UNIVERZITET U BEOGRADU

BIOLOŠKI FAKULTET

Danijela M. Paunović

Identifikacija *AGP* gena kičice (*Centaurium erythraea,* Gentianaceae) i praćenje njihove ekspresije u odgovoru na mehaničke povrede biljnog tkiva gajenog *in vitro*

Doktorska disertacija

Beograd, 2022

UNIVERSITY OF BELGRADE

FACULTY OF BIOLOGY

Danijela M. Paunović

Identification of *AGP* genes in centaury (*Centaurium erythraea,* Gentianaceae) and monitoring of their expression in response to tissue mechanical wounding *in vitro*

Doctoral Dissertation

Belgrade, 2022

Mentori:

_____

dr Milan Dragićević, viši naučni saradnik

Univerzitet u Beogradu, Institut za biološka istraživanja „Siniša Stanković“, Institut od nacionalnog značaja za Republiku Srbiju

_____

dr Ivana Dragićević, vanredni profesor

Univerzitet u Beogradu, Biološki fakultet

Članovi komisije za odbranu doktorske disertacije:

_____

dr Tijana Cvetić Antić, vanredni profesor

Univerzitet u Beogradu, Biološki fakultet

_____

dr Marko Đorđević, vanredni profesor

Univerzitet u Beogradu, Biološki fakultet

_____

dr Marija Marković, viši naučni saradnik

Univerzitet u Beogradu, Institut za biološka istraživanja „Siniša Stanković“, Institut od nacionalnog značaja za Republiku Srbiju

Datum odbrane:_____

Zahvalnica

Eksperimentalni deo doktorske disertacije urađen je u laboratorijama Odeljenja za fiziologiju biljaka Instituta za biološka istraživanja „Siniša Stanković" Univerziteta u Beogradu, Instituta od nacionalnog značaja za Republiku Srbiju.

Ovom prilikom želim da izrazim zahvalnost:

Mom mentoru, dr Milanu Dragićeviću za uložen trud za osmišljavanje disertacije i planiranje eksperimenata, za sve savete, učenje novih stvari sa kojima se do sada nisam susretala, pre svega rad u R programskom jeziku i pomoć pri rešavanju svih problema tokom izrade disertacije. Hvala što je preuzeo odgovornost i pristao da bude mentor na polovini mojih doktorskih studija, što me je gurao napred kada sam gubila volju i što je najzaslužniji za to što je ova teza dobila svoj finalni oblik na način prikazan ovde.

Mentorki, dr Ivani Dragićević, za sve korisne savete i primedbe koje su doprinele kvalitetu disertacije, na izdvojenom vremenu i brzini kojom je to urađeno, kao i za saradnju i podršku tokom doktorskih studija.

Dr Tijani Cvetić Antić, na pomoći i vrednim savetima tokom izrade teze kao i na prijatnoj saradnji tokom doktorskih studija. Dr Marku Đorđeviću, na vremenu izdvojenom za pregled teze i savetima za njeno poboljšanje. Dr Mariji Marković, za lepu saradnju i korisne smernice prilikom uobličavanja teksta doktorske disertacije.

Dr Angelini Subotić, na pruženoj prilici da budem deo njenog projekta prethodnih godina, a dr Ani Simonović na budućoj saradnji kao deo njenog tima.

Posebnu zahvalnost dugujem koleginicama, Katarini Ćuković i njenoj mentorki dr Milici Bogdanović, na ustupljenim uzorcima cDNK koji su značajno doprineli poboljšanju rezultata ove teze kao i stvaranju celovitije slike o ulogama izabranih AGP gena.

Dr Milani Trifunović-Momčilov, na zajedničkoj saradnji tokom prvih godina mog rada na institutu i prvim koracima u radu sa in vitro kulturom i kičicom.

Koleginicama dr Jeleni Božunović i Dragani Antonić Reljin, posebnu zahvalnost za sve elektroforeze koje su uradile za potrebe ove teze kada sam bila u drugom stanju. Vanja vam je beskrajno zahvalan za to.

Dragani Antonić Reljin i dr Mariji Marković, za sve godine druženja, podrške, saveta. Bez vas bi bilo manje smeha..

Mojim prijateljima i koleginicama sa doktorskih studija Jeleni, Ani, Bojani, Jeleni, Aleksandri, Tamari hvala za sve godine iskrenog prijateljstva. Posebnu zahvalnost dugujem Jeleni Stanković, čiji su me mudri saveti u kritičnim trenucima izrade ove disertacije spasili.

Zahvalnost dugujem i svim svojim kolegama Odeljenja za Fiziologiju biljaka na kolegijalnosti i prijatnoj radnoj atmosferi.

Svojoj porodici, Dragi, Milanu, Ružici i Vladimiru, neizmerno hvala za podršku i ljubav svih ovih godina godina školovanja. Možete da odahnete, sada je kraj.

Milošu i Vanji, hvala što su tu i što svaki dan čine neizmerno lakšim.

Ovu tezu posvećujem dedi Mazalu i baba Milki, jer znam da će se tome najviše obradovati.

Identifikacija AGP gena kičice (*Centaurium erythraea*, Gentianaceae) i praćenje njihove ekspresije u odgovoru na mehaničke povrede biljnog tkiva gajenog *in vitro*

Rezime

Arabinogalaktanski proteini (AGP) su ekstenzivno glikozilovani proteini ćelijskog zida i predstavljaju podklasu O-glikozilovanih glikoproteina bogatih hidroksiprolinom (HRGP). Odlikuje ih visoka raznolikost primarne strukture, velika familija gena koja ih kodira, kao i širok spektar uloga tokom rastenja i razvića biljaka. AGP su proteini sa neuređenom strukturom što otežava njihovu identifikaciju na osnovu homologije sekvenci. Sa ciljem poboljšanja metodologije za identifikaciju i analizu HRGP sekvenci u sklopu ove doktorske disertacije razvijen je pristup zasnovan na mašinskom učenju. Ovaj pristup koristi glavnu odliku HRGP, a to je prisustvo nekarakteristične aminokiseline hidroksiprolina. Model za predviđanje verovatnoće hidroksilacije prolina na osnovu lokalne sekvence proteina inkorporiran je u *ragp* R paket uz brojne druge alate koji omogućavaju analize proteinskih sekvenci kao što su: klasifikacija HRGP sekvenci, fleksibilna pretraga karakterističnih motiva, efikasna komunikacija sa serverima za predviđanje N-terminalne signalne sekvence, mesta dodavanja glikozil-fosfatidilinozitolnog sidra, određivanje položaja neuređenih regiona i anotaciju domena u sekvencama. Navedena metodologija omogućava prilagodljivu identifikaciju sekvenci pri kojoj izbor nekoliko parametara utiče na rigoroznost procesa, a time i broj identifikovanih sekvenci. Navedeni softver omogućava efikasnu identifikaciju HRGP sekvenci u celim biljnim proteomima što predstavlja preduslov za njihovo dalje ispitivanje.

Kičicu (*Centaurium erythraea*) karakteriše velika razvojna plastičnost i snažan morfogenetski potencijal, što je čini pogodnim model organizmom za istraživanja u razvojnoj biologiji. AGP su identifikovani kao značajan biohemijski marker somatske embriogeneze i organogeneze kičice i kao takvi mogu biti značajni za proces regeneracije kako biljaka gajenih u uslovima *in vitro*, tako i onih u prirodi. Povrede predstavljaju sastavni deo manipulacije biljnim tkivom *in vitro* i mogu indukovati morfogenetske procese. Odgovor biljke indukovan Yariv reagensom, koji specifično precipituje AGP, najsličniji je odgovoru indukovanom mehaničkim povredama biljnog tkiva. Može se pretpostaviti da postoji veza između procesa koji se odigravaju u i na ćeliji, a koji su inicirani mehaničkim povredama ili primenom Yariv reagensa kao i tokom indukcije somatske embriogeneze i organogeneze kičice.

Nedavno sekvenciran transkriptom kičice uz razvoj *ragp* R paketa omogućio je identifikaciju velikog broja HRGP i AGP sekvenci kičice. Od identifikovanih sekvenci odabrano je osamnaest predstavnika za ispitivanje ekspresije u različitim uslovima. Ekspresija odabranih gena praćena je: a) 48 h nakon mehaničke povrede lista i korena biljaka kičice gajenih u uslovima *in vitro* b) u eksplantatima lista i korena kičice gajenih na različitim koncentracijama Yariv reagensa c) u različitim uzorcima biljaka gajenih *in vitro*, biljaka iz prirode i iz različitih morfogenetskih procesa kičice (somatske embriogeneze i organogeneze). Od odabranih osamnaest gena, *CeAGp6*, koji kodira kratak AG peptid, bio je izraženo indukovan nakon mehaničke povrede lista, a reprimiran nakon povrede korena. Moguće je da ovaj AG peptid uključen u određen tip prenosa signala nakon povrede biljnog tkiva. Pored njega *CeFLA1*, koji je pokazao povećanje ekspresije tokom 48 h nakon povrede lista, a indukovan je i u embriogenom kalusu, mogao bi biti deo mreže koja povezuje povrede i somatsku embriogenezu.

Ključne reči: hidroksiprolin, biljni glikoproteini bogati hidroksiprolinom, arabinogalaktanski proteini, mašinsko učenje, predviđanje posttranslacionih modifikacija proteina, HRGP, morfogeneza, Yariv reagens

Naučna oblast: Biologija

Uža naučna oblast: Fiziologija i molekularna biologija biljaka

Identification of *AGP* genes in centaury (*Centaurium erythraea*, Gentianaceae) and monitoring of their expression in response to tissue mechanical wounding *in vitro*

Abstract

Arabinogalactan proteins (AGPs) are extensively glycosylated cell wall proteins belonging to the super family of O-glycosylated hydroxyproline-rich glycoproteins (HRGPs). They are characterized by a high diversity of the primary structure, they are encoded by a large gene family and are associated with a wide array of physiological roles in plant growth and development. AGPs are intrinsically disordered proteins which hinders their homology-based identification. In order to improve the methodology for identification and analysis of HRGP sequences, a new machine learning based approach was developed as part of this thesis. This approach exploits the main feature of HRGPs, the presence of the uncharacteristic amino acid hydroxyproline. A model for predicting proline hydroxylation probability based on local protein sequence has been incorporated into the *ragp* R package along with a number of diverse tools that allow the analysis of protein sequences such as: classification of HRGPs sequences, flexible scan for characteristic motifs, efficient communication with web servers for prediction of N-terminal signal peptides, glycosylphosphatidylinositol modification sites, disordered regions and domain annotation. The implemented pipeline enables adaptable identification of sequences where the choice of several parameters affects the stringency of the process and thus the number of sequences identified. This software allows efficient identification of HRGPs sequences in whole plant proteomes, which is a prerequisite for their further study.

Common centaury *(Centaurium erythraea*) is characterized by an exceptional developmental plasticity and strong morphogenetic potential, which makes it a suitable model organism for developmental biology studies. AGPs have been identified as a significant factor during somatic embryogenesis and organogenesis of common centaury, and as such can be important for the regeneration process in both plants grown *in vitro* and plants from nature. Mechanical wounding is an integral part of plant tissue manipulation *in vitro* and can induce morphogenetic processes. The plant response induced by Yariv reagent, which specifically precipitates AGPs, is most similar to the response induced by mechanical wounding of plant tissue. It can be assumed that there is a connection between the processes that take place in and on the cell, which are initiated by mechanical wounding, application of Yariv reagents and during the induction of somatic embryogenesis and organogenesis.

The recently sequenced *C. erythraea* transcriptome combined with the development of *ragp* R package allowed the identification of a large number of *C. erythraea* HRGP and AGP sequences. Eighteen representatives were selected from the identified sequences for expression analyses in different experimental conditions. The expression of selected genes was recorded: a) 48h following leaf and root mechanical wounding *in vitro* b) in leaf and root explants cultivated on different concentrations of Yariv reagent in medium c) in samples from plants cultivated *in vitro*, plants from nature and from different morphogenetic processes (somatic embryogenesis i organogenesis). Among the selected eighteen genes, *CeAGp6*, which encodes a short AG peptide, was significantly induced after mechanical wounding of leaf, and down-regulated after mechanical wounding of root. It is possible that this AG peptide is involved in some type of signal transduction after plant tissue wounding. *CeFLA1* which showed a clear trend of increasing expression over time after leaf wounding, and is induced in embryogenic callus, could be part of a network linking wounding with somatic embryogenesis.

Keywords: hydroxyproline, plant hydroxyproline rich glycoproteins, arabinogalactan proteins, machine learning, prediction of posttranslational modifications, HRGP, morphogenesis, Yariv reagent

Scientific field: Biology

Scientific subfield: Plant physiology and molecular biology

Skraćenice:

2,4-D - 2,4-dihlorfenoksisirćetna kiselina

18s - 18s ribozomalna RNK

abl - uzorci adventivnih pupoljaka formirani na eksplantatima listova koji su gajeni u uslovima dugog dana, na hranljivoj podlozi bez dodatih regulatora rastenja

ablh - uzorci adventivnih pupoljaka formirani na eksplantatima listova koji su gajeni u uslovima dugog dana, na hranljivoj podlozi sa regulatorima rastenja (0,2 mgL^{-1} 2,4-D i 0,5 mgL^{-1} CPPU)

abr - uzorci adventivnih pupoljaka formirani na eksplantatima korenova koji su gajeni u uslovima dugog dana, na hranljivim podlogama bez regulatorima rastenja

ACC - metrika za merenje performansi modela MU za klasifikaciju. Udeo tačno predviđenih opservacija, tačnost (engl. „*accuracy*“)

AG - arabinogalaktan

AGP - arabinogalaktanski proteini

AGp - arabinogalaktanski peptidi

AK - adenozin kinaza

APS - amonijum persulfat

AUC - metrika za merenje performansi modela MU za klasifikaciju. Površina ispod ROC krive

BACC - metrika za merenje performansi modela MU za klasifikaciju. Balansirana tačnost (engl. „*balanced accuracy*“) prosek osetljivosti i specifičnosti

βGlcY - β-D-glukozil Yariv reagens

BR – brasinosteroidi

cDNK - komplementarni lanac DNK

PAC - prolinom bogati arabinogalaktanski protein bogat cisteinom (engl „*PRP-AGP containing Cys“)*

CDD - baza konzerviranih domena (engl. „*conserved domain database*“)

CL-EXT - unakrsno povezani ekstenzini (engl: „*cross linking* EXT“)

CPPU - *N*-fenil-*N*’-(2-hloro-4-piridil) urea

cse - uzorak kotiledonarnog embriona indukovan na eksplantatima listova na hranljivoj podlozi sa regulatorima rastenja (0,2 mgL^{-1} 2,4-D i 0,5 mgL^{-1} CPPU) i u kontinualnom mraku

Ct - prag ciklus (engl. „*cycle threshold*“)

CTAB - cetil trimetil amonium bromid

DEPC - dietilpirokarbonat

DNK - dezoksiribonukleinska kiselina

DUF - domen nepoznate funkcije (engl. „*domain of unknown function*“)

ec - uzorak embriogenog kalusa indukovan na eksplantatima listova na hranljivoj podlozi sa regulatorima rastenja (0,2 mgL^{-1} 2,4-D i 0,5 mgL^{-1} CPPU) i u kontinualnom mraku

EDTA - etilendiamintetrasirćetna kiselina

EF2 - elongacioni faktor 2

eNOD-AGP - AGP nalik ranom nodulinu (engl. „*early nodulin-like AGPs*“)

ER - endoplazmatični retikulum

EXT - ekstenzini

F1 - slikanje sekvence aminokiselina u sekvencu numeričkih vrednosti na osnovu fizičko-hemijskih svojstava aminokieselina: CIDH920105, BHAR880101, CHAM820102, BIGC670101, CHAM810101 i DAYM780201

F2 - slikanje sekvence aminokiselina u sekvencu numeričkih vrednosti na osnovu multidimenzionalnih obrazaca aminokiselina

F3 - Moreau-Broto autokorelacioni deskriptor baziran na standardizovanim fizičko-hemijskim atributima amino kieselina: CIDH920105, BHAR880101, CHAM820102, BIGC670101, CHAM810101 i DAYM780201

F4 - Moreau-Broto autokorelacioni deskriptor baziran na standardizovanim multidimenzionalnim obrascima aminokiselina

F5 - Moran autokorelacioni deskriptor baziran na standardizovanim fizičko-hemijskim svojstvima aminokieselina: CIDH920105, BHAR880101, CHAM820102, BIGC670101, CHAM810101 i DAYM780201

F6 - Moran autokorelacioni deskriptor baziran na standardizovanim multidimenzionalnim obrascima aminokiselina

F7 - Geary autokorelacioni deskriptor baziran na standardizovanim fizičko-hemijskim svojstvima aminokieselina: CIDH920105, BHAR880101, CHAM820102, BIGC670101, CHAM810101 i DAYM780201

F8 - Geary autokorelacioni deskriptor baziran na standardizovanim multidimenzionalnim obrascima aminokiselina

F9 - kuplujući broj redosleda sekvence

F10 - kvazi deskriptor redosleda sekvence

F11 - deskriptor združenih trijada

F12 - pseudo-aminokiselinski sastav

F13 - amfifilni pseudo-aminokiselinski sastav

F14 - kompozicioni deskriptor

F15 - tranzicioni deskriptor

F16 - distribucioni deskriptor

FAS - domen nalik fasciklinu

FH-EXT - ekstenzini sa forminskim domenom (engl. „*formin-homolog EXTs*“)

FLA - arabinogalaktanski proteini nalik fasciklinu (engl. „*fasciclin like arabinogalactan*“)

FN - metrika za merenje performansi modela MU za klasifikaciju. Broj opservacija koji pripadaju pozitivnoj klasi, a dodeljena im je negativna klasa; lažno negativni (engl. „*false negative*“)

FP - metrika za merenje performansi modela MU za klasifikaciju. Broj opservacija koji pripadaju negativnoj klasi, a dodeljena im je pozitivna klasa; lažno pozitivni (engl. „*false positive*“)

GA - Goldžijev aparat

GPI - glikozil-fosfatidilinozitol

GPI-sp - C-terminalna signalna sekvenca za dodatak glikozil-fosfatidilinozitola

gse – uzrak globularnog embriona indukovan na eksplantatima listova na hranljivoj podlozi sa regulatorima rastenja (0,2 mgL^{-1} 2,4-D i 0,5 mgL^{-1} CPPU) i u kontinualnom mraku

H3 - histon 3

HAE - hibrid arabinogalaktana i ekstenzina (engl. „*hybrid AGP-EXT*“)

HGRP - glikoproteini bogati hidroksiprolinom (engl. „*hydroxyproline-rich glycoproteins*“)

HPGT - hidroksiprolin O-galaktoziltransferaze

IDP - proteini neuređene strukture (engl. „*intrinsically disordered proteins*“)

IGr - udeo dobitka informacije (engl. „*information gain ratio*“)

imf - uzorak nezrelih, zatvorenih cvetova biljaka iz prirode

K – pozitivan ceo broj. Koristi se u nekoliko konteksta: hiperparamater algoritma najbližih suseda, broj aminokiselina u lokalnoj aminokiselinskoj sekvenci, broj podela unakrsne validacije.

