Predviđanje funkcije proteina metodama binarne klasifikacije

Anja Bukurov

Matematički fakultet

26.06.2019

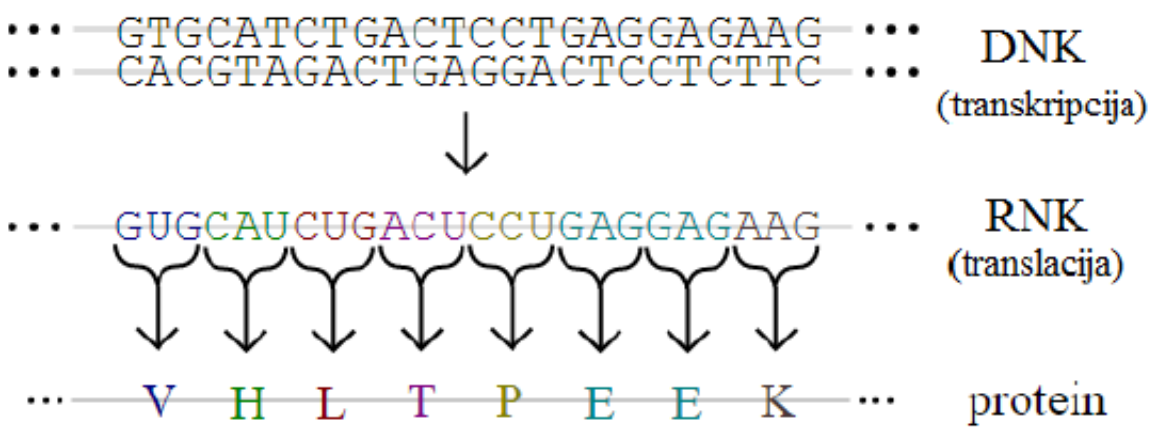
Sadržaj

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 standarnih 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 može biti mnogo, ali nisu sve podjednako dobre
- Traži se ona koja maksimizuje rastojanje između instanci dve klase

Metod potpornih vektora

- Tehnika zasnovana na pronalasku razdvajajuće hiperravni
- Sve instance iste klase treba da se nađu sa iste strane hiperravni
- Takvih ravni može 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

Predstavljanje proteina

- Aminokiseline su u računaru 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 ak_i iz intervala $[0, 19]$

Predstavljanje 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 acikli£ki 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 najspecifičnije fukcije

Skupovi za obučavanje i evaluaciju

- Početni skup 20960 proteina podeljen je na trening i test skup
- Za trening je izdvojeno 20860 proteina koji su korišćeni za obučavanje pojedinačnih binarnih modela
- Svaki binarni model predstavlja najbolji od nekoliko modela koji se razlikuju po korišćenim parametrima, a odabrani su na osnovu rezultata postignutih na podskupu trening skupa koji je izdvojen za validaciju
- Na test skupu od 100 proteina upoređeni su konačni prediktori čiji odgovor predstavlja uniju odgovora svih binarnih klasifikatora

Rezultati

Hvala na pažnji!