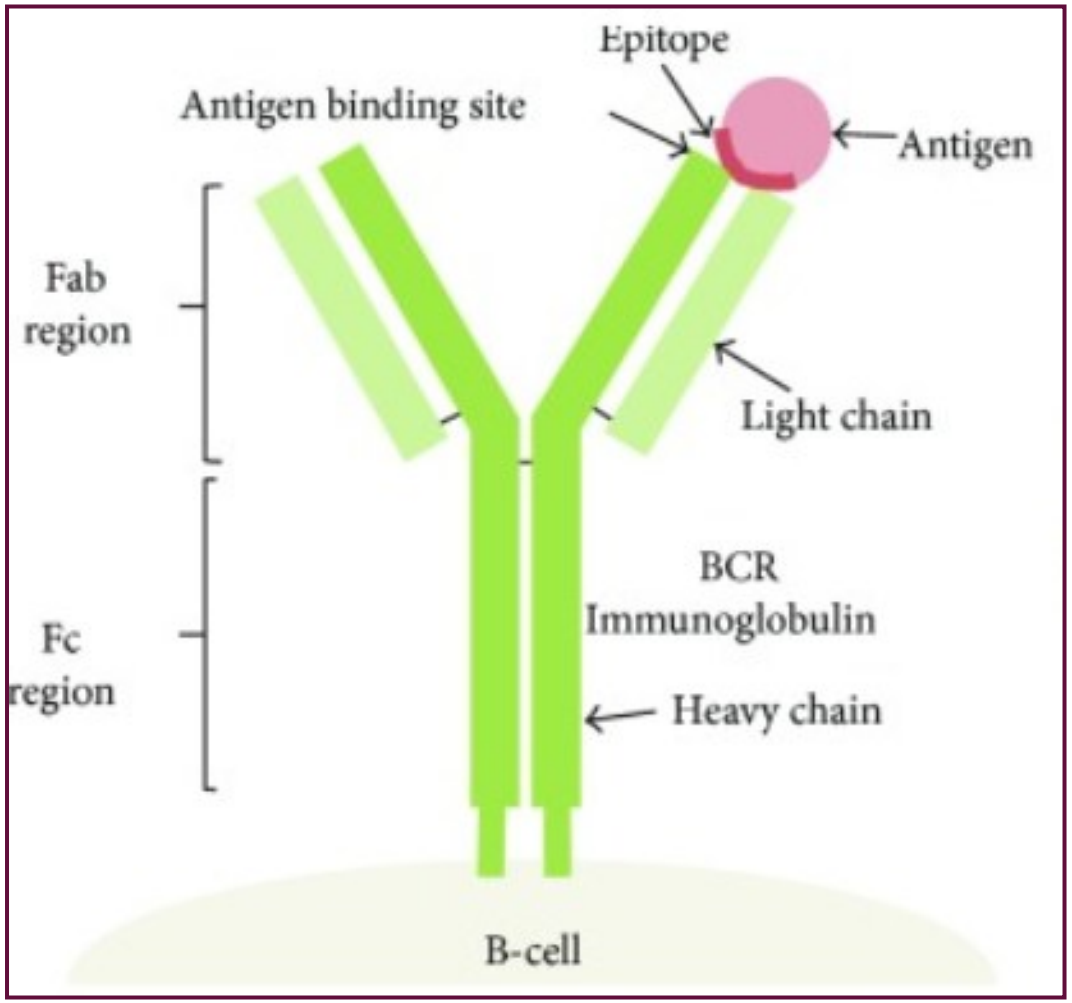
Predikcija B-cell epitopa u antigenima

Ante Sorić
Domagoj Sviličić
Martin Bugarin
Anđelko Prskalo
Diego Mišetić
Lana Barić

Uvod – Što su B-cell epitopi?