KLA - arabinogalaktani nalik receptornim kinazama (engl. „*kinase like AGPs*“)

KNN - algoritam MU koji klasifikuje tačku posmatranja u odnosu na to kako su susedi klasifikovani, K najbližih suseda (engl. „*k-nearest neighbors*”)

ln - uzorak listova biljaka iz prirode

LRR - leucinom bogati regioni (engl. „*leucine rich regions*“)

MAAB - klasifikacija glikoproteina bogatih hidroksiprolinom prema aminokiselinskom sastavu i sadržaju motiva (engl. „*the motif and amino acid bias*“)

mf - uzorak zrelih, otvorenih cvetova biljaka iz prirode

mRMR - minimalna suvišnost maksimalna značajnost (engl. „*minimum redundancy maximum relevance*“)

MS - Murashige & Skoog hranljiva podloga

MU - mašinsko učenje

NJ - metoda pridruživanja suseda (engl. „*neighbour joining*“)

N-sp - N-terminalni signalni peptid

nsLTP-AGP - arabinogalaktani nalik nespecifičnim proteinima za transfer lipida (engl. „*non-specific lipid transfer protein-like AGPs*“)

oc - uzorak organogenog kalusa indukovan na eksplantatima listova na hranljivoj podlozi sa regulatorima rastenja (0,2 mgL^{-1} 2,4-D i 0,5 mgL^{-1} CPPU) u uslovima dugog dana

PCR - lančana reakcija polimeraze (engl. „*polymerase chain reaction*“)

PK - protein kinazni domen

PLA - arabinogalaktani nalik plastocijaninu (engl. „*plastocyanin-like*“)

PMEI - inhibitor pektin metil esteraze (engl. „*pectin methyl esterase inhibitor*“)

PRP - proteini bogati prolinom

PTK - protein tirozin kinazni domen

PTM - posttranslaciona modifikacija

PVPP - polivinil polipirolidon

qRT-PCR - kvantitativni RT-PCR

RAN - nuklearni protein homolog RAS (engl. „*rat oncogene sarcoma*“) virusnom proteinu

RBF - kernel sa radijalnom osnovom (engl. „*radial basis function*”)

rc - uzorci korenova iz kulture korenova *in vitro*

RF - ansambl algoritam MU koje se bazira na agregaciji predviđanja nezavisnih stabala odlučivanja (engl. „*random forests*“)

rl - uzorci listova biljaka gajenih u uslovima *in vitro* u fazi rozete

RMSD - koren srednje kvadratne greške (engl. „*root mean square deviation*“)

rn - uzorak korenova biljaka iz prirode

RNK - ribonukleinska kiselina

ROC - kriva zavisnosti udela stvarno pozitivnih opservacija (senzitivnost) od udela lažno pozitivnih opservacija (1 – specifičnost) (engl. „*receiver operating characteristic*“)

RPL2 - ribozomalni protein L2 (engl. „*ribosomal protein L2*“)

rr - uzorci korenova biljaka gajenih u uslovima *in vitro* u fazi rozete

RT - reakcija reverzne transkripcije (engl. „*reverse transcription”)*

sd - uzorci klijanaca biljaka gajenih u uslovima *in vitro*

SE - somatska embriogeneza (engl. „*somatic embryogenesis*“)

sfs - sekvenciona pretraga unapred (engl. „*sequential forward selection*“)

SO - organogeneza izdanaka (engl. „*shoot organogenesis*“)

SOD - superoksid dismutaza

st - uzorak stabla biljaka iz prirode

SVM - metod potpornih vektora (engl. „*support vector machine*“)

T-DNA - deo DNK sekvence tumor indukujućeg plazmida bakterija koji se ugrađuje u genom domaćina (engl. „*transfer DNA*“)

TBP1 – protein koji se vezuje za TATA sekvencu, konsenzus sekvencu u promotoru koju prepoznaje transkripcioni faktor

TDZ – tidiazuron, 1-fenil-3-tidiazol-5-urea

TEMED - N,N,N',N'-tetrametiletan-1,2-diamin

TN - metrika za merenje performansi modela MU za klasifikaciju. Broj opservacija koji pripadaju negativnoj klasi, a dodeljena im je negativna klasa; stvarno negativni (engl. „*true negative*“)

TNR - metrika za merenje performansi modela MU za klasifikaciju. Udeo stvarno negativnih opservacija, specifičnost (engl. „*true negative rate*“)

TP - metrika za merenje performansi modela MU za klasifikaciju. Broj opservacija koji pripadaju pozitivnoj klasi, a dodeljena im je pozitivna klasa; stvarno pozitivni (engl. „*true positive*“)

TPR - metrika za merenje performansi modela MU za klasifikaciju. Udeo stvarno pozitivnih opservacija; osetljivost ili senzitivnost (engl. „*true postive rate*“)

TRIS - tris-hidroksimetil aminometan

TUA - α-tubulin

TUB - β-tubulin

XGB – efikasna implementacija Gradient Boost algoritma, ansambl algoritma MU koje se bazira na pojačavanju stabala odlučivanja (engl. „*extreme gradient boosting*”)

Skraćenice aminokiselina:

troslovna skraćenica	jednoslovna skraćenica	naziv
Ala	A	Alanin
Arg	R	Arginin
Asn	N	Asparagin
Asp	D	Aspartat
Cys	C	Cistein
Glu	E	Glutamat
Gln	Q	Glutamin
Gly	G	Glicin
Hyp	O	Hidroksiprolin
His	H	Histidin
Ile	I	Izoleucin
Leu	L	Leucin
Lys	K	Lizin
Met	M	Metionin
Phe	F	Fenilalanin
Pro	P	Prolin
Ser	S	Serin
Thr	T	Treonin
Trp	W	Triptofan
Tyr	Y	Tirozin
Val	V	Valin

Sadržaj

1. Uvod

1.1. Glikoproteini bogati hidroksiprolinom

Ćelijski zid je dinamična struktura koja predstavlja strukturnu potporu biljne ćelije i deluje kao prva linija odbrane od različitih faktora abiotičkog i biotičkog stresa. To je kompozitna struktura uglavnom sačinjena od polisaharida celuloze, hemiceluloza i pektina (Hijazi i sar., 2014), svaki sa specifičnim ulogama. Unutar ove isprepletane mreže polisaharida umetnuti su biljni glikoproteini bogati hidroksiprolinom (HRGP, engl. „*hydroxyproline-rich glycoproteins*“) (Lamport i Northcote, 1960) koji čine skoro 10% suve mase ćelijskog zida (Nguema-Ona i sar., 2013). HRGP imaju širok spektar funkcija u rastenju i razviću, odgovoru biljke na uslove sredine, signalizaciji i odbrani biljke od spoljašnjih faktora (Deepak i sar., 2010; Hijazi i sar., 2014; Kieliszewski i sar., 2010).

Kao što im naziv sugeriše, HRGP su karakteristični po visokom sadržaju modifikovane amino kiseline 4-hidroksiprolina (Hyp, O) koja nastaje hidroksilacijom prolina dejstvom enzima prolil 4-hidroksilaze tokom sazrevanja proteina. Za razliku od većine drugih posttranslacionih modifikacija (PTM) hidroksilacija prolina je ireverzibilna i prisutna kod životinja, biljaka i bakterija (Gorres i Raines, 2010). U biljnim HRGP ova hidroksilna grupa služi kao mesto dodatka različitih tipova glikozida O-glikozilacijom (Ellis i sar., 2010).

Zbog visokog sadržaja polarnih i naelektrisanih aminokiselina koje promovišu neuređenost, a pre svega prolina/hidroksiprolina koji je, zbog svoje rigidne konformacije nepogodan za najfrekventnije sekundarne strukture proteina kao što su alfa-heliks i beta-ravan, proteinski deo HRGP je neuređene strukture (IDP, engl. „*intrinsically disordered proteins*“). Neuređenost je takođe promovisana glikozilacijom (Johnson i sar., 2017a). Ovo praktično znači da HRGP nemaju jedinstvenu globularnu strukturu sa hidrofobnom unutrašnjošću. Zbog svojih brojnih fizioloških uloga i nedostatka stabilne strukture ovi proteini praktično narušavaju biohemijsku dogmu da je uloga proteina određena njegovom trodimenzionom strukturom odnosno konformacijom polipeptidne kičme.

Pošto su proteini ćelijskog zida, HRGP na N-terminusu imaju signalnu sekvencu (N-terminalni signalni peptid, N-sp) koja ih usmerava ka sistemu endomembrana endoplazmatičnog retikuluma (ER) i Goldžijevog aparata (GA) i na kraju do ekskrecije kroz ćelijsku membranu. Tokom sazrevanja proteina ova signalna sekvence se odseca. Određene grupe HRGP su pričvršćene za ćelijsku membranu pomoću glikolipidnog sidra modifikacijom C-terminusa sa glikozil-fosfatidilinozitolom (GPI, Seifert i Roberts, 2007). Dodatak GPI je uslovljen specifičnom C-terminalnom signalnom sekvencom (GPI-sp) koja se u reakciji transamidacije menja sa GPI, a aminokiselina na koju se dodaje GPI sidro se označava kao omega (ω) mesto (Kinoshita i Fujita, 2016; Galian i sar., 2012).

Zavisno od tipa i stepena glikozilacije, HRGP se mogu podeliti na visoko glikozilovane arabinogalaktanske proteine (AGP), umereno glikozilovane ekstenzine (EXT) i slabo glikozilovane proteine bogate prolinom (PRP) (Showalter i sar., 2010; Hijazi i sar., 2014). Prema hipotezi kontinuiteta hidroksiprolina, tip glikozilacije HRGP je moguće predvideti na osnovu Hyp-karakterističnih motiva u sekvenci (Ellis i sar., 2010; Tan i sar., 2012; Johnson i sar., 2017a). Regioni sa više kontinualnih Hyp karakteristični za EXT, su O-glikozilovani linearnim oligoarabinozidima. PRP imaju relativno nizak stepen glikozilacije arabinozom i kratkim linearnim arabinozidima na specifičnim aminokiselinskim motivima bogatim sa aminokiselinama V i K. Za AGP su karakteristični diksontinualni Hyp grupisani u dipeptide sa aminokiselinama A, S i T koji su O-glikozilovani razgranatim oligo- i polisaharidima (Johnson i sar., 2017a).

1.1.1. Ekstenzini

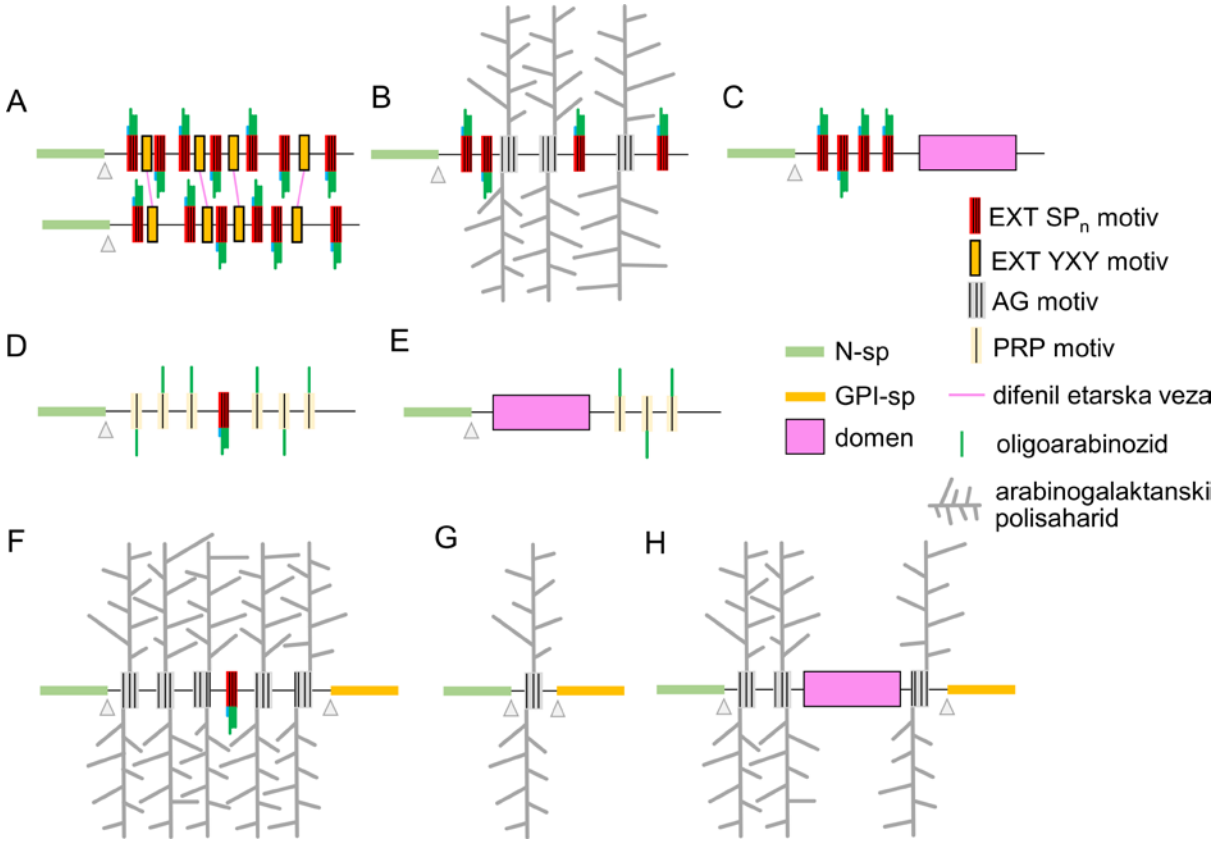
Ekstenzini su grupa umereno glikozilovanih proteina kod kojih šećeri čine oko polovine molekulske mase. Proteinski deo je prožet ponovljenim motivima SO_n (n je najčešće 3-5), a često sadrže ditirozinske YXY hidrofobne motive (Showalter i sar., 2010). Pomenuti ditirozinski motivi učestvuju u intramolekulskom, a moguće i intermolekulskom, kovalentnom umrežavanju formiranjem difenil etarskih veza (izoditirozina) (Slika 1A) koje povećavaju rigidnost konformacije EXT (Kieliszewski i Lamport, 1994; Showalter i sar., 2010; Held i sar., 2004). Kao što je napomenuto u odeljku 1.1. O-glikozilacija EXT se pretežno odigrava na Ser-Hyp_n motivima. Hyp u ovim motivima su O-glikozilovani kratkim linearnim oligoarabinozidima koji se najčešće sastoje od četiri do pet monosaharidnih jedinica, dok je Ser koji im prethodi O-glikozilovan monosaharidom galaktozom (Kieliszewski i Lamport, 1994; Ogawa-Ohnishi i sar., 2013). Pokazano je da je O-glikozilacija EXT esencijalna za sastavljanje ćelijskog zida i elongaciju korenskih dlaka *Arabidopsis thaliana* (Velasquez i sar., 2011).

