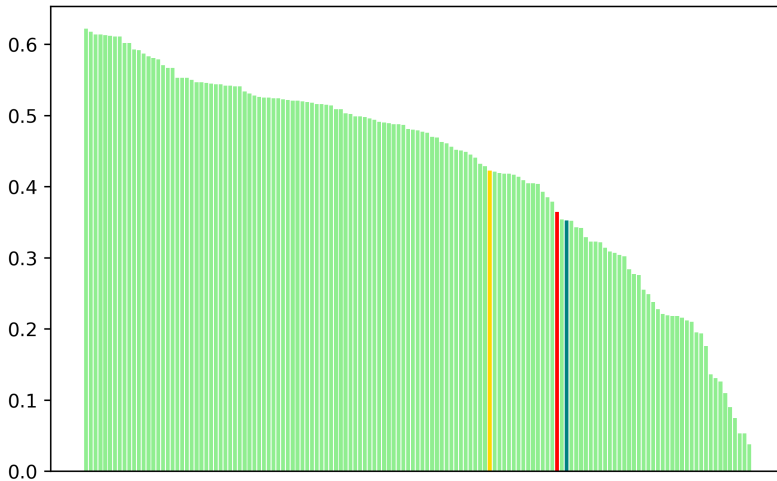
F1-mera



Poređenje sa CAFA

- Prediktori su dodatno testirani na *benchmark* skupu koji je korišćen u okviru CAFA3 takmičenja
- Skup sadrži 453 proteina
- Poređenje je izvršeno na osnovu f_1 -mere

F1 mera



Zaključak

- Trenirani prediktori nemaju približnu moć predviđanja u poređenju sa rezultatima prikazanim na poslednjem CAFA takmičenju
- Planovi za njihovo poboljšanje uključuju treniranje pojedinačnih modela i prediktora za svaki organizam pojedinačno, povećanje trening skupa, izmenu ulaznih podataka, korišćenje raznovrsnijih metoda binarne klasifikacije

Hvala na pažnji!