- Specifični dijelovi antigena koji aktiviraju imunološki odgovor tijela vezanjem na B-cell
- Antigeni se sastoje od peptida izgrađenih od 20 standardnih aminokiselina
- B-cell epitopi su podnizovi unutar tih sekvenci na koje se vežu B-cell
- Dizajn cjepiva i istraživanja imunoloških odgovora
- Dvije kategorije epitopa: **linearni** i konformacijski
- Linearni su sačinjeni od uzastopnih aminokiselinskih ostataka



Predviđanje B-cell epitopa

Primjena rekurentnih neuronskih mreža (RNN) u predviđanju tih epitopa

Glavni izazovi: varijabilna duljina sekvenci i kompleksne interakcije između aminokiselina koje utječu na imunološka svojstva

RNN idealne jer uzimaju u obzir sekvencijalne odnose i pamte prethodne elemente unutar niza

Naš projekt

- Baza IEDB za stvaranje skupa podataka
- Pozitivni uzorci obuhvaćaju poznate B-cell epitope, a negativni su generirani iz nasumičnih proteinskih sekvenci
- Biološka relevantnost i pouzdanost modela
- Primjenom RNN-a istražuje se kako duljina epitopa i okruženje unutar proteinske sekvence utječu na imunološka svojstva

Postojeći pristupi

- Sustav ABCpred temeljen na RNN
- Baza Bcipep: 2479 kontinuiranih B-staničnih epitopa
- RNN (bolji rezultati) i unaprijedne neuronske mreže (FNN)
- Maks. duljina epitopa: 20 aminokiselina
- Nadopunjavanje kraćih epitopa sa susjednim aminokiselinama iz njihovih sekvenci
- Metoda peterostruke unakrsne validacije
- Najveća točnost: prozor duljine od 16 aminokiselina i jedan sloj od 35 skrivenih jedinica

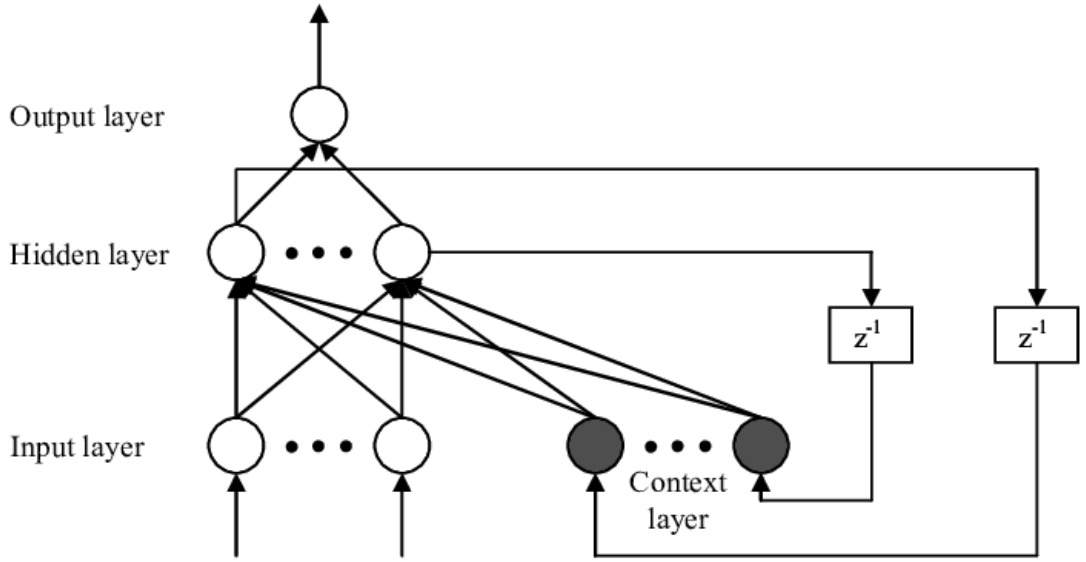
Opis rješenja problema

Podaci

- Izvučeno 4097 proteinskih sekvenci koje sadrže B-stanične epitope i isto toliko onih koji ih ne sadrže
- Nadopunjavanje epitopa s obje strane (simetrično) aminokiselinama iz njihove izvorišne proteinske sekvence
- Npr. AAALPGKCGV -> PNNAAALPGKCGVHIP
- One-hot enkodiranje za kodiranje aminokiselina
- Svaka aminokiselina pretvorena u binarni vektor duljine 21 (20 aminokiselina i oznaka X)
- 60:20:20 podjela skupa podataka