EXT se prema literaturi dele na nekoliko klasa (Showalter i sar., 2010): klasični EXT, kratki EXT, hibridni EXT, i nekoliko klasa himernih ekstenzina (Slika 1):

- Klasični EXT (Slika 1A, CL-EXT, engl. „*cross linking*“) su visokorepetitivne sekvence koje sadrže kratke hidrofilne blokove O-glikozilovanih $SO_{(3–5)}$ motiva i hidrofobne YXY motive uključene u umrežavanje EXT (Cannon i sar., 2008).
- Kratki EXT imaju manje od 200 aminokiselina i mogu imati i GPI-sidro (Showalter i sar., 2010), što nije slučaj sa ostalim klasama EXT.
- Hibridni EXT (Slika 1B) imaju motive karakteristične za AGP (retko PRP) i EXT motive. Četiri AGP-EXT hibrida (HAE, engl. „*hybrid AGP-EXT*") identifikovana su kod *Arabidopsis thaliana* (Showalter i sar., 2010) i svi su asocirani sa pojedinim domenima (himerni HRGP): HAE1 ima PMEI domen (*Pfam*, engl „*Protein family*“: PF04043, engl. „*pectin methyl esterase inhibitor*“), HAE4 ima domen za transfer biljnih lipida (PF00234, engl. „*plant lipid transfer domain*“), dok HAE2 i HAE3 imaju domene nepoznate funkcije (DUF, engl. „*domain of unknown function*“).
- Himerni EXT (Slika 1C) imaju regione bogate karakterističnim motivima za EXT i jedan ili više specifičnih konzerviranih domena (Showalter i sar., 2010; Liu i sar., 2016) od kojih su najpoznatiji:
  - EXT sa leucinom bogatim ponovcima (LRX, engl. „*leucine rich repeat extensins*“) najčešće pri N-terminusu imaju leucinom-bogate regione (LRR, PF00560, engl. „*leucine rich regions*”), dok su regioni sa EXT motivima lokalizovani ka C-terminusu. Smatra se da su LRR uključeni u protein-protein interakcije (Kobe i Deisenhofer, 1994), što ih čini kandidatima za regulatorne funkcije na površini ćelije. Pokazano je da su kod *A. thaliana* LRX uključeni u morfogenezu korenskih dlačica (Baumberger i sar., 2001). Kod *A. thaliana* je otkriveno 11 LRR-EXT, od čega je sedam eksprimirano u vegetativnim, a četiri u reproduktivnim tkivima (Stratford i sar., 2001).
  - Prolinom bogate receptor kinaze nalik ekstenzinima (PERK, engl. „*proline-rich extensin-like receptor kinases*“) predstavljaju transmembranske himerne EXT kod kojih se SP_n ponovci nalaze u ekstracelularnom delu ka N-terminusu, dok je receptor kinazni domen (PF07714 i PF00069) na intracelularnom C-terminusu. Pokazano je da postoje dve velike grupe PERK kod *A. thaliana*: specifično eksprimirani u polenu i oni koji se eksprimiraju u svim tkivima (Nakhamchik i sar., 2004). PERK su povezani sa odgovorom *A. thaliana* na abscisinsku kiselinu (Bai i sar., 2009), a BnPERK1 iz *Brassica napus* je uključen u percepciju signala i odgovor na povrede i patogene (Silva i Goring, 2002).
  - EXT sa forminskim domenom (FH-EXT, engl. „*formin-homolog EXTs*“) sadrže konzervirani FH2 domen (PF02181) pored EXT motiva. FH2 domen je kod biljaka uključen u interakciju između aktina i mikrotubula citoskeleta (Chalkia i sar., 2008;

Borassi i sar., 2016). Članovi ove familije su neophodni za pravilan rast korenskih dlačica i izduživanje polenove cevi (Huang i sar., 2013; Cheung i sar., 2010)

EXT povećavaju mehaničku čvrstoću ćelijskog zida formiranjem izoditirozina kroz reakciju katalizovanu peroksidazama. Usled formiranja velikog broja ditirozinskih mostova koji grade ekstenzini prilikom akumulacije u ćelijskom zidu, ćelijski zid postaje rigidan, čime se ograničava njegovo rastenje. EXT imaju važnu ulogu u obezbeđivanju zaštite ćelijskog zida biljne ćelije, što zahteva njihovu pravilnu glikozilaciju (Castilleux i sar., 2021). Pokazano je da je arabinozilacija EXT neophodna za odbranu tokom ranih faza infekcije korena *A. thaliana* sa *Phytophthora parasitica* (Castilleux i sar., 2019). Ekstenzinska mreža asocirana sa polisaharidima ćelijskog zida, kao što su pektini i AGP, ojačava ćelijski zid i ograničava kolonizaciju patogena (Tan i sar., 2020).



Slika 1. Šematski prikaz struktura članova HRGP familije: A - CL-EXT, B – hibridni EXT/AGP, C – himerni EXT, D – PRP, E – himerni PRP, F – AGP, G – AG peptid, H – himerni AGP. Hyp u motivima su predstavljeni crnim vertikalnim linijama, AG motivi su predstavljeni sivom pozadinom, EXT SP_n motivi su predstavljeni crvenom pozadinom, EXT Tyr motivi su predstavljeni narandžastim pravougaonicima, a difenil etarske veze između njih sa ljubičastim linijama, PRP motivi su predstavljeni krem pravougaonicima. Himerni HRGP poseduju *Pfam* domene (ljubičasti pravougaonici). N-sp je predstavljen zelenim segmentom na N-terminusu, a GPI-sp žutim segmentom na C-terminusu. Sive strelice označavaju da se ovi signalni peptidi odstranjuju. O-glikozilovani oligo i polisaharidi su šematski predstavljeni na svim klasama HRGP (modifikovano prema Johnson-u i sar. (2017a)).

1.1.2. Proteini bogati prolinom

Proteini bogati prolinom (PRP) su najmanje glikozilovani članovi HRGP familije koji sadrže 2-27% šećera (Shpak i sar., 1999). Za PRP su karakteristični aminokiselinski motivi PPVX[KT], KKPCPP i KPP[VHTQA][YTPE]KPP (Sommer-Knudsen i sar., 1998). Određen udeo Pro u ovim

motivima su hidroksilovani i O-glikozilovani arabinozom i oligoarabinozidima (Showalter i sar., 2016).

PRP se u literaturi najčešće dele na klasične (Slika 1D), himerne PRP (Slika 1E) i kratke PR peptide (Showalter i sar., 2010). Klasični PRP se identifikuju kao sekvence koje sadrže 45% ili više PVKCYT, sa dva ili više KKPCPP ili PVX[KT] ponovaka (Showalter i sar., 2010). Himerni PRP sadrže PRP karakteristične regione asocirane sa određenim konzerviranim domenima kao što je cistein bogati C-terminalni PAC (engl. „*PRP-AGP containing Cys*“) domen (Hijazi i sar., 2012).

Tačna funkcija PRP nije utvrđena ali se smatra da su uključeni u procese rastenja i razvića, tokom formiranja nodula, ekspanzije ćelija (Showalter, 1993; Dvoráková i sar., 2012) i u odbrani od biotičkog i abiotičkog stresa, posebno kod ćelija koje su u fazi rastenja (Battaglia i sar., 2007; Hijazi i sar., 2014).

1.1.3. Arabinogalaktanski proteini

Među različitim grupama HRGP, AGP privlače najviše pažnje pre svega zbog strukturne raznovrsnosti i uloga u različitim fiziološkim procesima (Ellis i sar., 2010). Pronađeni su kod algi i svih biljaka, od mahovina (Lee i sar., 2005) do skrivenosemenica (Majewska i Nothnagel, 2000; Schultz i sar., 2000).

Arabinogalaktanski proteini su proteoglikani ćelijskog zida biljaka kod kojih proteinski deo čini do 10% molekulske mase. Za proteinski deo AGP su karakteristični arabinogalaktanski (AG) motivi ili glikomoduli koji se sastoje od diskontinualnih Hyp grupisanih u dipeptide sa aminokiselinama A, S, T i ređe G i V (npr: AO, SO, TO, OT, OG, OV i drugi). AG motivi su O-glikozilovani na Hyp sa razgranatim arabino-3,6-galaktanskim polisaharidima tipa II (Slika 2, Johnson i sar., 2017a) koji se sastoje od (1→3)-β-D- galaktana sa (1→6)-β-D- galaktanskim bočnim lancima koji su supstituisani sa arabinozom i u manjoj meri sa L-ramnozom, L-fukozom, D-manozom, D-ksilozom, D-glukozom, D-glukozaminom, D-glukuronskom i D-galakturonskom kiselinom (Showalter i sar., 2001).

Podela AGP u sedam grupa izvršena je na osnovu primarne strukture (Gaspar i sar. 2001) i to su:

- Klasični AGP (Slika 1F) čine proteinske sekvence dužine preko 100 aminokiselina, koje se sastoje od N-sp na N-terminusu posle kog je centralni region varijabilne dužine bogat Pro, Ala, Ser i Thr aminokiselinama (PAST) koje su aranžirane u AG motive, dok se na C-terminusu nalazi GPI-sp.
- AG peptidi (Slika 1G) su po građi slični klasičnim AGP sa tom razlikom što su to kratke sekvence koje nakon odsecanja N-sp i GPI-sp imaju 10-15 aminokiselina.
- AGP bogati lizinom poseduju kratak bazni region bogat Lys (Zhang i sar., 2011). Kod *A. thaliana* su otkrivena dva AGP koji pripadaju ovoj grupi, AtAGP17/18 koji su lokalizovani na ćelijskoj membrani i uključeni u signalnu transdukciju.
- AGP nalik fasciklinu (FLA, engl. „*fasciclin like arabinogalactan*“) su himerni AGP koji su asocirani sa jednim ili dva fasciklinska domena (FAS, PF02469) duga oko 110-150 aminokiselina. FLA se razlikuju na osnovu broja FAS domena i AGP regiona (mogu sadržati jedan ili dva) i prisustva GPI-sidra (Johnson i sar., 2003; Ma i Zhao, 2010). Pokazano je da proteini sa FLA domenima imaju adhezivnu ulogu (Kim i sar., 2000; Kawamoto i sar., 1998) i da presudnu ulogu u njihovoj funkciji imaju dva visoko konzervirana regiona od oko 10 aminokiselina označena kao H1 i H2 (Kawamoto i sar., 1998).

- AGP nalik ranom nodulinu (eNOD-AGP, engl. „*early nodulin-like AGPs*“) (Mashiguchi i sar., 2008) i AGP nalik plastocijaninu (PLA, engl. „*plastocyanin-like*“) predstavljaju himerne AGP asocirane sa PLA domenom (PF02298) (Ma i sar., 2011).
- AGP nalik nespecifičnim proteinima za transfer lipida (nsLTP-AGP, engl. „*non-specific lipid transfer protein-like AGPs*“) su himerni AGP asocirani sa ns-LTP domenom (PF00234, engl. „*plant lipid transfer/seed storage/trypsin-alpha amylase inhibitor*”). ns-LTP sadrži osam Cys koji grade 4 disulfidna mosta (Motose i sar., 2004). Ksilogen je primer proteina koji sadrži AG regione i nsLTP domen (Kobayashi i sar., 2011).
- Neklasični AGP su velika grupa svih AGP koji se ne mogu svrstati u prethodno opisanih šest klasa. Za njih je karakteristično da najčešće nemaju signal za dodavanje GPI sidra na C-terminusu (Gaspar i sar., 2001).

Nove grupe AGP se i dalje otkrivaju, pri čemu mnoge sekvence imaju odlike različitih klasa tako da se pre može govoriti o kontinuumu sekvenci, ne samo AGP već i ostalih HRGP klasa.

1.1.3.1. Biosinteza arabinogalaktanskih proteina

Biosinteza AGP počinje translacijom N-terminalne signalne sekvence AGP u ribozomima, što im omogućava ulazak u endoplazmatični retikulum (ER) i sistem endomembrana. Tokom sazrevanja polipeptidni lanci AGP podležu brojnim modifikacijama od posttranslacionih modifikacija kao što su hidroksilacija prolina u okviru AG motiva i njihova glikozilacija, do uklanjanja sekretorne signalne sekvence na N-terminusu i dodatka GPI sidra.

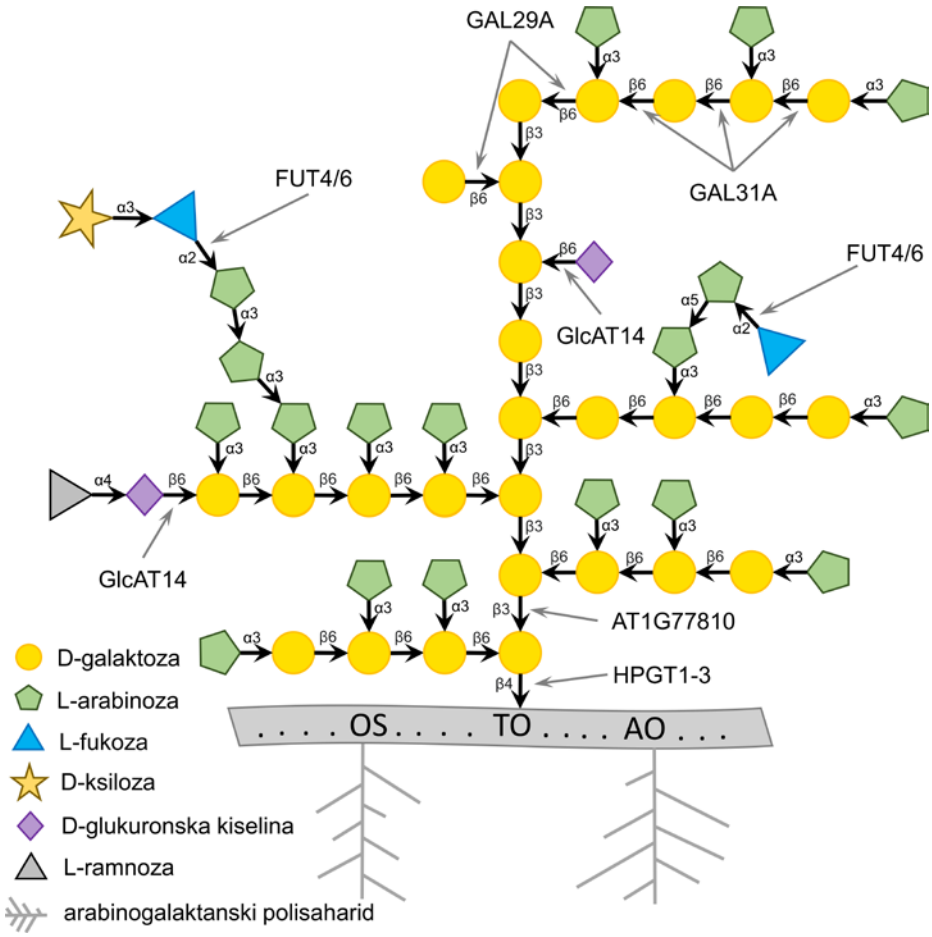
Hidroksilacija prolina se dešava u ER i katalisana je prolil-4-hidroksilazama, kodiranim multigenskom familijom (Walter i sar., 1994). Nekoliko biljnih prolil hidroksilaza je okarakterisano do sada (Velasquez i sar., 2015) i pokazano je da one najčešće hidroksiluju Pro koji se nalaze posle A, Q, O, P, S, T i V (Canut i sar., 2016). Na ovaj način se obezbeđuju reaktivne hidroksilne grupe koje su preduslov za O-glikozilaciju (Faye i sar., 2005; Hijazi i sar., 2014).

Propeptidima koji sadrže hidrofobni signal za dodavanje GPI sekvence na C-terminusu, dodaje se GPI sidro koje ih pričvršćuje za unutrašnju stranu ER membrane (Schultz i sar., 1998). Od ER peptidi se vezikulama transportuju do GA gde se odigravaju dalje PTM, a nakon toga do površine ćelije gde ih GPI-sidro pričvršćuje za površinu ćelije i omogućava njihovu lateralnu mobilnost u membrani, a može imati funkciju u signalnoj transdukciji (Schultz i sar., 1998; Ellis i sar., 2010). Peptidni signal za GPI-sidro na C-terminusu proteina sastoji se najčešće od 8 do 12 polarnih aminokiselina i hidrofobnog regiona promenljive dužine (najčešće od 9 do 24 aminokiselina) (Schultz i sar., 1998; Eisenhaber i sar., 2003; Ellis i sar., 2010). U polarnom regionu GPI-sp prve četiri aminokiseline najčešće imaju kratak bočni niz, i za jednu od njih (Gly, Ala, Cys, Ser, Asp ili Asn) se pričvršćuje GPI (ω mesto). Signal za GPI-sidro biva prepoznat od strane kompleksa GPI transamidaze koja u reakciji transamidacije kači GPI-sidro za ω aminokiselinu istovremeno odsecajući preostalu C-terminalnu signalnu sekvencu (Desnoyer i sar., 2020; Ellis i sar., 2010). Kod *A. thaliana*, od 85 identifikovanih AGP, za 55 je predviđeno da sadrže GPI-sidro (Showalter i sar., 2010).

