SVEUČILIŠTE U RIJECI

**TEHNIČKI FAKULTET**

Diplomski sveučilišni studij računarstva

Diplomski rad

**ANALIZA SLIČNOSTI PEPTIDA TEMELJENA NA LATENTNOM PROSTORU AUTOENKODERA**

Rijeka, rujan 2022. Bruno Novosel   
 0069078199

SVEUČILIŠTE U RIJECI

**TEHNIČKI FAKULTET**

Diplomski sveučilišni studij računarstva

Diplomski rad

**ANALIZA SLIČNOSTI PEPTIDA TEMELJENA NA LATENTNOM PROSTORU AUTOENKODERA**

Mentor: doc. dr. sc. Goran Mauša

Rijeka, srpanj 2020. Bruno Novosel   
 0069078199

IZJAVA

Izjavljujem da sam ovaj rad izradio samostalno, uz vodstvo i stručnu pomoć mentora doc. dr. sc. Gorana Mauše.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Bruno Novosel

ZAHVALA

Zahvaljujem mentoru doc. dr. sc. Goranu Mauši na pomoći, stručnom vođenju i velikom strpljenju kojim me vodio kroz izradu ovog rada.

Sadržaj

[1. UVOD 1](#_Toc110008986)

[2. PEPTIDI 2](#_Toc110008987)

[2.1. Što su peptidi? 2](#_Toc110008988)

[2.2. Antimikrobijska svojstva 4](#_Toc110008989)

[2.3. Antiviralna svojstva 5](#_Toc110008990)

[2.4. Sekundarne strukture 6](#_Toc110008991)

[2.5. Predviđanje sekundarne strukture 8](#_Toc110008992)

[3. AUTOENKODERI 9](#_Toc110008993)

[3.1. Varijacijski autoenkoderi 9](#_Toc110008994)

[4. MODEL 10](#_Toc110008995)

[5. ZAKLJUČAK 11](#_Toc110008996)

[LITERATURA 12](#_Toc110008997)

[POPIS KRATICA 14](#_Toc110008998)

[Sažetak 15](#_Toc110008999)

[Abstract 15](#_Toc110009000)

# UVOD

Razvojem medicine i znanosti došlo je do otkrića i masovne proizvodnje cjepiva i antivirusnih lijekova, no usprkos tome virusi i dalje predstavljaju glavni uzrok bolesti kod ljudi te se svakodnevno traže nova rješenja kojima bi se suzbili određeni virusi. Veliki potencijal u tome području imaju antiviralni peptidi (AVP). Iako oni predstavljaju pomalo nekonvencionalan pristup liječenju virusnih infekcija i otkriveno ih je relativno malo, u mnogim istraživanjima su pokazali obećavajuće rezultate [1]. Usprkos tome, vrlo malo se poduzima oko predviđanja svojstava AVP i poboljšanja već otkrivenih metoda predviđanja.

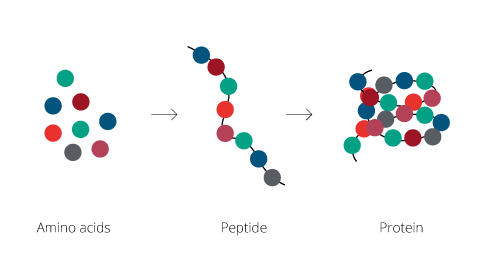
Pokazalo se da se mnoga svojstva AVP, kao što su naboj, gustoća naboja, indeks nestabilnosti, alifatski indeks, hidrofobni omjer itd., mogu upotrijebiti u predviđanju antiviralnih svojstava peptida [2]. Općenito govoreći, AVP su dio antimikrobijskih peptida (AMP), koji se bore protiv bakterija, gljivica, parazita i virusa, ali je otkriveno da predviđanja koja se koriste za AMP ne rade dobro i za AVP. Zbog toga se pokušava izraditi što bolji model koji bi predviđao svojstva izričito za AVP. Jedno takvo istraživanje proveli su i Kuan Y. Chang i Je-Ruei Yang  
na Odjelu za računarstvo sa Nacionalnog Sveučiliša Tajvanskog Oceana [1] gdje su dokazali da njihov model predviđanja koji se bazira na metodi *random forest* radi puno bolje od dotadašnjih modela te je to istraživanje i glavna motivacija za ovaj rad.