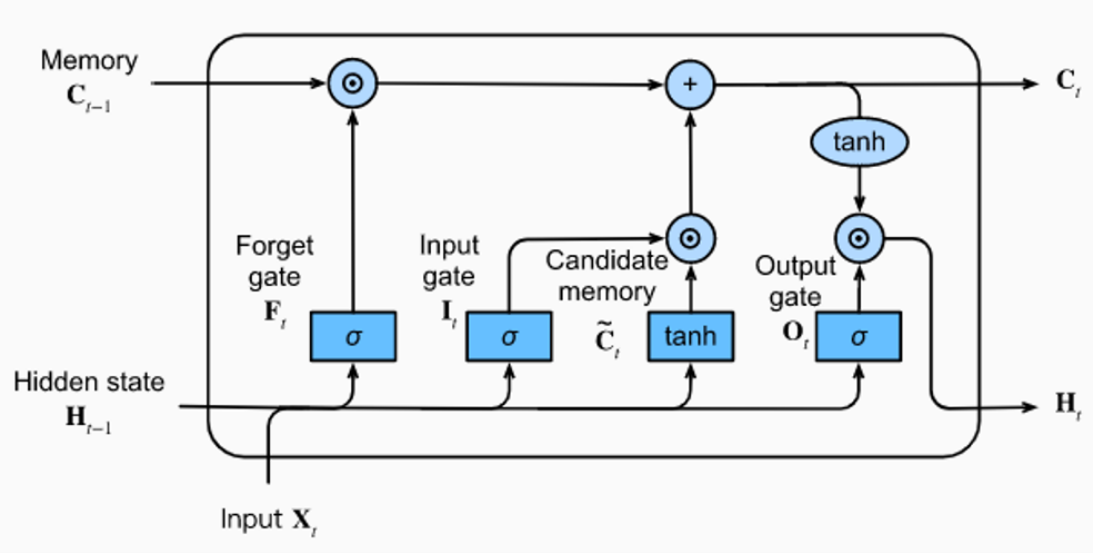
Elmanova mreža

- Ulazni, skriveni i izlazni sloj
- Prima ulazni vektor x_t koji se prosljeđuje skrivenom sloju
- Skriveni sloj povezan sam sa sobom → skriveno stanje prethodnog vremenskog koraka se prosljeđuje nazad u trenutni skriveni sloj
- Kratkoročna memorija
- Završno skriveno stanje h_t se lin. transformacijom pretvara u jedan broj
- Na izlazu sigmoida
- Problem nestajanja ili eksplodiranja gradijenta



LSTM mreža

- Dizajnirana da uhvate dugoročne ovisnosti između sekvencijalnih podataka
- Stanje ćelije c_t → pohranjivanje dugoročnih informacija koje su relevantne za zadatak
- Skriveno stanje h_t → prenosi informaciju o tome što je mreža dosad naučila u obradi sekvence i prenosi u sljedeći vremenski korak obrade
- Ažuriranje stanja ćelije i skrivenog stanja nakon svakog vremenskog koraka
- Vrata zaborava → koliko informacija iz ćelija stanja treba odbaciti
- Vrata ulaza → koliko novih informacija treba biti dodano ćeliji stanja
- Vrata izlaza → kako će izgledati novo skriveno stanje



Zajedničko

Postupak propagacije pogreške unazad kroz vrijeme (BPTT)

BPTT računa gradijente funkcije pogreške te ažurira težine mreže kako bi smanjio pogrešku

Evaluacija točnosti modela: osjetljivost, specifičnost, točnost, pozitivna prediktivna vrijednost (PPV)

Opis eksperimentalnih rezultata

TABLICA II - PERFORMANSE LSTM MODELA

Veličina prozora	Osjetljivost	Specifičnost	PPV	Točnost
16	70.18	76.20	69.38	73.59
18	71.76	76.46	69.21	74.47
20	75.64	77.07	73.44	76.41
22	81.74	75.83	73.26	78.47
24	72.59	82.73	76.95	78.24
26	76.68	83.69	79.66	80.51