Glikozilacija predstavlja najznačajniju posttranslacionu modifikaciju u živim organizmima, jer može uticati na fizičko-hemijske osobine samog proteina, uključujući otpornost na denaturaciju temperaturom, zaštitu od proteolitičke degradacije, rastvorljivost i biološke funkcije (Faye i sar., 2015). U zavisnosti od atoma za koji se glikozid pričvršćuje glikozilacija proteina može biti N- ili O-tipa. N-glikozilacija se odigrava na Asn u konzerviranim motivima Asn-X-Ser/Thr, gde X može biti bilo koja aminokiselina sem Pro (Faye i sar., 2015). Smatra se da klasični AGP nemaju motive za N-glikozilaciju, dok mnogi himerni, kao na primer FLA, imaju (Ellis i sar., 2010). O-glikozilacija se odigrava na aminokiselinama Ser, Thr i Hyp, pri čemu kod biljaka po frekvenciji prednjači O-

glikozilacija na Hyp (Duruflé i sar., 2017). Kod biljaka se galaktoza može vezati za Ser i Hyp, dok se arabinoza može vezati samo za Hyp (Duruflé i sar., 2017).

Glikozilacija AGP (Slika 2) počinje u ER dodavanjem molekula galaktoze na Hyp u reakciji katalisanoj Hyp-O-galaktoziltransferaza iz GT31 familije glikozil transferaza (Oka i sar., 2009). Kod *A. thaliana* su izolovane tri Hyp O-galaktoziltransferaze (HPGT1, HPGT2 i HPGT3, Slika 2) koje O-glikoziluju Hyp sa galaktozom, a pokazano je da je oko 90% aktivnosti svih endogenih O-galaktoziltransferaza rezultat aktivnosti ova tri enzima (Ogawa-Ohnishi i Matsubayashi, 2015). Glikozilacija se nastavlja u GA gde se odigrava nadograđivanje glikozida dejstvom različitih glikoziltransferaza (Showalter i Basu, 2016). Identifikovani su brojni članovi GT29, GT31 i GT14 familija glikozil transferaza uključeni u biosintezu AGP glikana (Qu i sar., 2008; Knoch i sar., 2013; Dilokpimol i Geshi, 2014). Galaktoziltransferaze, koje pripadaju GT31 i GT29 familijama, su uključene u sintezu AG glikozida dodavanjem galaktoze β-(1,3) i β-(1,6) vezama. Kod *A. thaliana* je okarakterisana jedna GT31 galaktoziltransferaza (At1g77810) koja je uključena u sintezu β-(1,3) galaktanskog lanca i pokazano je da je primarno lokalizovana u GA (Qu i sar., 2008). Pored nje, identifikovane su β-(1,6)-galaktoziltransferaze, GALT31A koja pripada GT31 famliji i uključena je ekstenziju linearnih β-(1,6) galaktanskih lanaca (Geshi i sar., 2013) i GAL29A, koja pripadala GT29 familiji, verovatno uključena u dodatak bočne β-(1,6) galaktoze na β-(1,3) galaktanski niz, na taj način formirajući račvanje (Dilokpimol i Geshi, 2014). Dodatak terminalnih uronskih kiselina na β-(1,6)- i β-(1,3)-galaktanske lance katalisan je GT14 familijom enzima, a kod *A. thaliana* su identifikovane tri enzima GlcAT14A, GlcAT14B i GlcAT14C (Knoch i sar., 2013). Terminalna fukoza se kod *A. thaliana* dodaje delovanjem α-fukoziltransferaza AtFUT4 i AtFUT6 iz GT37 familije glikozil transferaza (Tryfona i sar., 2014).



Slika 2. Biosinteza AGP kod *A. thaliana*. Šematska struktura arabinogalaktanskog polisaharida sa prikazanim mestima delovanja poznatih glikoziltransferaza. HPGT1-3 su Hyp O-

galaktoziltransferaze koje glikoziluju hidroksilnu grupu Hyp sa galaktozom. At1g77810 je uključen u sintezu β-(1,3) galaktanskog lanca. GAL29A i GALT31A katališu formiranje i ekstenziju β-(1,6) galaktanskih lanaca. GlcAT14 katališe dodatak terminalnih uronskih kiselina. AtFUT4/6 katališu dodatak terminalne fukoze (modifikovano prema Showalter i Basu (2016)).

Nakon glikozilacije, AGP se sekretornim vezikulama transportuju do ćelijskog zida, gde naknadno proteolitičko sečenje može da odvoji GPI sidro i oslobodi AGP, u slučaju GPI-usidrenih AGP (Schultz i sar., 1998).

1.1.3.2. Fiziološke uloge AGP

Izuzetna raznovrsnost primarne strukture peptidnog lanca AGP i činjenica da su kodirani velikom familijom gena, čine ovu klasu proteina podobnim za širok spektar uloga u regulaciji procesa rastenja i razvića biljaka (Coimbra i sar., 2009). Pokazano je da AGP učestvuju u procesima ćelijske proliferacije, embriogeneze, ćelijskog rastenja, razmnožavanja, diferencijacije ksilema, programirane ćelijske smrti, odgovora na delovanje hormona i opadanja listova (Serpe i Nothangel, 1994; Gao i Showalter, 1999; Ellis i sar., 2010; Seifert i Roberts, 2007; Majewska-Sawka i Nothnagel, 2000; Pereira i sar., 2015; Costa i sar., 2019). Postoje dokazi da AGP učestvuju u izduživanju polenove cevi (Lee i sar., 2008) i formiranju polena (Coimbra i sar., 2009; Li i sar., 2010), kao i u kontroli procesa zigotske embriogeneze, gametogeneze i androgeneze (Tang i sar., 2006). Veliki udeo AGP pričvršćen je za ćelijsku membranu preko GPI sidra, povezujući ih sa elementima citoskeleta i omogućavajući im da budu uključeni u obezbeđivanje oblika ćelije, polarizaciju i signalizaciju (Seifert i Roberts, 2007; Ellis i sar., 2010). Određeni GPI-AGP su asocirani sa receptornim kinazama pa se smatra da imaju ulogu u prenosu signala (Zhou, 2019). Pokazano je da su AGP uključeni u osmoregulaciju (Lamport i sar., 2006), u odgovor na različite tipove stresa kao što su niske temperature (Meng i sar., 2020), mehaničke povrede biljnog tkiva (Gilson i sar., 2001; Liu i sar., 2007; Fragkostefanakis i sar., 2012), interakcije sa mikroorganizmima (Nguema-Ona i sar., 2013) i odgovor na infekciju biljnog tkiva (Gaspar i sar., 2004; Xie i sar., 2011). Tokom interakcije biljke i patogena, proizvodi degradacije AGP od strane hidrolitičkih enzima patogena indukuju odbrambeni odgovor biljke (Villa-Rivera i sar., 2021). AGP su dobar model sistem za proučavanje adaptacija ćelijskog zida na različite uslove sredine zahvaljujući modifikacijama glikana identifikovanim u različitim tkivima koje menjaju funkciju AGP (Pfeifer i sar., 2020). Asocijacija sa konzerviranim domenima omogućava različitim klasama AGP još širi spektar funkcija. Tako je pokazano da FLA imaju uticaj na organizaciju polisaharida ćelijskog zida kao što su celuloza i pektini što utiče na svojstva ćelijskog zida i rastenje biljaka (MacMillan i sar., 2010).

Fiziološke uloge AGP se najčešće proučavaju korišćenjem β-glikozil Yariv reagenasa (Yariv i sar., 1967) i monoklonalnih antitela. β-glikozil Yariv reagensi su sintetička fenilazo glikozidna jedinjenja koja specifično vezuju AGP nekovalentnom interakcijom sa ugljenim hidratima i precipitiraju ih (Kitazawa i sar., 2013). Od β-glikozil Yariv reagenasa najčešće upotrebljavan za ispitivanje AGP je β-glukozil Yariv (βGlcY) koji najefikasnije precipitira AGP. βGlcY se koristi u histohemijskim bojenjima kojima se vizuelizuje distribucija AGP u tkivu, u radijalnoj imunodifuziji i elektroimunoesejima (imuno elektroforeze) za vizuelizaciju i kvantifikaciju AGP (Trifunović i sar., 2014). Osim toga pokazano je da dodatak βGlcY biljkama ili biljnom tkivu inhibira razvoj klijanaca, blokira somatsku embriogenezu i biosintezu polimera ćelijskog zida (Seifert i Roberts, 2007). Pretpostavlja se da vezivanjem i precipitacijom AGP u ćelijskom zidu, βGlcY dovodi do inaktivacije AGP što predstavlja jedan od načina za ispitivanje njihove biološke funkcije. Brojni su radovi koji ispituju interakcije Yariv reagensa i AGP (Paulsen i sar., 2014; Tang i sar., 2006; Hu i sar., 2006; Maurer i sar., 2010; Mashigushi i sar., 2008). Monoklonalna antitela specifična za određene epitope na ugljenim hidratima se rutinski koriste u histološkoj vizuelizaciji distribucije različitih AGP epitopa (Moller i sar., 2008). Obe pomenute metodologije omogućavaju procene fizioloških uloga cele

heterogene familije AGP, ali ne i individualnih AGP. U poslednjoj deceniji se za istraživanje specifičnih funkcija pojedinih AGP sve više koriste mutanti u kojima je ekspresija ispitivanih AGP sprečena. Ovakav pristup je inicijalno započet na model biljci *A. thaliana,* pre svega zahvaljujući velikim bibliotekama T-DNK insercionih mutanata kao što su SALK (Sessions i sar., 2002), GABI-KAT (Rosso i sar., 2003), SAIL i WISC (Alonso i sar., 2003; Woody i sar., 2007), dok sve više postaje dostupan i za istraživanje uloge određenih gena i kod drugih biljnih vrsta zahvaljujući razvoju metoda kao što su RNAi (Matthew, 2004) i CRISPR-Cas (Zhang i Showalter, 2020). Koristeći pomenute pristupe pokazano je da AtFLA3 iz *A. thaliana* učestvuje u depoziciji celuloze i da njegov nedostatak dovodi do abnormalnog razvoja polena (Li i sar., 2010), kao i da je AtAGP30, neophodan za regeneraciju korena u uslovima *in vitro* (van Hengel i sar., 2002). Studije rađene na mutantima *agp6* i *agp11 A. thaliana* pokazale su da ovi AGP imaju ulogu u sprečavanju nekontrolisane proizvodnje polenovih zrna i normalnom izduživanju polenove cevi (Coimbra i sar., 2009; Suzuki i sar., 2017).

1.1.3.3. Uloga AGP u razviću i odgovoru biljke na mehaničke povrede

Odgovori biljaka na povrede su intenzivno proučavani, pošto su biljke u prirodi konstantno izložene nekom vidu ovakvog stresa iz spoljašnje sredine (vetar, kiša, herbivori). Povrede nastale mehaničkim putem su potencijalna mesta infekcije patogena, a ekspresija odbrambenih gena na mestu povrede je jedan vid zaštitne barijere biljke prema mikroorganizmima. S obzirom da su AGP proteini ćelijskog zida, do promene u njihovoj ekspresiji dolazi prilikom mehaničkih povreda i napada patogena na biljku (Fragkostefanis i sar., 2012; Kjellbom i sar., 1997). Tom prilikom dolazi do akumulacije AGP na pogođenim mestima i njihovog umrežavanja sa ostalim strukturama ćelijskog zida (Nguema-Ona i sar., 2013). Sa druge strane i biljke gajene u uslovima *in vitro* su izložene mehaničkim povredama koje su neizbežna posledica *in vitro* manipulacija, a biljne ćelije na njih reaguju iniciranjem oksidativnog stresa (Yahraus i sar., 1995; Zhou i Thornburg, 1999) i odbrambenih odgovora koji su praćeni dediferencijacijom, reaktivacijom ćelijskog ciklusa i vaskularnom regeneracijom (Filipović i sar., 2015; Ćuković i sar., 2020; Lup i sar., 2016).

Glavni putevi regeneracije biljaka gajenih u uslovima *in vitro* su *de novo* organogeneza izdanaka (SO, engl. „*shoot organogenesis*") i somatska embriogeneza (SE, engl. „*somatic embryogenesis*"), kao i *de novo* organogeneza korenova. Ovi procesi zasnivaju se na totipotentnosti somatskih ćelija, koje nakon primanja odgovarajućeg signala reprogramiraju gensku ekspresiju i podležu nizu biohemijskih i morfoloških promena čiji je rezultat embriogena ili organogena ćelija (Duclercq i sar., 2011). SE ili SO mogu da se odvijaju sa ili bez stadijuma kalusa (indirektno ili direktno). Takođe, mogu biti indukovane regulatorima rastenja ili se dešavati spontano. Mehanička povreda koju prati ćelijsko reprogramiranje može spontano idukovati oba procesa (Méndez-Hernández i sar., 2019). Ovi putevi regeneracije se intenzivno koriste u osnovnim istraživanjima morfogeneze *in vitro* gajenih biljaka, organogeneze *in planta* i zigotske embriogeneze, kao i u biotehnologiji za propagaciju elitnih sorti biljaka ili transgenih biljaka, proizvodnju haploida, somatsku hibridizaciju i proizvodnju veštačkih semena (Duclercq i sar., 2011; Karami i sar., 2009).

Uloga AGP u SE i SO pokazana je brojnim istraživanjima na različitim biljnim vrstama. Ispitivanjem uloge AGP u nezrelim semena šargarepe utvrđeno je da se elektroforetska mobilnost AGP menja tokom razvića, kao i da su AGP epitopi u nezrelim semenima razvojno regulisani (van Hengel i sar., 2002). Zaključeno je i da se biološka aktivnost AGP tokom formiranja zigotskih embriona menja zavisno od starosti semena (van Hengel i sar., 2002). Poon i sar. (2012) su AGP kalusa pamuka (*Gossypium hirsutum*), nastalih u procesu SE, izolovali i dodali u hranljivu podlogu na koju su prebačeni eksplantati hipokotila *in vitro* gajene kulture pamuka da bi ispitao njihov uticaj na SE. Utvrđeno je da izolovani AGP dovode do promovisanja SE pamuka, a u slučaju kada su izolovani AGP bili delimično ili potpuno razgrađeni, SE je izostajala. Istraživanjem promena komponenti ćelijskog zida *Quercus suber* tokom SE, korišćenjem imunofluorescentnih eseja sa

monoklonalnim antitelima, detektovan je povećan nivo ukupnih AGP (Pérez-Pérez i sar., 2019). Ekstrakti bogati arabinogalaktanskim proteinima, izolovani iz suspenzija embriogenih i neembriogenih ćelija šećerne repe (*Beta vulgaris* L.) pozitivno utiču na organogenezu i povećavaju broj formiranih izdanaka u poređenju sa kontrolom (Wiśniewska i Majewska-Sawka, 2007). Banana (*Musa sp.*) je pokazivala slab razvoj embriogenih kalusa nakon transformacije, pa je pronalaženje načina da se poboljša SE postao kritičan korak u transformaciji banane. Nakon dodavanja rekombinantnog AGP proteina duvana, procenat embriogenih kalusa banane je povećan, što je prvi dokaz da se primenom stranih AGP može poboljšati SE banane (Shu i sar., 2014).