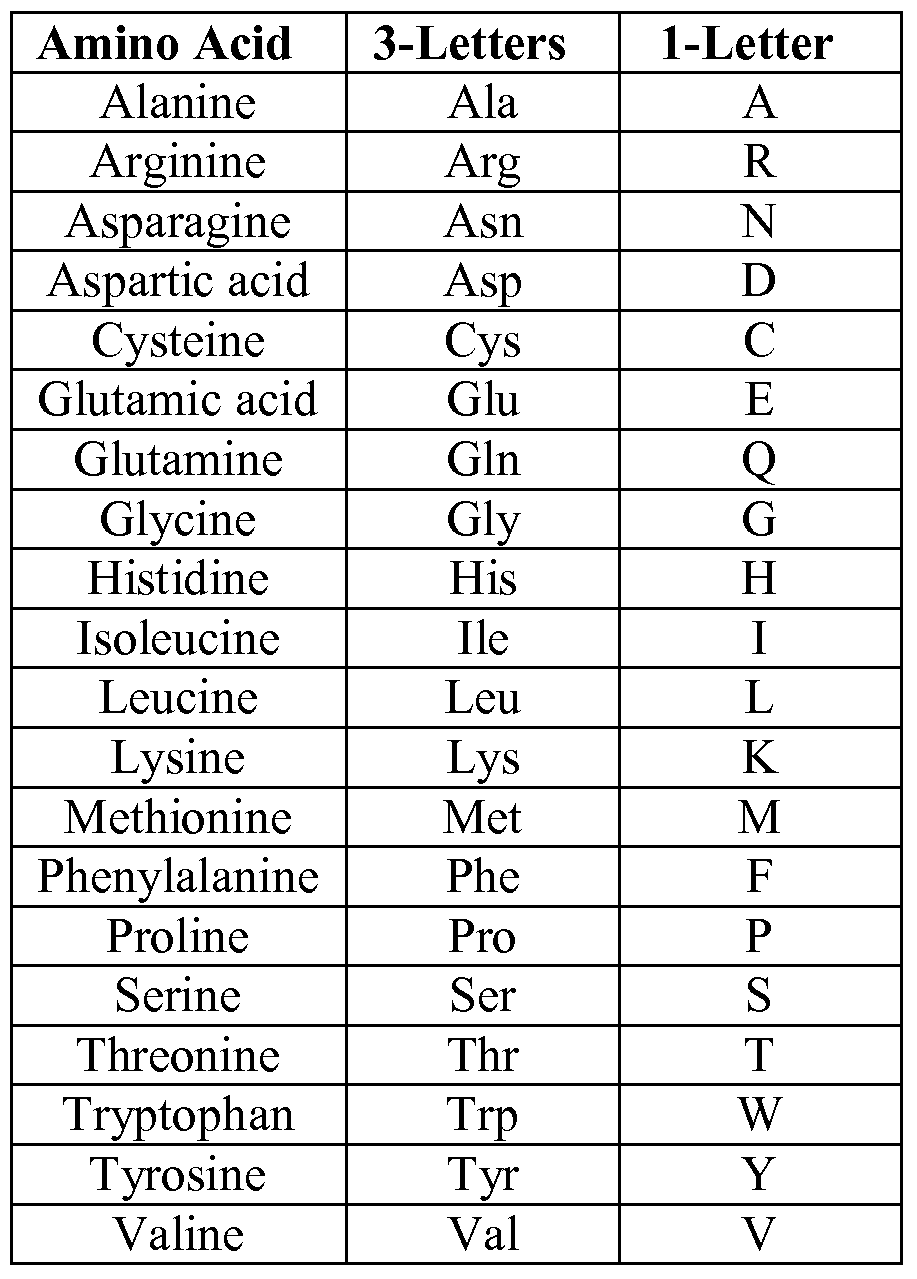
1. PEPTIDI

## Što su peptidi?

Peptidi su kraća verzija proteina, kako je i prikazano slikom 2.1. Prema definiciji, peptidi su kratki lanci sastavljeni od 2-50 aminokiselina, povezanih peptidnim vezama [3]. Prema izvoru i funkciji podijeljeni su u više grupa, od kojih su samo neke – biljni peptidi, antimikrobijski peptidi, moždani peptidi, endokrini peptidi, kardiovaskularni peptidi itd., iz čega odmah možemo vidjeti da imaju vrlo široku primjenu i koriste se za sve od ublažavanja upala, preko izgradnje mišića do uništavanja mikroba. Popis svih prirodnih aminokiselina koje grade peptide i njihovih oznaka prikazan je slikom 2.2.

Na peptide nailazimo svakodnevno i u namirnicama biljnog ili životinjskog porijekla kao što su mlijeko, jaja, meso, soja, pšenica itd., a kako su kraći od proteina to ih čini pogodnijima našem tijelu za procesiranje.