TABLICA I - PERFORMANSE ELMANOVOG MODELA RNN-A

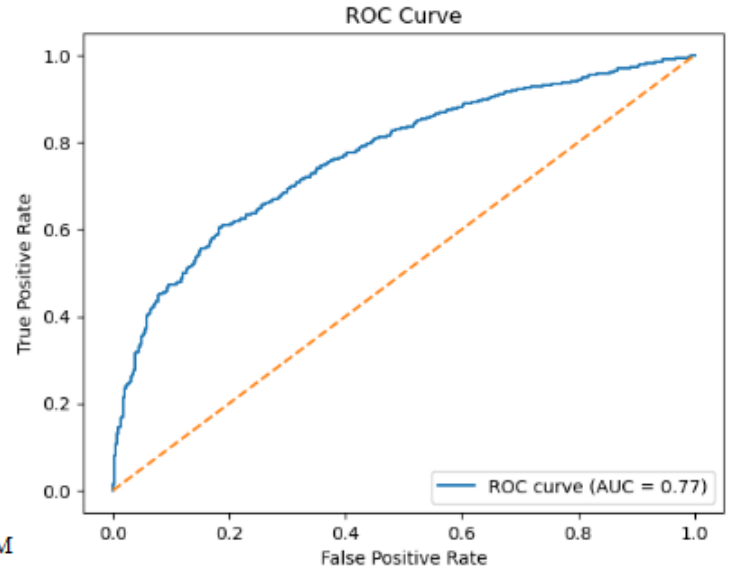
Veličina prozora	Osjetljivost(%)	Specifičnost(%)	PPV(%)	Točnost(%)
16	42.50	91.52	79.38	70.22
18	45.38	88.10	73.77	69.97
20	60.53	77.19	68.99	69.60
22	45.51	89.77	78.27	69.96
24	61.26	81.27	72.20	72.41
26	70.70	75.47	70.60	73.30

- Različite fiksne veličine epitopa te različite kombinacije hiperparametara
- Najbolja točnost nad fiksnim duljinama sekvenci: Elman 70% , LSTM 80%

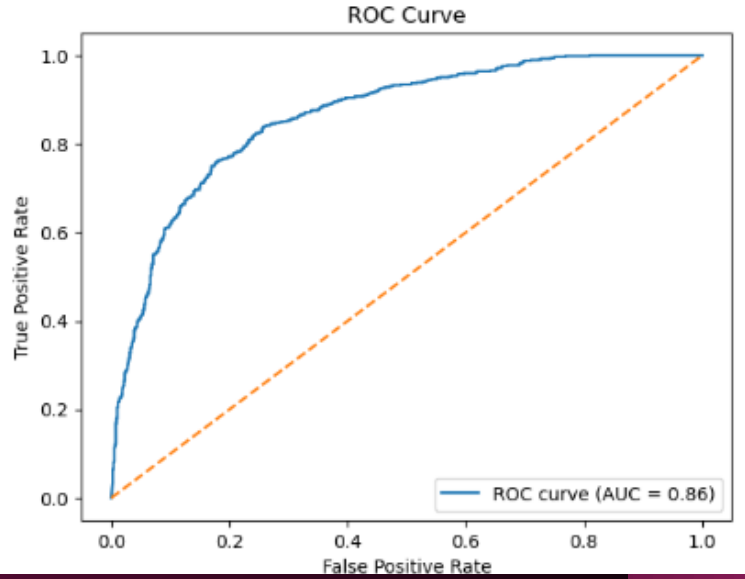
Veličina prozora 26

- Elmanov model: stopa učenja 0.005, vel. skrivenog sloja 30, batch size 64, Adam optimizator
- LSTM: stopa učenja 0.01, vel. skrivenog sloja 40, batch size 128, Adam optimizator
- Elman: AUC = 0.77 → umjerene performanse
- LSTM: AUC = 0.86 → superiorne performanse

SLIKA I – ROC KRIVULJA ZA ELMANOV RNN MODEL S VELIČINOM PROZORA 26



SLIKA II – ROC KRIVULJA ZA LSTM MODEL S VELIČINOM PROZORA 26



Sekvence varijabilnih duljina

TABLICA IV – PERFORMANSE LSTM MODELA

Veličina prozora	Osjetljivost	Specifičnost	PPV	Točnost
16	60.02	66.52	35.56	64.99
18	68.64	62.26	35.89	63.76
20	69.16	61.57	35.65	63.36
22	72.50	58.51	34.97	61.80
24	71.88	60.04	35.64	62.83
26	77.14	50.86	32.58	57.04