1.1.4. Pristupi identifikaciji HRGP sekvenci

S obzirom na raznolikost fizioloških i strukturnih uloga u koje su HRGP uključeni i njihovu prisutnost u svim biljkama, potraga za sekvencama HRGP u biljnim genomima i transkriptomima je značajan cilj biljne glikobiologije pošto predstavlja preduslov za njihovo dalje istraživanje. HRGP, kao i svi IDP, imaju slabija evolutivna ograničenja u odnosu na globularne proteine, pošto njihova funkcija nije predodređena specifičnom konformacijom peptidnog lanca, odnosno tercijernom strukturom. Visoka stopa mutacija i niska kompleksnost, odnosno relativno visoka repetitivnost njihovih proteinskih sekvenci ometaju njihovo pronalaženje na osnovu homologije (Johnson i sar., 2017a). Sa druge strane HRGP sekvence imaju visok sadržaj pojedinih aminokiselina i odlikuje ih prisustvo specifičnih kratkih aminokiselinskih motiva (glikomotiva) koji mogu poslužiti kao kriterijum za njihovu identifikaciju. Pretraga HRGP sekvenci se zbog ovoga bazira na kombinaciji ključnih odlika ovih proteina:

- Prisustvo N-terminalnog signalnog peptida je zajednička odlika HRGP pošto su to molekuli lokalizovani u ćelijskom zidu.
- Odnos aminokiselina u sastavu HRGP nije balansiran i prezastupljene su aminokiseline koje ulaze u sastav motiva za glikozilaciju. Klasični AGP su bogati sa P, A, S i T, te se u literaturi one sekvence koje sadrže više od 50% PAST aminokiselina identifikuju kao klasični AGP, dok se AG peptidima smatraju kratke sekvence sa više od 35% PAST (Schultz i sar., 2002; Showalter i sar., 2010). PRP se smatraju sekvence sa više od 45% PVKCYT (Showalter i sar., 2010) ili PVKY (Johnson i sar., 2017a), a EXT se smatraju sekvence sa više od 45% PSKY (Johnson i sar., 2017a).
- Prisustvo GPI-sp, hidrofobnog C-terminalnog peptida koji je signal dodavanja GPI, je odlika mnogih, ali ne svih AGP i AG peptida (Ellis i sar., 2010; Showalter i sar., 2010; Simonović i sar., 2016).
- Korisni motivi za pronalaženje HRGP sekvenci su $SO_{3–5}$ motivi karakteristični za EXT sekvence (Showalter i sar., 2010; Johnson i sar., 2017a), AG motivi karakteristični za AGP sekvence (Ma i sar., 2017) i motivi kao što su PPVX[KT] i KKPCPP koji služe za identifikaciju PRP sekvenci (Showalter i sar., 2010).
- Pretraga na osnovu homologije može dati upotrebljive rezultate i u slučaju himernih HRGP koji sadrže konzervirane domene. Konzervirani domeni koji su nađeni isključivo kod HRGP nisu poznati, sem za AG peptide (PF06376, prethodno poznat kao DUF1070 (Simonović i sar., 2016)). Ovaj domen je nađen na C-terminusu nekih AG peptida i veći deo njega čini konzerviran GPI-sp koji se seče tokom sazrevanja peptida. Domeni prisutni kod HRGP, kao što su fasciklinski (PF02469), proteini za transfer lipida (*ns-LTP like, engl. „non-specific lipid-transfer proteins“*, PF00234), plastocijanin (PF02298) i nekoliko drugih, su korisni za identifikaciju himernih HRGP sekvenci zasnovanu na homologiji.

Do sada je literaturi opisano nekoliko različitih metodologija za filtriranje HRGP sekvenci koje se baziraju na kombinaciji navedenih osobina HRGP. Jedan od prvih programa za identifikaciju HRGP bio je *BIO OHIO* (Showalter i sar., 2010) koji je kombinovao predviđanje prisustva N-sp i

GPI sa analizom zastupljenosti pojedinih aminokiselina i pretragom domena karakterističnih za AGP. MAAB (engl. „*the motif and amino acid bias*“) metoda razvijena je od strane Johnson i sar. (2017a) ne samo za pronalaženje već i klasifikaciju HRGP u deskriptivne podklase (Tabela 1). Ova klasifikacija je rezultat spoznaje da veliki broj hibridnih HRGP sekvenci nije korektno obuhvaćen inicijalnom podelom HRGP na tri klase (EXT, PRP i AGP) pošto sadrže odlike dve ili čak sve tri klase, jer superfamilija HRGP predstavlja čitav spektar glikoproteina sa različitim stepenom i tipom glikozilacije (Johnson i sar., 2017a). MAAB klasifikacija se bazira na predviđanju N-sp nakon čega sledi određivanje sadržaja karakterističnih aminokiselina, brojanje karakterističnih motiva i predviđanje prisustva GPI-sp (Tabela 1). Pored prethodne dve metode koje su korisne za identifikaciju i klasifikaciju „prototipskih“ HRGP sa visokim sadržajem karakterističnih aminokiselina, razvijena je i metoda *Finding AGP* (Ma i sar., 2017) isključivo za identifikaciju AGP sekvenci kojom je omogućena i identifikacija AGP sekvenci sa niskim sadržajem karakterističnih aminokiselina i motiva. Ova metoda se zasniva na kombinaciji ukupnog i parcijalnog aminokiselinskog sastava u delu sekvence proteina sa najvećom učestalošću karakterističnih aminokiselina, dužini proteina, kao i prisustvu i razdaljini AG glikomotiva.

Tabela 1. MAAB klasifikacija HRGP prema Johnson i sar., (2017a). U tabeli su prikazane klase 0-24 sa opisom kriterijuma na osnovu kojih su klasifikovane.

MAAB klasa	opis
0	Nije HRGP, sadržaj aminokiselina ne ukazuje da je sekvenca HRGP
1	klasični GPI-AGP, > 45% PAST, predviđen GPI-sp
2	klasični CL-EXT, > 45% PSKY, nije predviđen GPI-sp
3	PRP (> 45% PVKY)
4	klasični AGP bez GPI, > 45% PAST, nije predviđen GPI
5	AGP (> 45% PAST) sa visokim sadržajem EXT motiva (SPn + Y)
6	AGP (> 45% PAST) sa prezastupljenim SPn motivima (odnos SPn/Y motiva > 4)
7	AGP (> 45% PAST) sa prezastupljenim Y motivima (odnos SPn/Y motiva < 0,25)
8	AGP (> 45% PAST) sa prezastupljenim PRP motivima
9	GPI-usidreni EXT, >45% PSKY i predviđen GPI-sp
10	EXT (> 45% PSKY) sa prezastupljenim AGP motivima
11	EXT (> 45% PSKY) sa prezastupljenim SPn motivima (odnos SPn/Y motiva > 4)
12	EXT (> 45% PSKY) sa prezastupljenim Y motivima (odnos SPn/Y motiva < 0,25)
13	EXT (> 45% PSKY) sa prezastupljenim PRP motivima
14	GPI-usidreni PRP, > 45% PVKY, predviđen GPI-sp
15	PRP (> 45% PVKY) sa prezastupljenim AGP motivima
16	PRP (> 45% PVKY) sa prezastupljenim EXT motivima (SPn + Y)
17	PRP (> 45% PVKY) sa prezastupljenim SPn motivima (odnos SPn/Y motiva > 4)
18	PRP (> 45% PVKY) sa prezastupljenim Y motivima (odnos SPn/Y < 0,25)
19	AGP (> 45% PAST) koji imaju visok sadržaj aminokiselina karakterističnih za druge HRGP
20	EXT (> 45% PSKY) koji imaju visok sadržaj aminokiselina karakterističnih za druge HRGP
21	EXT (> 45% PSKY) sa prezastupljenim SPn motivima (SPn/Y > 4) koji imaju visok saržaj aminokiselina karakterističnih druge HRGP
22	EXT (> 45% PSKY) sa prezastupljenim Y motivima (SPn/Y < 0,25) koji imaju visok saržaj aminokiselina karakterističnih za druge HRGP
23	PRP (> 45% PVKY) koji imaju visok sadržaj aminokiselina karakterističnih za druge HRGP
24	HRGP sa visokom sadržajem karakterističnih aminokiselina i slabom pokrivenošću sekvence HRGP motivima (< 15% pokrivenost sekvence sa poznatim HRGP motivima)

Nedostatak opisanih metoda se ogleda ili u nemogućnosti identifikacije/pretrage ne-prototipskih HRGP sekvenci sa niskim sadržajem HRGP karakterističnih aminokiselina i malom zastupljenošću HRGP motiva, naročito ukoliko ne sadrže domene za koje se zna da asociraju sa HRGP, ili u potencijalno visokoj frekvenciji lažno identifikovanih sekvenci kod metoda koje kao kriterijum za identifikaciju koriste mali broj HRGP karakterističnih motiva.

1.2. Opšte karakteristike kičice (*Centaurium erythraea* Rafn.)

Kičica (*Centaurium erythraea* Rafn.) je jednogodišnja ili dvogodišnja zeljasta biljka sa staništem u Evropi, Jugoistočnoj Aziji i Severnoj Africi (Mroueh i sar., 2004; Subotić i sar., 2006). Pripada rodu Centaurium, familiji Gentianaceae (familiji lincura), redu Gentianales. Odlikuje je stablo visine 10-50 cm, sa izduženim listovima organizovanim u rozete i cvetovima u štitastim cvastima. Naseljava šumska staništa, pašnjake, puteve i staze do 1400 m nadmorske visine (Van Rossum, 2009).

Zahvaljujući prisustvu sekoiridoida: svercijamarina, sverozida i (Šiler i sar., 2012) i ksantona, eustomina i dimetileustomina (Aberham i sar., 2011), kičica ima značajna lekovita svojstva i koristi se za tretiranje gastrointestinalnih tegoba, anemije i drugih stanja. Ekstrakti nadzemnih delova *C. erythraea* poseduju hepatoprotektivna, diuretička, protivupalna, antioksidativna, antibakterijska, antigljivična kao i antidijabetična svojstva (Šiler i sar., 2014; Mroueh i sar., 2004; Hamza i sar., 2010). Zbog medicinske vrednosti dolazi do nekontrolisanog branja sa prirodnih staništa pa otuda i potreba za uzgajanjem kičice.

Kičicu karakteriše izvanredna razvojna plastičnost pošto poseduje veliki potencijal za regeneraciju u uslovima *in vitro* što olakšava istraživanja različitih morfogenetskih puteva (Filipović i sar., 2015), kao i snažan morfogenetski potencijal što omogućava njenu upotrebu za genetske transformacije pomoću *A. rhizogenes* (Subotić i sar., 2003) *i A. tumefaciens* (Trifunović i sar., 2013; Trifunović i sar., 2015). Nedavnim istraživanjima Bogdanović i sar. (2021) ustanovljeni su protokoli za uspešnu sekundarnu somatsku embriogenezu kičice.

Zbog svoje interesantne populacione dinamike, fitohemijskih osobina, lekovitih svojstava, razvojne plastičnosti i lake manipulacije u uslovima *in vitro,* kičica je jedna od najistraživanijih biljnih vrsta na odeljenju za fiziologiju biljaka Instituta za biološka istraživanja „Siniša Stanković“, sa potencijalom da bude šire prihvaćena kao model organizam. Jedna od glavnih prepreka ovome je nedostatak genetske sekvence kičice zbog koje su istraživanja na molekularnom nivou ograničena. Veliki korak u otklanjanju ove prepreke je nedavno sekvenciran transkriptom kičice (Ćuković i sar., 2020).

1.2.1. Istraživanja HRGP kod kičice

Ispitivanja HRGP kod kičice su do sada bila koncentrisana na ulogu i dinamiku AGP tokom razvojnih procesa kičice. Ova istraživanja su se u najvećem delu bazirala na korišćenju βGlcY ili antitela specifičnih za glikanske epitope AGP. Pokazano je da sa povećanjem koncentracije βGlcY u podlozi za gajenje biljaka dolazi do smanjenja učestalosti formiranja somatskih embriona i adventivnih pupoljaka kičice (Simonović i sar., 2015). Dodatak βGlcY je najviše uticao na indirektnu SE, i na direktno razviće izdanaka, dok na proces direktnog razvića korenova nije imao uticaj. Tokom pomenutih procesa detektovan je porast koncentracije AGP pri čemu je najbrži porast detektovan tokom indirektne SE koja je ujedno i najosetljivija na dodatak βGlcY. Sa druge strane pokazano je da određene koncentracije βGlcY u podlozi dovode do stimulacije indukcije i regeneracije izdanaka u kulturi korenova kičice, pri čemu povećanje koncentracije βGlcY u podlogama za indukciju SE dovodi do porasta koncentracije AGP u regenerisanim korenovima i izdancima (Trifunović i sar.,

2014). Upotrebom antitela JIM4, JIM8, JIM13, JIM15, LM2, 28 LM14 i MAC207 koja su specifična za glikane AGP pokazano je da se distribucija AGP epitopa menja tokom formiranja somatskih embriona i adventivnih pupoljaka kičice i da se u kasnijim fazama ovih procesa količina detektovanih AGP epitopa smanjuje (Filipović i sar., 2021).

Navedene studije identifikovale su AGP kao značajan faktor tokom SE i SO kod kičice. AGP identifikovani kao učesnici SE, SO ili signalnog puta izazvanog povredama mogu biti značajni za proces regeneracije ne samo za biljke gajene u uslovima *in vitro*, već i u prirodi. Arabinogalaktanski proteini koji učestvuju u odgovoru na povrede su najverovatnije kodirani genima koji su povezani sa regeneracijom, samim tim moguća je potencijalna upotreba tih gena u poboljšanju ili povećanju tolerancije useva na vetar, poplave, povrede izazvane insektima ili biljojedima koji izazivaju mehaničke povrede.

1.3. Mašinsko učenje – kratak pregled

Pošto se deo ove teze bavi unapređenjem indentifikacije HRGP sekvenci upotrebom mašinskog učenja (MU) u narednim odeljcima biće napravljen kratak pregled osnovnih pojmova vezanih za MU koji su neophodni za razumevanje preduzetog procesa MU, rezultata i diskusije.

1.3.1. Mašinsko učenje – osnovni pojmovi

Mašinsko učenje predstavlja razvoj i upotrebu računarskih algoritama koji nisu eksplicitno programirani za rešavanje nekog problema već to uče na osnovu iskustva bilo kroz podatke o problemu ili ponavljanjem (Nikolić i Zečević, 2019). Raznovrsnost i konstantno povećanje broja podataka sa kojima se manipuliše u savremenom svetu zahtevaju tehnološki napredak u praktično svim sferama informacionih tehnologija, a naročito algoritama za MU koji omogućavaju automatizovano dobijanje korisnih informacija iz velike količine sirovih podataka ili predviđanje budućih trendova korišćenjem trenutno dostupnih podataka. Algoritmi MU su u poslednjoj deceniji ostvarili nezamisliv napredak i danas su u stanju da rešavaju visoko kompleksne zadatke kao što su sofisticirana generacija teksta, koji je teško razlikovati od teksta koji je pisao čovek (Brown i sar., 2020), nadljudsko igranje igara kao što su šah (Silver i sar., 2017) i go (Silver i sar., 2017), visoko precizno predviđanje tercijerne proteinske strukture na osnovu sekvence (Jumper i sar., 2021), autonomna vožnja automobila (Grigorescu i sar., 2020) i mnogi drugi.

Tehnike mašinskog učenja imaju široku primenu u bioinformatici (Baldi i Brunak, 2001; Larrañaga i sar., 2006), genetici (Libbrecht i Noble, 2015), medicini (Kourou i sar., 2015; Rajkomar i sar., 2019) i farmaceutskoj industriji (Duvenaud i sar, 2015). U pomenutim naukama najčešće se koriste dva tipa MU - nadgledano (engl. „*supervised learning*“) i nenadgledano (engl. „*unsupervised learning*“) učenje (James i sar., 2017). Ova, relativno neformalna, podela ne obuhvata sve MU tipove i algoritme čija se kompleksnost i heterogenost povećava iz dana u dan, ali je po mišljenju autora dovoljna za razumevanje metoda koje su korišćene u okviru ove disertacije:

- Nadgledano MU (engl. „*supervised learning*“) je najčešće primenjivan tip mašinskog učenja, u kome je ulaznom skupu podataka X, koji predstavlja nezavisne promenljive ili atribute koji opisuju svaku opservaciju, pridružena izlazna vrednost y, odnosno zavisna promenljiva. Može se posmatrati kao da su ulazni podaci X obeleženi sa y. Zadatak algoritma MU je da na osnovu ovih podatka (X, y) proceni funkciju koja slika X u y. Sam čin procene pomenute funkcije naziva se obučavanje, odnosno treniranje modela (Nikolić i Zečević, 2019), i o ovome će biti reči nešto kasnije. U zavisnosti od distribucije y, razlikuju se dve osnovne vrste problema koje

se mogu rešiti ovim tipom mašinskog učenja: regresioni, kada je y kontinualna vrednost i klasifikacioni problemi, kada je y diskretna vrednost (James i sar., 2017).

- Nenadgledano MU (engl. „*unsupervised learning*“) predstavlja tip mašinskog učenja u kome su dati ulazni podaci X, ali ne i izlazna vrednost y, a zadatak je pronaći pravilnosti koje postoje u podacima. Najčešći tipovi zadataka u nenadgledanom MU su grupisanje (engl. „*clustering*“) i smanjenje dimenzionalnosti odnosno učenje reprezentacija (engl. „*dimensionality reduction*“) (James i sar., 2017; Nikolić i Zečević, 2019).

Osim navedenih tipova postoji i polunadgledano MU (engl. „*semi-supervised learning*“) kao njihova kombinacija, gde je za manji deo ulaznih podataka X dat y (obeleženi podaci), dok za veći deo nije (neobeleženi podaci). Zatim učenje potkrepljivanjem (engl. „*reinforcement learning*“) koje se koristi kada je za rešavanje nekog problema potrebno preduzeti niz akcija i drugi (Nikolić i Zečević, 2019). Sa razvojem neuronskih mreža, koje danas čine najpopularniji skup algoritama za MU, splet zadataka koji se rešavaju MU je drastično povećan pa samim tim mnoge od tih algoritama nije lako rasporediti u prethodno navedene tipove.