  
*Slika 2.1. Pojednostavljeni prikaz aminokiselina, peptida i proteina (preuzeto iz [4])*



*Slika 2.2. Popis svih prirodnih aminokiselina i njihove oznake (preuzeto iz [5])*

## Antimikrobijska svojstva

Antimikrobijska svojstva su vrlo bitna kod peptida jer AMP su kao prvo molekule koje se prirodno pojavljuju i kreću se proizvoditi kao prva linija obrane od raznih bolesti kod svih višestaničnih organizama. Postoje i sintetički stvoreni AMP, naravno po uzoru na prirodne, a obje vrste se mogu boriti protiv raznih vrsta virusa, bakterija, gljivica i parazita.  
Većina AMP otkrivenih do danas se mogu okarakterizirati kao jedan od sljedeća četiri tipa prema njihovim sekundarnim strukturama: *β-sheet*, *α-helix*, *extended* i *loop*. Među tim strukturnim grupama, *β-sheet* i *α-helix* su najčešće strukture, a *α-helix* je najviše proučavana skupina do sada [6].  
Iako većina AMP pripada jednoj od četiri navedene grupe prema sekundarnim strukturama, neki AMP ne pripadaju ni jednoj od tih grupa. Također, neki AMP se mogu sastojati i od dvije vrste strukturnih komponenata.

Usprkos mnogim korisnim karakteristikama AMP, još uvijek postoje neke prepreke u njihovom korištenju kao što je moguće otrovno djelovanje na ljude, osjetljivost na grube uvjete okoline, manjak selektivnosti protiv određenih vrsta, visoki troškovi proizvodnje itd.

Možemo spomenuti neke osobito važne fizikalno-kemijske karakteristike AMP [6]:

* duljina – ona je važna za aktivnost AMP jer je potrebno barem 7-8 aminokiselina da bi se stvorila peptidna molekula koja je hidrofobna i hidrofilna na različitim krajevima; obično je za strukturu *α-helix* potrebno najmanje 22 aminokiseline, dok je za *β-sheet* potrebno 8; duljina može imati i utjecaj na otrovna svojstva pa se ona mora uzeti u obzir kod sintetiziranja novih peptida
* ukupni naboj – čini ga suma svih naboja ioniziranih grupa u peptidu; varira od negativnog do pozitivnog i glavni je faktor za inicijalnu interakciju peptida s membranom ćelije koju napada
* hidrofobnost – može utjecati na aktivnost i selektivnost AMP; gotovo 50% aminokiselina u primarnim sekvencama AMP su hidrofobne; u većini slučajeva povećanje hidrofobnosti na pozitivno nabijenoj strani AMP može povećati antimoikrobijsku aktivnost i obrnuto
* topljivost – kako bi AMP trebali djelovati tako da ulaze kroz membrane lipida, trebali bi biti topljivi u okolini gdje je prisutna voda

Kao što vidimo, mnogi faktori utječu na aktivnost AMP, a postoje i neke povezanosti između tih faktora. Pri stvaranju novih AMP to treba uzeti u obzir jer promjenom jednog od tih parametara da bi dobili željenu modifikaciju AMP možemo promijeniti i ostale parametre. Čak i jednostavna promjena u primarnoj sekvenci može uvelike utjecati na fizikalno-kemijska svojstva koja su presudna za aktivnost AMP i spektar ćelija koje on napada, a predviđanje rezultata modifikacije AMP ili funkcije sintetičkog peptida prije nego je on stvoren još uvijek predstavlja određeni problem.

## Antiviralna svojstva

Podvrsta antimikrobijskih peptida – antiviralni peptidi (AVP), je poznata po tome što se bori protiv raznih virusa. Postoji više mehanizama koje AVP koriste u borbi protiv virusa, npr. blokiranje vezanja virusa, prevencija spajanja virusa sa stanicama domaćina, prekid procesa signaliziranja virusa ili sprječavanje razmnožavanja virusa u stanicama domaćina [1].