TABLICA III – PERFORMANSE RNN MODELA

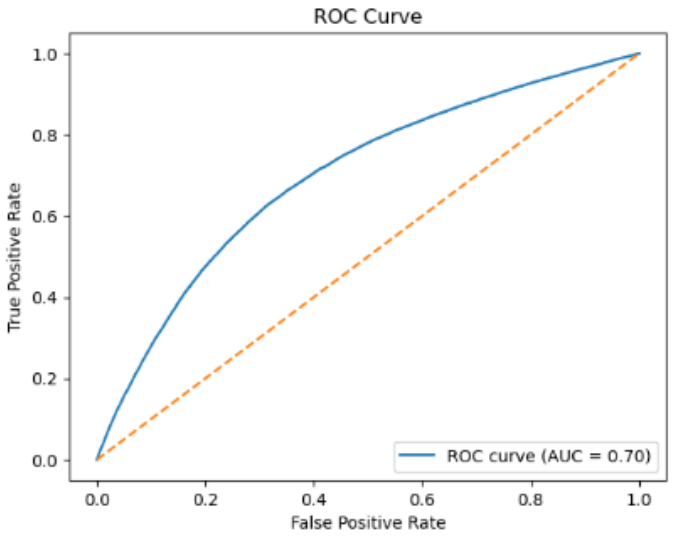
Veličina prozora	Osjetljivost	Specifičnost	PPV	Točnost
16	35.93	81.53	37.45	70.80
18	28.77	87.16	40.81	73.42
20	47.74	72.55	34.87	66.71
22	69.09	41.55	26.68	48.03
24	66.82	54.24	31.01	57.20
26	60.41	57.77	30.57	58.39

- Niža točnost predikcija, povećani broj lažno pozitivnih slučajeva
- LSTM daje bolje performanse u svim važnim metrima u odnosu na Elmanov RNN model

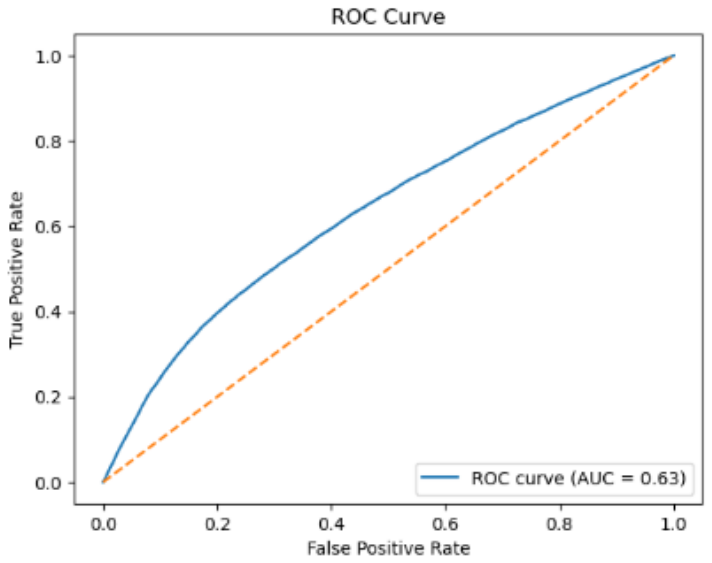
Najbolji model ?

- Prozor od 20 sekvenci:
  - RNN: dobar balans između prepoznavanja pozitivnih uzoraka i minimiziranja lažno pozitivnih, solidna točnost
  - LSTM: hvatanje šireg konteksta epitopa uz visoku osjetljivost, prihvatljiv broj lažno pozitivnih rezultata
- Prozor 24 približno dobar

SLIKA IV – ROC KRIVULJA ZA LSTM MODEL S VELIČINOM PROZORA 20



SLIKA III – ROC KRIVULJA ZA ELMANOV RNN MODEL S VELIČINOM PROZORA 20



Diskusija / Zaključak

- Najveća točnost: LSTM 80.51% za duljinu sekvence 26
- LSTM arhitektura omogućuje bolje razumijevanje dugoročnih ovisnosti unutar sekvenci
- Elmanov RNN je jednostavniji pristup s nešto nižim performansama
- Razlog lošijeg predviđanja: B-cell epitopi nemaju fiksnu duljinu, a koristi se prozor s fiksnom duljinom
- Poboljšanje: bolje upravljanje epitopa s promjenjivim duljinama, uz dinamičko podešavanje veličine prozora ili korištenje dužih sekvenci

Literatura

- Saha, Sudipto & Raghava, Gajendra. (2006). Prediction of Continuous B-cell Epitopes in an Antigen Using Recurrent Neural Network. Proteins. 65. 40-8. 10.1002/prot.2107
- https://pmc.ncbi.nlm.nih.gov/articles/PMC5763123/
- https://www.researchgate.net/figure/Architectural-graph-of-Elman-neural-network_fig1_224393607
- https://medium.com/@ottaviocalzone/an-intuitive-explanation-of-lstm-a035eb6ab42c