1.3.2. Kratak pregled obučavanja i procene modela u nadgledanom MU

U prethodnom odeljku (1.3.1.) nadgledano MU je definisano kao skup algoritama koji na osnovu ulaznih podataka X i asocirane zavisne promenljive y imaju za zadatak da procene funkciju koja slika X u y. Kao krajnji cilj ovoga je dobijanje modela koji je u stanju da na osnovu nekih drugih ulaznih podataka X_2 i procenjene funkcije (modela) predvidi nepoznati y_2. Ukoliko je predviđanje nepoznatog y_2 tačno ili blizu istini onda kažemo da model ima dobre performanse. Iz ovoga je jasno da je cilj nadgledanog MU dobijanje modela koji ima dobru sposobnost generalizacije, odnosno koji će davati tačna predviđanja ne samo za podatke koji su korišćeni za njegovo obučavanje već i za nove podatke koje algoritam nije koristio tokom procene modela (James i sar., 2017).

Osnovne faze tokom nadgledanog MU su: prikupljanje i priprema podataka, treniranje različitih algoritama i njihova evaluacija sa ciljem odabira najboljeg, validacija performansi na podacima koji nisu korišćeni za treniranje algoritma i na kraju korišćenje modela za rešavanje problema.

Prikupljanje podataka, njihov pregled i priprema za nadgledano MU su kritični koraci od kojih zavisi kvalitet celog procesa i konačnog modela. Prikupljanje podataka predstavlja usko grlo u MU zbog čega predstavlja aktivnu temu istraživanja (Roh i sar., 2019). Za neke aplikacije jednostavno nema dovoljno obeleženih podataka, a dodatak novih podataka je skup, nepraktičan ili nemoguć, dok za druge ulazni podaci nisu u iskoristivom obliku pa ih je potrebno na neki način transformisati, što zahteva dodatne resurse. Skup podataka na osnovu kojih algoritam MU formira model naziva se skup za obučavanje ili treniranje (engl. „*training set*“).

Pod treniranjem modela podrazumeva se upotreba algoritama MU koji na osnovu raspoloživih parova (X, y) procenjuju funkciju koja slika skup nezavisno promenljivih (X) u zavisnu promenljivu (y). Ta funkcija predstavlja model, i u optimalnom slučaju ona opisuje realan proces koji povezuje X i y. Često ovo nije slučaj pošto na y mogu uticati i faktori koji nisu opisani/uključeni u X, odnosno realan proces koji povezuje X i y može biti tako kompleksan da nije moguća njegova približna procena na osnovu raspoloživih podataka (James i sar., 2017). Tokom treniranja algoritam MU određuje parametre modela; takvi su na primer koeficijent pravca i odsečak u linearnoj regresiji. Parametri se razlikuju u zavisnosti od tipa algoritma MU, a zajedničko im je da ih sam algoritam MU određuje tokom procesa obučavanja kako bi minimizovao funkciju greške (engl. „*loss function*“), koja predstavlja meru odstupanja predviđenih i stvarnih vrednosti[1] zavisne promenljive y (James i

[1] Razlika predviđene i stvarne vrednosti se naziva rezidual.

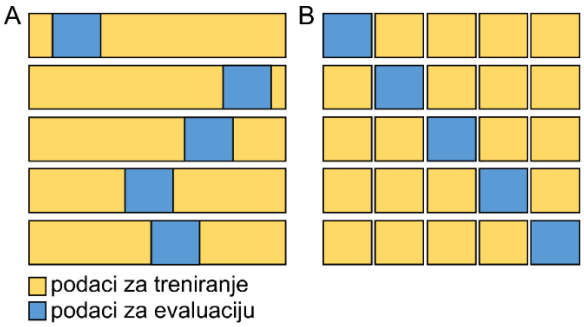
sar., 2017; Nikolić i Zečević, 2019). U problemima regresije kao srednja vrednost funkcije greške (srednja greška) se često koristi RMSD (engl. „*root mean square deviation*“) koji predstavlja kvadratni koren sume kvadrata razlika između stvarnih i predviđenih vrednosti, dok se u problemima klasifikacije sa dve klase često koristi *log loss* koji predstavlja negativni prosek logaritma korigovanih predviđenih verovatnoća i kao takav predstavlja meru koja opisuje koliko su predviđene verovatnoće bliske istinskim klasama. Minimizacija srednje greške izborom parametara modela predstavlja prilagođavanje modela podacima (Nikolić i Zečević, 2019). Postoji relativno veliki broj različitih algoritama MU, od onih koji određuju mali broj parametara tokom prilagođavanja podacima (niska fleksibilnost), do izuzetno fleksibilnih algoritama čiji modeli mogu biti visoko nelinearni. Pri treniranju modela potrebno je izbeći nedovoljnu prilagođenost modela podacima (engl. „*underfittin*g“) kada funkcija koja se procenjuje nema dovoljno parametara, odnosno nema dovoljnu fleksibilnost da opiše realan proces koja povezuje X i y. Takođe potrebno je izbeći i preprilagođavanje modela podacima (engl. „*overfitting*“) kada model koja se procenjuje ima previše parametara pa umesto da opisuje opšte pravilnosti u podacima, počinje da ih „memoriše“ što se loše odražava na mogućnost ovakvih modela da generalizuju (James i sar., 2017; Nikolić i Zečević, 2019). Pronalaženje balansa je poznato i kao problem kompromisa između sistematskog odstupanja i varijanse (engl. „*bias/variance trade-off*“). Varijansa opisuje koliko bi se kreirani model MU promenio, kada bi došlo do promena podataka u skupu za obučavanje. Sistematsko odstupanje se odnosi na grešku koja se javlja kada se jednostavan model MU koristi za rešavanje složenijeg problema. Što je metoda MU fleksibilnija odnosno prilagodljivija podacima, to će varijansa biti veća, a sistematsko odstupanje manje (James i sar., 2017; Nikolić i Zečević, 2019).

Minimizacija srednje greške na skupu podataka za treniranje nije optimalan kriterijum za dobijanje modela koji dobro generalizuje, naročito kada se koriste algoritmi MU koji proizvode fleksibilnije modele. U takvim slučajevima minimizacija srednje greške podrazumeva i preprilagođavanje podacima i gubitak generalizacije modela. Dakle za evaluaciju performansi modela potrebno je koristiti neku metriku koja kvantifikuje njegovu moć generalizacije, odnosno pravilnog predviđanja na podacima koji nisu korišćeni za njegovu obuku (James i sar., 2017; Nikolić i Zečević, 2019). Najčešća dva pristupa kojima se ovo postiže su validacija i unakrsna validacija:

- Kod validacije se skup dostupnih podataka inicijalno podeli na dva dela. Nešto veći deo (50 – 80% podataka) se koristi za treniranje modela, dok se performanse modela evaluiraju na preostalom delu podataka (James i sar., 2017). Odabira se model koji postiže najbolje performanse na skupu podataka za evaluaciju (naziva se i skup za validaciju). Glavni nedostatak ovakvog pristupa je taj što performanse na skupu za validaciju mogu dosta da variraju u zavisnosti od toga koji su podaci odabrani za treniranje, a koji za validaciju. Dakle ovakav pristup je subjektivan i ne kvantifikuje varijansu u performansama modela u zavisnosti od skupa podataka za evaluaciju (James i sar., 2017).
- Unakrsna validacija podrazumeva ponavljanje treniranja modela i njegove evaluacije nekoliko puta (James i sar., 2017). U zavisnosti od toga kako se formira podela između instanci za treniranje i evaluaciju razlikuje se nekoliko tipova unakrsne validacije:
  - Monte Carlo unakrsna validacija podrazumeva određen broj ponavljanja treniranja i evaluacije modela na različitim nasumičnim podelama dostupnih podataka (Slika 3A)
  - Unakrsna validacija sa k podela[2] (engl. „*k-fold cross-validation*“) podrazumeva podelu podataka na k delova (gde je k na primer 5 ili 10) da bi se u k iteracija ponovilo treniranje modela na k-1 delova podataka i evaluacija modela na preostalom delu podataka (Slika 3B). Ovaj proces može da se ponovi nekoliko puta sa različitom podelom podataka na k delova i tada se naziva ponovljena unakrsna validacija sa k podela. U ekstremnom slučaju kada je k jednako broju opservacija u dostupnom skupu podataka metoda se naziva *jackknife* unakrsna validacija (James i sar., 2017).

[2] U literaturi na srpskom jeziku čest je i termin unakrsna validacija u k-slojeva. U ovom tekstu taj termin je rezervisan za jedan drugi tip unakrsne validacije, o kome će biti reči nešto kasnije.

Glavni nedostatak unakrsne validacije je taj što se odvija u nekoliko iteracija treniranja i evaluacije modela što u zavisnosti od vremena potrebnog za treniranje i dostupnih računarskih resursa može biti nepraktično.



Slika 3. Ilustracija podele podataka na skup za treniranje i skup za evaluaciju u unakrsnoj validaciji. A. Monte Carlo unakrsna validacija sa 5 ponavljanja. B. Unakrsna validacija sa 5 podela.

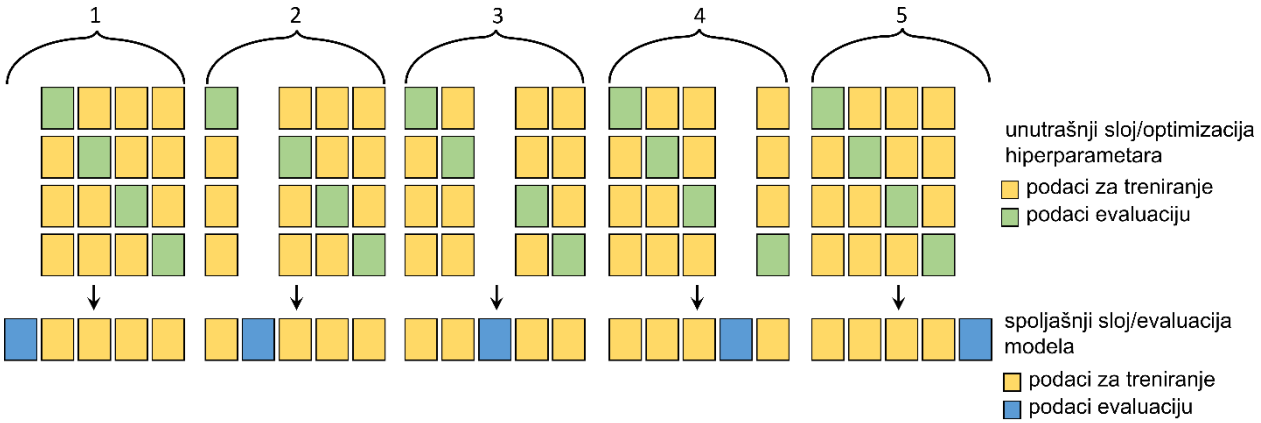
Brojni algoritmi MU poseduju i takozvane hiperparametre[3] (engl. „*hyperparameter*“), od kojih zavisi način procene parametara odnosno prilagođavanje modela podacima (James i sar., 2017). Za razliku od parametara modela koje algoritam MU određuje na osnovu skupa podataka za treniranje, vrednosti hiperparametara se zadaju pre samog treniranja pošto algoritmi MU najčešće nemaju mogućnost da ih odrede na osnovu podataka. Pravilan odabir hiperparametara ključan je za treniranje modela sa visokim performansama. Odabir hiperparamatera se najčešće svodi na isprobavanje određenog broja konfiguracija, odnosno različitih kombinacija njihovih vrednosti, i odabira one, koja proizvodi model sa najboljim performansama. Postoji nekoliko pristupa izboru konfiguracija hiperparametara koji će se evaluirati od kojih su najrasprostranjeniji nasumična pretraga (engl. „*random search*“), pretraga po pravilnoj mreži (engl. „*grid search*“) i Bajesovska optimizacija (Claesen i De Moor, 2015):

- Kod nasumične pretrage se unapred definiše opseg vrednosti za svaki hiperparametar, a sam izbor vrednosti hiperparametara je nasumičan, najčešće baziran na uniformnoj raspodeli, kada je izbor bilo koje vrednosti podjednako verovatan. Broj konfiguracija hiperparametara koje će se evaluirati može biti unapred zadat, ili je moguće evaluirati konfiguracije dok se ne postigne određen nivo performansi modela, ili dok se performanse modela ne povećavaju određen broj iteracija.
- Kod pretrage po pravilnoj mreži najčešće se unapred biraju vrednosti za svaki hiperparametar, a evaluiraju se sve moguće kombinacije ovih vrednosti.
- Bajesovska optimizacija podrazumeva postojanje sekundarnog modela MU. Inicijalno se određen broj kombinacija hiperparametara ispita korišćenjem nasumične pretrage. Korišćene konfiguracije hiperparametara i performanse modela koji su obučeni koriste se kao podaci za sekundarni model, koji na osnovu njih predlaže buduće konfiguracije hiperparametara koje će se evaluirati. Proces se nastavlja najčešće unapred određen broj iteracija ili kada se performanse modela ne povećavaju određen broj iteracija.

Performanse različitih konfiguracija hiperparametara se najčešće evaluiraju validacijom ili unakrsnom validacijom sa k podela kako bi se odabrao model (definisan određenom kombinacijom hiperparametara) koji najbolje generalizuje. Prilikom ovoga se javlja dodatan problem, a to je da performanse različitih konfiguracija hiperparametara neće zavisiti samo od kapaciteta odgovarajućih modela da generalizuju već i od unapred definisane podele podataka na skupove za treniranje i validaciju (Varma i Simon, 2006). Drugim rečima hiperparametari se prilagođavaju unapred definisanoj podeli podataka. Ovaj problem postaje izraženiji sa povećanjem broja iteracija pretrage hiperparametara. Usled ovoga potrebno je odabranu konfiguraciju hiperparametara dodatno proveriti

[3] U literaturi na srpskom jeziku čest je i termin metaparametri. U ovom tekstu će se isključivo koristiti anglicizam hiperparametri.

na nezavisnom skupu podataka ili više skupova kako bi se stekao nepristrasan utisak o performansama modela. Jedan od načina da se ovo postigne je unakrsna validacija u dva sloja[4] (Slika 4, engl. „*nested cross-validation*“). Spoljašnji sloj služi za nepristrasnu evaluaciju modela i on se najčešće sastoji od unakrsne validacije sa k podela. U svakom od odgovarajućih skupova podataka za treniranje spoljašnjeg sloja odigrava se unutrašnja unakrsna validacija koja služi za odabir hiperparametera modela kroz određen broj iteracija. Kada se odabere kombinacija hiperparametara na osnovu performansi u unutrašnjem sloju unakrsne validacije, ona se koristi za obučavanje modela na celom skupu za treniranje odgovarajuće podele spoljašnjeg sloja, i evaluira na odgovarajućem skupu za evaluaciju. Proces se ponavlja dok se ne iscrpi k podela spoljašnjeg sloja.



Slika 4. Ilustracija unakrsne validacije u dva sloja. Spoljašnji sloj se sastoji od unakrsne validacije sa k (5 prikazano na slici) podela i služi za nepristrasnu evaluaciju modela. U svakoj od 5 podela podataka spoljašnjeg sloja odigrava se dodatna unutrašnja unakrsna validacija koja se u ovom slučaju sastoji od 4 podele, a koja služi za evaluaciju različitih konfiguracija hiperparametara kroz određen broj iteracija. Kada se na osnovu performansi u unutrašnjem sloju unakrsne validacije odabere odgovarajuća kombinacija hiperparametara one se koristi za treniranje modela na odgovarajućom podacima za treniranje spoljašnjeg sloja i evaluira na podacima za evaluaciju spoljašnjeg sloja.

1.3.2.1. Binarna klasifikacija

Binarna klasifikacija predstavlja posebnu grupu klasifikacionih problema kada se predviđa pripadnost jednoj od dve klase. U ovakvim problemima jedna od klasa se naziva pozitivna klasa i ona je najčešće asocirana sa vrednošću 1, dok se druga naziva negativna klasa i ona je najčešće asocirana sa vrednošću 0. Modeli za binarnu klasifikaciju kao predviđanje najčešće daju numeričku vrednost između 0 i 1, a ova vrednost se na osnovu nekog praga odluke prevodi u klase (Nikolić i Zečević, 2019). Prag odluke (engl. „*probability trehshold*“) dakle predstavlja vrednost od 0 do 1 na osnovu koje se predviđene verovatnoće pretvaraju u predviđene klase. Za vrednost se najćešće bira 0,5, kada se radi o klasifikacionim problemima u kojima su frekvencije obe klase slične, dok se u problemima kada je jedna od klasa više zastupljena u skupu podataka za treniranje često pristupa optimizaciji ove vrednosti kako bi se maksimizovala neka metrika koja kvantifikuje performanse binarne klasifikacije.