Kao i ostali peptidi, mogu biti prirodni (dobiveni iz biljaka, insekata, vodozemaca, podmorskih organizama ili sisavaca) ili umjetno sintetizirani. Postoje tri glavne metode za racionalno stvaranje novih peptida [2]:

* dizajn temeljen na predlošku (*template-based design)* – pokušava dodati selektivnost i poboljšati aktivnost već postojeće sekvence peptida dodavanjem neke aminokiseline ili premještanjem već postojeće; ovakve modifikacije mogu dovesti do stvaranja novog AVP čak i od neaktivnih peptida
* fizikalno-kemijska metoda – također koristi već poznate sekvence kako bi stvorili sličnu sekvencu s drugačijim fizikalno-kemijskim svojstvima
* ‘de novo’ metoda – stvara nove peptide koristeći uzorke i učestalosti peptida

Bitno je u ovom radu napomenuti kako je istraživanjem pokazano da je predviđanje AVP prema fizikalno-kemijskim svojstvima dobivenih iz sekvence peptida moguće te da takvo predviđanje bazirano na modelu *random forest* [1] radi bolje od ostalih, do sada isprobanih, modela predviđanja.

# AUTOENKODERI

## Varijacijski autoenkoderi

# MODEL

# ZAKLJUČAK

Peptidi s antiviralnim svojstvima sve više obećavaju u liječenju bolesti izazvanih virusima, a da bi se mogli koristiti i primijeniti treba ih prvo otkriti. Kako ne bi morali testirati svaki peptid za svaki virus, što je vremenski zahtjevno, skupo i neefikasno, tu je vrlo obećavajuće postalo predviđanje antiviralne aktivnosti peptida prema njihovim fizikalno-kemijskim svojstvima.

Takvo predviđanje može biti vrlo učinkovito, ako se dobro izvede. Tako je i ovim jednostavnim modelom strojnog učenja postignuta točnost od preko 80% što je vrlo dobro za neke jednostavne potrebe. Ugađanjem nekih parametara i uključivanjem nekih dodatnih svojstava prema kojima bi se mogla predviđati antiviralna aktivnost, npr. sekundarnih struktura, koje su bile zamišljene kao dio ovog modela, ali zbog određenih problema na kraju nisu uključene u model, mogla bi se i povećati točnost predviđanja i tako bi ono moglo imati i širu primjenu i to čak za neka veća i važnija znanstvena istraživanja. Predviđanje antiviralne aktivnosti peptida predstavlja jedan težak, kompleksan, ali i zanimljiv zadatak te kada bi se shvatio i iskoristio potencijal koji nudi, to bi moglo vrlo pozitivno utjecati na budućnost i ostvariti velike korake u medicini, a samim time i ljudskim životima.

# LITERATURA

[1] Kuan Y. Chang, Je-Ruei Yang, s Interneta, <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3734225/>, 25. lipnja 2020.

[2] Liana Costa Pereira Vilas Boas, Marcelo Lattarulo Campos, Rhayfa Lorrayne Araujo Berlanda, Natan de Carvalho Neves, Octávio Luiz Franco, s Interneta, <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7079787/>, 26. lipnja 2020.

[3] Wikipedia, s Interneta, <https://en.wikipedia.org/wiki/Peptide>, 25. lipnja 2020.

[4] World of Peptides, s Interneta, [world-of-peptides.com](http://www.world-of-peptides.com/what-is-the-difference-between-a-peptide-and-a-protein/), 5. srpnja 2020.

[5] Courtney Simons, s Interneta, <https://cwsimons.com/structure-of-amino-acids-and-proteins/>, 5. srpnja 2020.

[6] [Ali Adem Bahar](https://www.ncbi.nlm.nih.gov/pubmed/?term=Bahar%20AA%5BAuthor%5D&cauthor=true&cauthor_uid=24287494),[Dacheng Ren](https://www.ncbi.nlm.nih.gov/pubmed/?term=Ren%20D%5BAuthor%5D&cauthor=true&cauthor_uid=24287494), s Interneta, <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3873676/>, 26. lipnja 2020.

[7] Wikipedia, s Interneta, <https://en.wikipedia.org/wiki/Protein_secondary_structure>, 27. lipnja 2020.

[8] ThoughtCo., s Interneta, <https://www.thoughtco.com/protein-structure-373563>, 5. srpnja 2020.

[9] Golifescience, s Interneta, <https://www.golifescience.com/secondary-structure-of-proteins/>, 27. lipnja 2020

[10] C.Mirabello, G.Pollastri, s Interneta, <http://distillf.ucd.ie/porterpaleale/>, 29. lipnja 2020.

[11] Wikipedia, s Interneta, <https://en.wikipedia.org/wiki/Relative_accessible_surface_area>, 6. srpnja 2020.

[12] Claus AF Andersen & Burkhard Rost, s Interneta, <https://rostlab.org/papers/2003_rev_assign/paper_slow.html>, 6. srpnja 2020.

[13] Guy Yachdav, Peter Hoenigschmid, Laszlo Kajan, s Interneta, <https://github.com/Rostlab/reprof>, 29. lipnja 2020.