Metrike kojima se kvantifikuju performanse binarne klasifikacije se mogu definisati na osnovu matrice konfuzije (Tabela 2) koja predstavlja tabelarni prikaz broja elemenata određene klase

[4] U literaturi na srpskom jeziku često se pod terminom unakrsna validacija u k-slojeva podrazumeva metoda koja se u ovoj tezi naziva unakrsna validacija sa k-podela (engl. „*k-fold cross-validation*”). Termin unakrsna validacija u dva sloja (engl. „*nested cross-validation*“) ovde podrazumeva da se unutar svake podele unakrsne validacije sa k-podela odigrava dodatna unakrsna validacija sa k-podela. Odnosno da postoji spoljašnji sloj unakrsne validacije sa k-podela koji se koristi za procenu performansi modela i unutrašnji sloj unakrsne validacije sa k-podela koji se koristi za optimizaciju modela.

koji su klasifikovani tačno, odnosno pogrešno (James i sar., 2017; Nikolić i Zečević, 2019). Elementi matrice konfuzije su:

- stvarno pozitivni (TP, engl. „*true positive*“) predstavlja broj opservacija koji pripadaju pozitivnoj klasi, a dodeljena im je pozitivna klasa.
- lažno negativni (FN, engl. „*false negative*“) predstavlja broj opservacija koji pripadaju pozitivnoj klasi, a dodeljena im je negativna klasa.
- lažno pozitivni (FP, engl. „*false positive*“) predstavlja broj opservacija koji pripadaju negativnoj klasi, a dodeljena im je pozitivna klasa.
- stvarno negativni (TN, engl. „*true negative*“) predstavlja broj opservacija koji pripadaju negativnoj klasi, a dodeljena im je negativna klasa.

Tabela 2. Matrica konfuzije za problem binarne klasifikacije.

		predviđena klasa	
		klasa 1	klasa 0
stvarna klasa	klasa 1	stvarno pozitivni (TP)	lažno negativni (FN)
	klasa 0	lažno pozitivni (FP)	stvarno negativni (TN)

Najčešće upotrebljavane metrike definisane na osnovu vrednosti iz matrice konfuzije su:

- osetljivost ili senzitivnost (TPR, engl. „*true postive rate*“) predstavlja udeo predviđenih stvarno pozitivnih opservacija od ukupnog broja pozitivnih opservacija:

$$TPR = \frac{TP}{TP + FN}$$

- specifičnost (TNR, engl. „*true negative rate*“) predstavlja udeo predviđenih stvarno negativnih opservacija od ukupnog broja negativnih opservacija:

$$TNR = \frac{TN}{TN + FP}$$

- tačnost (ACC, engl. „*accuracy*“) predstavlja udeo tačno predviđenih opservacija:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}$$

- balansirana tačnost (BACC, engl. „*balanced accuracy*“) predstavlja prosek osetljivosti i specifičnosti:

$$BACC = \frac{TPR + TNR}{2}$$

- Matthews koeficijent korelacije (Matthews, 1975) predstavlja koeficijent korelacije između stvarnih i predviđenih klasa i često se koristi u problemima kada je jedna od klasa frekventnija u skupu podataka za treniranje:

$$MCC = \frac{TP\ x\ TN - FP\ x\ FN}{\sqrt{(TP + FP)\ x\ (TP + FN)\ x\ (TN + FP)\ x\ (TN + FN)}}$$

- *Cohen*-ov *kappa* koeficijent (Cohen, 1960) uzima u obzir mogućnost da poklapanje između predviđanja i istinske klase bude slučajno (verovatnoća raste sa disbalansom u frekvencija klasa u skupu podataka za treniranja):

$$\kappa = \frac{p_0 - p_c}{1 - p_c}$$

gde je p_0 tačnost:

$$p_0 = \frac{TP + TN}{N}$$
$$N = TP + TN + FP + FN$$

a p_c očekivano slučajno poklapanje između predviđanja i stvarne klase

$$p_c = \frac{TP + FP}{N}\ x\ \frac{TP + FN}{N} + \frac{TN + FN}{N}\ x\ \frac{FP + TN}{N}$$

Osim pomenutih metrika koje se oslanjaju na matricu konfuzije, u binarnoj klasifikaciji često se koristi i površina ispod ROC (engl. „*receiver operating characteristic*“) krive (AUC). Ova metrika se koristi kada model predviđa verovatnoće pripadnosti klasama (0 - 1) i ne zavisi od praga odluke kojom se predviđene verovatnoće prevode u klase, već kvantifikuje performanse uzimajući u obzir sve moguće pragove odluke (James i sar., 2017). ROC kriva predstavlja zavisnost udela stvarno pozitivnih opservacija (senzitivnost) od udela lažno pozitivnih opservacija (1 - specifičnost) za svaki mogući prag odluke (vrednosti verovatnoća koju je model predvideo na nekom skupu podataka). Površina ispod ove krive predstavlja verovatnoću da pri slučajnom izboru dve opservacije iz različitih klasa, opservacija koja pripada negativnoj klasi ima manju predviđenu verovatnoću od one koja pripada pozitivnoj klasi (Nikolić i Zečević, 2019). Vrednost od AUC = 1 (maksimalna vrednost) podrazumeva da su predviđene verovatnoće za sve opservacije koje pripadaju pozitivnoj klasi veće od verovatnoće predviđene za bilo koju opservaciju koja pripada negativnoj klasi, odnosno da su senzitivnost i specifičnost predviđanja klasa maksimalne (jednake 1).

1.3.2.2. Odabrani algoritmi mašinskog učenja

Danas postoji jako puno različitih algoritama za MU, čak i pregled najvažnijih bi daleko prevazišao okvire ovog teksta. Ovaj odeljak pre svega služi za predstavljanje osnovnih karakteristika četiri algoritma MU, i to isključivo u okviru klasifikacije, čije su performanse evaluirane u okviru ove teze:

- Metod potpornih vektora (SVM, engl. „*support vector machine*“) je algoritam za nadgledano MU koji se koristi za rešavanje klasifikacionih i regresionih problema (Cortes i Vapnik, 1995). U slučaju klasifikacije SVM za cilj ima nalaženje optimalne hiperravni koja na najbolji mogući način razdvaja instance koje pripadaju različitim klasama skupa za treniranje u vektorskom prostoru. Optimalna hiperravan, odnosno hiperravan najšireg pojasa (maksimalne margine) je u SVM definisana kao hiperravan koja je podjednako udaljena od najbližih predstavnika obe klase (Nikolić i Zečević, 2019). U realnim problemima često nije moguće pronaći hiperravan koja u potpunosti razdvaja predstavnike klasa pa je uveden dodatni parametar C koji reguliše stepen tolerisanja grešaka, a ovakva metoda se naziva metod potpornih vektora sa mekim pojasom (engl. „*soft margin*”). Kada je C hiperparameter mali, algoritam ima tendenciju da traži hiperravan koja razdvaja klase sa širokom pojasom tolerišući

pogrešno klasifikovane opservacije, a kada je C hiperparametar visok, pojas koji razdvaja klase oko hiperravani je uži, a tako dobijeni modeli imaju tendenciju da bolje minimizuju prosečnu grešku klasifikacije podataka koji se koriste za treniranje modela. U SVM se često koriste kernel funkcije koje imaju ulogu da preslikaju ulazne podatke X u neki vektorski prostor sa ciljem da se potencijalno nelinearna površina razdvajanja klasa pretvori u hiperravan koja se može opisati linearnom jednačinom u tom vektorskom prostoru i rešiti sa SVM. Najčešće primenjivan tip kernela, kada se o podacima za treniranje malo zna, je kernel sa radijalnom osnovom (RBF, engl. „*radial basis function*") za koji je karakterističan hiperparametar γ koji definiše širinu uticaja svake opservacije (James i sar., 2017).

- k najbližih suseda (Fix i Hodges, 1951, KNN, engl. „*k-nearest neighbors*") je algoritam inicijalno zamišljen za klasifikaciju, a zasniva se na upotrebi rastojanja između instanci skupa za obučavanje predstavljenih kao tačke u vektorskom prostoru. Algoritam polazi od pretpostavke da će se slične instance nalaziti bliže u prostoru, a nepoznata instanca se klasifikuje na osnovu klasa k najbližih suseda iz skupa podataka za treniranje (Nikolić i Zečević, 2019). Pomenuti k predstavlja hiperparametar ovog algoritma koji je potrebno naštimovati, niže vrednosti k kao proizvod daju fleksibilnije modele koji su bolje prilagođeni podacima za treniranje. Kao metrika bliskosti u prostoru najčešće se koristi Euklidsko rastojanje, mada postoje generalizacije kada se koriste i druga, kao na primer Minkowski rastojanje koje je definisano parametrom p. U slučaju kada je p = 1 Minkowski rastojanje je jednako Manhattan rastojanju, a u slučaju kada je p = 2 jednako je Euklidskom rastojanju, tako da se Minkowski rastojanje posmatra kao generalizacija pomenuta dva tipa rastojanja. U ovakvim implementacijama k najbližih suseda parametar Minkowski rastojanja p predstavlja hiperparametar koji se može optimizovati.
- Algoritam *random forests* (RF, Breiman, 2001) pripada klasi ansambl metoda. Ansambli u MU predstavljaju skupove većeg broja modela MU koji se obučavaju na podacima za treniranje, a njihova predviđanja se na određen način agregiraju u konačno predviđanje. Algoritam RF se zasniva na agregaciji predviđanja nezavisnih stabala odlučivanja (engl. „*decision trees*"). Stabla odlučivanja su jednostavan i pre svega lako interpretabilan algoritam MU koja se zasniva na dendogramskoj strukturi. Svaki čvor stabla predstavlja logički test nad jednom nezavisnom promenljivom koji generiše podelu na dva podskupa (grane); sve instance koje su tačne za ovaj test se sortiraju u jednu granu, a sve instance koje nisu tačne se sortiraju u drugu granu (Nikolić i Zečević, 2019). Test za numeričku promenljivu je najčešće veće > ili manje < od neke kritične vrednosti, dok test za diskretne promenljive podrazumeva pripadnost pojedinim klasama. Na ovaj način se instance raspoređuju do listova koji predstavljaju predviđanja modela. Zadatak algoritma je da izabere optimalnu nezavisnu promenljivu i formuliše odgovarajući test u okviru datog čvora stabla odlučivanja čija kombinacija najbolje razvrstava instance tako da dobijene podgrupe budu što homogenije. Ovo se najčešće postiže pohlepnom indukcijom od korena ka listovima (engl. „*top-down induction of decision trees*", Quinlan, 1986) koja podrazumeva da se u okviru svakog čvora radi iscrpna pretraga nad svim dostupnim nezavisnim promenljivim dok se ne pronađe kombinacija nezavisne promenljive i logičkog testa koja maksimizuje homogenost u podgrupama čvora. Pristup je pohlepan jer ne uzima u obzir optimalne kombinacije nezavisnih promenljivih i testova u okviru celog stabla odlučivanja, već samo u okviru svakog pojedinačnog čvora. Osim toga za ovaj tip algoritama potrebno je definisati i kriterijum za zaustavljanje granjanja, koji može biti heuristički, kao na primer minimalan broj instanci u čvoru posle koga dalje nema podele, može biti baziran na nekom statističkom testu, ili kada nehomogenost listova postigne neku malu unapred definisanu vrednost, a implementirani su i drugi pristupi (Nikolić i Zečević, 2019). Ansambl RF se sastoji od određenog broja stabala odlučivanja treniranih na različitim podskupovima skupa za treniranje, koji se najčešće formiraju *bootstrap* metodom (nasumično uzorkovanje instanci skupa sa ponavljanjem) kako bi se smanjila korelacija u predviđanjima dobijenih modela. Dodatno smanjenje korelacije između modela se postiže ograničavanjem broja nezavisnih promenljivih koje su dostupne u

svakom čvoru u okviru stabala odlučivanja. Broj nezavisnih promenljivih koje su nasumično izabrane iz skupa za treniranje, a u okviru kojih se mora definisati određeni čvor je hiperparametar u RF algoritmu, kao i broj stabala odlučivanja koja će biti treniran u okviru RF ansambla. Ostali hiperparametri RF algoritma se najčešće odnose na kriterijume za zaustavljanje granjanja u okviru stabala odlučivanja.

- Algoritam XGBoost (XGB, engl. „*extreme gradient boosting*", Chen i Guestrin, 2016) pripada klasi MU algoritama koji se nazivaju pojačavanje (engl. „*boosting*", Friedman, 2001) i takođe predstavlja ansambl metodu koja se sastoji od puno modela MU najčešće baziranih na stablima odlučivanja. Za razliku od RF gde se koristi prosta agregacija predviđanja nezavisnih modela MU, kod metoda pojačavanja modeli nisu nezavisni već se ansambl gradi tako što se svaki naredni model trenira tako da nadomesti slabosti tekućeg skupa modela (Nikolić i Zečević, 2019). Ovo se postiže praktično treniranjem modela na greškama prethodnog skupa modela. Algoritam XGB predstavlja visoko optimizovanu implementaciju gradijentnog pojačavanja pomoću stabala, i predstavlja jedan od najuspešnijim algoritama MU današnjice. Ovaj algoritam ima jako veliki broj hiperparametara od kojih deo utiče na treniranje pojedinačnih stabala odlučivanja, deo definiše način i tempo učenja celog ansambla, a deo služi da unese dozu nasumičnosti kako bi se smanjila korelacija između grešaka predviđanja pojedinačnih stabala.

1.3.3. Upotreba mašinskog učenja za predviđanja na osnovu proteinske sekvence sa akcentom na predviđanje hidroksilacije prolina

Poslednje dve decenije svedoci smo neverovatnog razvoja tehnika za sekvenciranje nukleinskih kiselina (NGS tehnike, Goodwin i sar., 2016). Baze bioloških sekvenci rastu eksponencijalnom brzinom, a sa njima i pokušaji da se na osnovu sekvence predvidi neka biološki značajna informacija. Sa razvojem NGS tehnika, podudara se i vrtoglava popularnost i razvoj tehnika MU, koje su od male grupe, pre svega akademskih entuzijasta pre dve decenije postale toliko sveprisutne, da praktično svaka osoba, sa savremenim telefonom, sa sobom svuda nosi barem nekoliko modela MU za različite namene. Naizgled kulminacija ova dva procesa, barem iz perspektive bio-nauka, je rešenje glavnog problema u biohemiji, odnosno precizno predviđanje 3D strukture proteina na osnovu sekvence (Jumper i sar., 2021; Baek i sar., 2021). Živimo u uzbudljivim vremenima, a teško je predvideti šta nas očekuje u budućnosti.

Predviđanje mesta posttranslacionih modifikacija u proteinima na osnovu sekvence je jedan od glavnih izazova u biohemiji. Sama formulacija ovog problema je dosta prostija od prethodno pomenutog predviđanja 3D strukture. Za predviđanje PTM značajna je sekvenca motiva oko aminokiseline koja se modifikuje (Amanchy i sar., 2011), a ukoliko se radi o modifikacijama na zrelim proteinima koji su preuzeli nativnu konformaciju, od značaja je dostupnost aminokiseline koja se modifikuje, a potencijalno i interakcije peptidnog motiva oko nje sa udaljenim delovima sekvence koje zajedno čine lokalnu 3D strukturu (Jung i sar., 2010; Yang i sar., 2021).

Nasuprot eksponencijalnom rastu deponovanih bioloških sekvenci u različitima bazama za ovu namenu, broj proteina čije su PTM eksperimentalno anotirane u najboljem slučaju zabeležava linearan rast tokom vremena. Ovo je posledica relativne ograničenosti u propusnosti, kombinovane sa visokom cenom metoda koje služe za eksperimentalnu anotaciju PTM u proteinima. Iako značajan, napredak u masenoj spektrometriji proteina kombinovanoj sa tečnom hromatografijom visokih performansi, jednoj od najmodernijih metoda za detekciju PTM (Parker i sar., 2010; Doll i Burlingame, 2015), neuporediv je sa napretkom koji su doživele NGS tehnologije u istom periodu. Ako se predviđanje pozicija PTM postavi kao supervizovan MU problem, što je u najvećem broju dosadašnjih pokušaja odnosno publikacija i slučaj, za uspešno treniranje modela neophodne su proteinske sekvence (X) za koje se nedvosmisleno zna koje aminokiseline su modifikovane sa

određenom PTM, a koje ne (y). Prema tome glavna prepreka za ovakav tip zadatka je mala dostupnost obeleženih podataka koji se mogu koristiti za treniranje i evaluaciju modela MU. Iako je skorašnji napredak različitih pristupa neuronskim mrežama omogućio rešavanje zadataka izuzetne kompleksnosti, generalni je konsenzus da su impresivni rezultati postignuti modernim algoritmima MU omogućeni kroz upotrebu velike količine podatka za treniranje (Adadi, 2021). Predviđanje PTM na osnovu sekvence je dodatno zakomplikovano činjenicom da je poreklo proteinskih sekvenci kod kojih su PTM anotirane jako pristrasno, odnosno isključivo su anotirane sekvence poreklom od model organizama i to pre svega čoveka. U budućnosti se može očekivati porast dostupnih podataka za treniranje kao i razvoj pristupa MU koji postižu jako dobre performanse i kada su trenirani sa malim skupovima obeleženih podataka.