[14] RPBS Web portal, s Interneta, <https://mobyle.rpbs.univ-paris-diderot.fr/cgi-bin/portal.py#forms::psipred>, 29. lipnja 2020.

[15] J Yang, R Yan, A Roy, D Xu, J Poisson, Y Zhang. The I-TASSER Suite: Protein structure and function prediction. Nature Methods, 12: 7-8 (2015),

A Roy, A Kucukural, Y Zhang. I-TASSER: a unified platform for automated protein structure and function prediction. Nature Protocols, 5: 725-738 (2010),

Y Zhang. I-TASSER server for protein 3D structure prediction. BMC Bioinformatics, vol 9, 40 (2008), s Interneta, <https://zhanglab.ccmb.med.umich.edu/I-TASSER/>, 29. lipnja 2020.

[16] Niklas Donges, s Interneta, <https://builtin.com/data-science/random-forest-algorithm>, 30. lipnja 2020.

[17] Wikipedia, s Interneta, <https://en.wikipedia.org/wiki/Decision_tree>, 5. srpnja 2020.

[18] Tutorialspoint, s Interneta, <https://www.tutorialspoint.com/machine_learning_with_python/machine_learning_with_python_classification_algorithms_random_forest.htm>, 30. lipnja 2020.

[19] Tony Yiu, s Interneta, <https://towardsdatascience.com/understanding-random-forest-58381e0602d2>, 6. srpnja 2020.

[20] Abilash R., s Interneta, <https://medium.com/@ar.ingenious/applying-random-forest-classification-machine-learning-algorithm-from-scratch-with-real-24ff198a1c57>, 5. srpnja 2020.

[21] Wikipedia, s Interneta, <https://en.wikipedia.org/wiki/Weka_(machine_learning)>, 30. lipnja 2020.

[22] Eibe Frank, Mark A. Hall, and Ian H. Witten (2016). The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann, Fourth Edition, 2016.

[23] Thakur N, Qureshi A, Kumar M (2012) [AVPpred: collection and prediction of highly effective antiviral peptides.](http://nar.oxfordjournals.org/content/40/W1/W199.long) Nucleic Acids Res. 2012 Jul;40(Web Server issue):W199-204. doi: 10.1093/nar/gks450.  
PMID: [22638580](http://www.ncbi.nlm.nih.gov/pubmed/22638580), s Interneta, <http://crdd.osdd.net/servers/avppred/>, 30. lipnja 2020.

[24] Modlamp dokumentacija, s Interneta, <https://modlamp.org/modlamp.html#module-modlamp.descriptors>, 6. srpnja 2020.

[25] Gufosowa - Own work, CC BY-SA 4.0, <https://commons.wikimedia.org/wiki/File:K-fold_cross_validation_EN.svg>, 1. srpnja 2020.

# POPIS KRATICA

AVP – antiviralni peptidi

AMP – antimikrobijski peptidi

MSA – *multiple sequence alignment*

SVM – *support vector machines*

WEKA - *Waikato Environment for Knowledge Analysis*

# Sažetak

Virusi i dalje predstavljaju glavni uzrok bolesti kod ljudi usprkos masovnoj proizvodnji cjepiva i antivirusnih lijekova, a velik potencijal za nove pomake u tom području predstavljaju peptidi s antiviralnom aktivnošću. Kako bi efikasno otkrivali nove takve peptide potrebno je koristiti neku vrstu predviđanja, a pokazano je da je to moguće na osnovi njihovih fizikalno-kemijskih svojstava. Takvo predviđanje, ako se dobro izvede, može biti vrlo učinkovito. Posebno se istaknulo strojno učenje metodom *random forest*, kakvo je implementirano i u ovom radu. Daljnjim trudom i ulaganjem u ovakva istraživanja može se ostvariti velik napredak u medicini i poboljšati kvaliteta ljudskog života.

Ključne riječi: antiviralni peptidi, predviđanje sekundarne strukture, algoritam *random forest*, strojno učenje

# Abstract

Viruses are still the main cause of human diseases despite the mass production of vaccines and antiviral drugs. Antiviral peptides have a great potential to change things for the better in that particular field. For efficient discovery of such new peptides, some kind of prediction should be used and it was shown that this was possible based on their physicochemical properties. Such prediction, if done right, can be very effective. Machine learning using the random forest method was proven especially efficient, which was also implemented in this study. With further effort and investments in such research, a great progress in medicine can be achieved and the quality of human lives improved.

Keywords: antiviral peptides, secondary structure prediction, random forest algorithm, machine learning