Za predviđanje PTM na osnovu proteinske sekvence, ukoliko je već dostupan skup za treniranje, potrebno je na neki način niz slova koja simbolizuju redosled aminokiselina u proteinu transformisati u oblik razumljiv računaru, odnosno u numeričke vrednosti. U literaturi je opisano jako puno načina da se ovo postigne od proste zamene aminokiselina sa numeričkim vrednostima koje odgovaraju određenim fizičko-hemijskim osobinama, zatim različitih transformacija koje na neki način oslikavaju pravilnost promene fizičko-hemijskih osobina kroz proteinsku sekvencu, kao što su autokorelacioni deskriptori i drugi. Koncizan pregled metoda za transformaciju proteinske sekvence u oblik razumljiv računaru dat je u Xiao i sar. (2015). Osim pomenutih metoda transformacije koji su nezavisni od samog algoritma za MU, sve su popularnije metode učenja reprezentacija (engl. „*embedings*“) proteinskih sekvenci tokom samog traniranja modela MU. Ovakve metode najčešće kao ulaz prihvataju *one-hot* predstavljenu proteinsku sekvencu gde je svaka pozicija u peptidnom lancu opisana sa dvadeset vektora koji odgovaraju aminokiselinama, od kojih jedan odgovarajući ima vrednost 1, a ostali 0. Ovako predstavljeni podaci imaju nizak sadržaj informacije, veliki broj dimenzija, odnosno jako su proređeni (engl. „*sparse*“) i oni se tokom inicijalnog ili nekoliko inicijalnih slojeva neuronske mreže transformišu u vektorski prostor sa manje dimenzija (Chollet i Allaire, 2019). Ova transformacija se trenira zajedno sa celokupnim modelom MU, te je zbog toga ona blizu optimalnog načina transformacije proteinske sekvence u oblik razumljiv računaru za dati problem. Nedostatak ovakvog pristupa je taj što je za učenje dobrih reprezentacija proteinskih sekvenci neophodno postojanje velike količine obeleženih podataka. U poslednje dve godine je popularno takozvano samo-supervizovano učenje reprezentacija proteinskih sekvenci (Elnaggar i sar., 2020; Rives i sar., 2021). Samo-supervizovano učenje podrazumeva da je za učenje reprezentacija potrebna samo informacija o sekvencama. Algoritmu se kao ulaz daju proteinske sekvence kojima je određen procenat aminokiselina maskiran (nepoznat), a cilj je predvideti ih. Ovakvi pristupi imaju potencijal da drastično povećaju tačnost metoda MU koje uče na osnovu proteinskih sekvenci u budućnosti.

Kada se razmatra predviđanje pozicija hidroksilacije prolina korišćenjem MU u dosadašnjoj literaturi, može se zapaziti da postoji relativno mali broj publikacija. Pre svega zbog relativno malog značaja ove PTM u odnosu na druge u humanoj biohemiji i fiziologiji. Ova PTM se javlja pre svega u strukturnim proteinima, najviše u kolagenu i potrebna je za njegovo pravilno funkcionisanje jer povećava stabilnost trostrukog heliksa kolagena (Kotch i sar., 2008). Poznato je da do hidroksilacije prolina u kolagenu dolazi u motivima XPG te za daljim predviđanjem korišćenjem MU nema potrebe. Dostupni serveri za predviđanje pozicije hidroksiprolina u proteinima su:

- *iHyd-PseAAC* (Xu i sar., 2014) koji kao atribute za predviđanje hidroksilacije Pro koristi sklonost dipeptida ka specifičnoj poziciji unutar lokalne sekvence kojim je modifikovan prethodno definisan deskriptor proteinske sekvence pseudo-aminokiselinski sastav (PseAAC, engl. *„pseudo-amino acid composition“*, Chou, 2001). Kao algoritam MU autori su koristili linearnu diskriminantnu analizu.
- *PredHydroxy* (Shi i sar., 2015) kombinuje informacije o težini položaja aminokiselinskog sastava sa vrednostima osam fizičko-hemijskih osobina aminokiselina koji opisuju lokalnu sekvencu. Kao algoritam MU je korišćen SVM.

- *RF- Hydroxysite* (Ismail i sar., 2016) kao atribute za opisivanje lokalne sekvence koristi fizičko-hemijske, strukturne i evolutivne informacije. Kao algoritam MU je korišćen *RF*.
- *iHyd-PseCp* (Qiu i sar., 2016) je prediktor nastao inkorporiranjem informacije o redosledu sekvence u pseudo-aminokiselinski sastav (PseACC, Chou, 2001) korišćenjem *RF* algoritma.

Sve pomenute metode su kao ulaz koristile proteinske sekvence iz različitih carstava organizama (životinjskih i biljnih), pri čemu su dominantno zastupljene životinjske sekvence. Takođe nijedna od metoda nema kao ulazni atribut da li je sekvenca poreklom iz biljke, životinje ili eventualno bakterije. Kao što će biti predstavljeno u rezultatima, ove metode pokazuju ograničen uspeh u predviđanju hidroksilacije prolina u biljnim proteinima.

2. Cilj rada

Kao osnovni ciljevi ove doktorske disertacije postavljeni su:

- Razvoj nove metodologije za identifikaciju i analizu HRGP sekvenci, sa akcentom na AGP sekvence.
- Identifikacija AGP sekvenci kičice.
- Analiza ekspresije *AGP* gena kičice u odgovoru na stres izazvan mehaničkim povredama.
- Analiza ekspresije *AGP* gena kičice u odgovoru na različite koncentracije βGlcY.
- Analiza ekspresije odabranih *AGP* gena tokom različitih razvojnih procesa kičice.

Realizacija ciljeva izvedena je kroz sledeće korake:

- Treniranje i procena performansi modela mašinskog učenja za predviđanje verovatnoće hidroksilacije prolina na osnovu sekvence u biljnim proteinima.
- Uspostavljanje metodologije za precizno pronalaženje AGP sekvenci koja kao centralni element sadrži model MU za predviđanje hidroksiprolina kombinovan sa uspešnim primerima iz literature.
- Implementacija metodologije za identifikaciju i analizu HRGP sekvenci u softverski paket otvorenog koda.
- Identifikacija gena za *AGP* iz *de novo* sastavljenog transkriptoma kičice korišćenjem pomenute metodologije.
- Ispitivanje ekspresije odabranih *AGP* gena tokom odgovora biljke na abiotički stres izazvan mehaničkom povredom biljnog tkiva kičice gajenog u uslovima *in vitro*.
- Ispitivanje ekspresije odabranih *AGP* gena tokom odgovora biljnog tkiva kičice gajenog u uslovima *in vitro* na produženo izlaganje različitim koncentracijama βGlcY.
- Ispitivanje ekspresije odabranih *AGP* gena u uzorcima tkiva iz različitih faza somatske embriogeneze, organogeneze, biljaka gajenih u uslovima *in vitro* i biljaka iz prirode.

3. Materijal i metode

3.1. Uspostavljanje nove metodologija za filtriranje i analizu HRGP sekvenci

Sa ciljem poboljšanja metodologije za identifikaciju i analizu HRGP sekvenci osmišljen je novi pristup baziran na mašinskom učenju.

3.1.1. Predviđanje hidroksilacije prolina

Model mašinskog učenja koji služi za predviđanje verovatnoće hidroksilacije prolina predstavlja glavnu inovaciju pristupu filtriranja HRGP-a koja je razvijena u ovoj tezi, stoga će detaljno biti objašnjena metodologija treniranja modela, od same pripreme podataka za treniranje i evaluaciju, transformacije proteinskih sekvenci u numerički oblik, odabira atributa korišćenih za treniranje preko same optimizacije hiperparametara i evaluacije modela.

Priprema podataka.

Da bi se algoritam MU obučio da predviđa verovatnoću hidroksilacije prolina u biljnim proteinima korišćeni su podaci o eksperimentalno potvrđenim hidroksiprolinima u biljnim proteinima iz UniProtKB/Swiss-Prot baze podataka (UniProt release 2017_07, www.uniprot.org) (The UniProt Consortium, 2017). Preuzete su sekvence 40 biljnih proteina sa eksperimentalno potvrđenim hidroksiprolinima. Nakon manuelne provere anotiranih regiona i poređenja sa citiranom literaturom, uklanjanja nesekvenciranih regiona kao i hidroksiprolina/prolina za koje se na osnovu literature nije moglo sa sigurnošću utvrditi hidroksilacioni status, odnos hidroksiprolina i prolina u ovom skupu podataka je bio blizu dva (hidroksiprolini su bili duplo frekventniji). Pošto taj odnos verovatno ne oslikava pravi odnos ove dve aminokiseline u sekretovanim biljnim proteinima, uključen je dodatni skup od 269 biljnih proteinskih sekvenci iz UniProtKB/Swiss-Prot baze podataka sa eksperimentalno utvrđenim postojanjem N-sp signalne sekvence, a bez eksperimentalno utvrđenih Hyp prilikom sekvenciranja. Dodatna grupa proteinskih sekvenci (Hyp negativna grupa) je uključivala samo sekretorne proteine, da bi se model trenirao na sekvencama koje imaju sličnost sa sekvencama na kojima će model biti korišćen.

Sve proteinske sekvence su detaljno proverene i upoređene sa UniProtKB/Swiss-Prot anotacijama i literaturnim podacima o sekvenciranju, nakon čega su uklonjeni nesekvencirani regioni i regioni koji pokazuju odstupanja među različitim izvorima podataka da bi se minimizovao broj pogrešno anotiranih prolina/hidroksiprolina. Nakon ovoga, iz preostalih delova proteinskih sekvenci ekstrahovane su lokalne sekvence dužine 21 aminokiselinu (21-merne sekvence, 21-mere) koje su sadržale ± 10 aminokiselina oko svakog Hyp/Pro koji je nakon provere podataka preostao. Nakon toga pristupilo se smanjenju redundantnosti sekvenci – prvo su uklonjene duplirane 21-mere, a za dalje smanjenje redundantnosti korišćena je Levenštajnova distanca određena korišćenjem R paketa *stringdist* (van der Loo, 2014). Levenštajnova distanca predstavlja minimalan broj zamena (supstitucija), insercija ili delecija aminokiselina neophodnih da se jedna 21-merna sekvenca pretvori u drugu. Smanjenje redundantnosti je urađeno posebno za Hyp pozitivne 21-mere (sekvence koje sadrže Hyp u sredini) i Hyp negativne 21-mere (sekvence koje sadrže Pro u sredini), postepenim uklanjanjem sekvenci koje se od ostalih razlikuju u tačno jednoj poziciji (ili za tačno jednu poziciju) na osnovu Levenštajnove distance. Sekvence su uklanjane jedna po jedna, kroz nekoliko iteracija, tako što se nakon određivanja Levenštajnovih distanci između svih parova prvo ukloni sekvenca sa maksimalnim brojem homologa (broj sekvenci koje su od nje udaljene za Levenštajnovu distancu l), nakon čega se vrši ponovna procena i uklanjanje sekvence sa maksimalnim brojem homologa, ovaj pristup se nastavlja dok se iz skupa podataka ne uklone sve sekvence koje su za Levenštajnovu

distancu l udaljene od neke druge, kao što je opisano u Schwartz i sar. (2009). Kao rezultat, dobijen je skup od 225 proteinskih sekvenci sa 1093 21-merom od kojih je 182 21-mere imalo pozitivan hidroskilacioni status (Hyp u sredini). Minimalna Levenštajnova distanca između bilo koje dve 21-mere u ovom skupu podataka je bila 2, odnosno maksimalno je 18 od 20 aminokiselina (90%) oko centralnog Hyp/Pro moglo da bude identično između bilo koje dve 21-mere. Dobijeni skup je zatim nasumično podeljen na deo za treniranje modela koji je činilo oko 80% sekvenci (181 jedinstvena proteinska sekvenca sa 150 21-merom sa Hyp u sredini i 737 21-mera sa Pro u sredini) i deo za evaluaciju modela koji je činilo oko 20% proteinskih sekvenci (44 jedinstvene proteinske sekvence sa 32 21-mere sa Hyp u sredini i 174 21-mere sa Pro u sredini). Da bi se procenio uticaj dužine lokalne sekvence na performanse modela, 21-merne sekvence su smanjenje na 19, 17, 15 i 13-merne sekvence, uklanjanjem po jedne aminokiseline sa krajeva uz isti način uklanjanja homologa kao što je opisano za skup 21-mernih sekvenci.

Transformacija proteinskih sekvenci u numerički oblik

Da bi se algoritam mašinskog učenja trenirao da predviđa hidroksilaciju prolina na osnovu lokalnog konteksta sekvence, potrebno je na neki način prevesti proteinske sekvence u numerički oblik koji će oslikavati sekvence, a sa kojima će algoritam moći da radi. U literaturi je opisano puno načina da se ovo postigne, a pošto nije moguće unapred pretpostaviti koji je bolji za zadatak predviđanja verovatnoće hidroksilacije prolina, korišćeno je 16 (obeleženi sa F1 – F16, F – od engleske reči „*feature*“ – odlika, atribut) različitih načina transformacije proteinskih sekvenci u numerički oblik:

- Prvi skup atributa, F1 podrazumevao je slikanje sekvence aminokiselina u 21-merama u sekvencu numeričkih vrednosti na osnovu međusobno nezavisnih fizičko-hemijskih svojstava aminokiselina: normalizovana prosečna hidrofobnost - CIDH920105, prosečan indeks fleksibilnosti - BHAR880101, slobodna energija rastvora u vodi [kcal/mol] - CHAM820102, zapremina bočnog lanca - BIGC670101, sterički parametar - CHAM810101 i relativna promenljivost - DAYM780201. Preslikane su aminokiseline sa obe strane centralnog Hyp/Pro u 21-merama (20 aminokiselina) što je rezultiralo skupom podataka od 120 dimenzija (6 osobina × 20 aminokiselina – 10 oko centralnog Hyp/Pro).
- Drugi skup atributa, F2 konstruisan je slikanjem sekvence aminokiselina u 21-merama u sekvencu numeričkih vrednosti na osnovu 5 multidimenzionih obrazaca dobijenih Faktor analizom fizičko-hemijskih osobina aminokiselina (Atchley i sar., 2005) čime je dobijen skup atributa od 100 dimenzija (5 osobina × 20 aminokiselina – 10 oko centralnog Hyp/Pro).
- Skupovi atributa F3 i F4 predstavljaju normalizovane *Moreau-Broto* autokorelacione deskriptore (grupe atributa) izračunate na osnovu fizičko-hemijskih osobina korišćenih za F1 (F3) i pet multidimenzionih obrazaca korišćenih za konstrukciju F2 (F4). Pre konstrukcije F3 i F4 su atributi aminokiselina standardizovani tako da im srednja vrednost bude 0, a standardna devijacija 1. Ovaj autokorelacini dekriptor se računa na sledeći način:

$$ATS(d) = \frac{\sum_{i=1}^{N-d} P_i \, P_{i+d}}{N-d} \quad d = 1,2,3 \ldots \max lag = 12$$

  Gde su P_i i P_{i+d} vrednosti korišćene fizičko-hemijske osobine aminokiselina u pozicijima i i $i+d$, N je broj aminokiselina u sekvenci, a d je maksimalna *lag* vrednost koje je za ove potrebe izabrana da bude 12.

- Skupovi atributa F5 i F6 predstavljaju *Moran* autokorelacione deskriptore. F5 je konstruisan na osnovu fizičko-hemijskih osobina korišćenih za F1, a F6 na osnovu pet multidimenzionalnih obrazaca korišćenih za konstrukciju F2. Pre konstrukcije F5 i F6 su atributi aminokiselina standardizovani tako da im srednja vrednost bude 0, a standardna devijacija 1. Ovaj autokorelacini deskriptor se računa na sledeći način:

$$I(d) = \frac{\frac{1}{N-d}\sum_{i=1}^{N-d}(P_i - P')(P_{i+d} - P')}{\frac{1}{N}\sum_{i=1}^{N}(P_i - P')^2} \quad d = 1,2,3 \ldots \max lag = 12$$

Gde je $P' = \frac{\sum_{i=1}^{N} P_i}{\mathrm{N}}$, a P_i, P_{i+d}, d i N imaju isto značenje kao i za *Moreau-Broto* autokorelacioni deskriptor.

- Skupovi atributa F7 i F8 predstavljaju *Geary* autokorelacione deskriptore. F7 je konstruisan na osnovu fizičko-hemijskih osobina korišćenih za F1, a F8 na osnovu pet multidimenzionih obrazaca korišćenih za konstrukciju F2. Pre konstrukcije F7 i F8 su atributi aminokiselina standardizovani tako da im srednja vrednost bude 0, a standardna devijacija 1. Ovaj autokorelacini deskriptor se računa na sledeći način:

$$C(d) = \frac{\frac{1}{2(N-d)}\sum_{i=1}^{N-d}(P_i - P_{i+d})^2}{\frac{1}{N-1}\sum_{i=1}^{N}(P_i - P')^2} \quad d = 1,2,3 \ldots \max lag = 12$$

Gde je $P' = \frac{\sum_{i=1}^{N} P_i}{\mathrm{N}}$, a P_i, P_{i+d}, d i N imaju isto značenje kao i za *Moreau-Broto* autokorelacioni deskriptor.

Dimenzije pomenutih autokorelacionih deskriptora atributa jednake su proizvodu broja atributa i 12 (odabrani max *lag* parametar).

- Skupovi atributa F9 i F10 predstavljaju deskriptore redosleda sekvence (engl. „*sequence-order descriptors*“) koji su izračunati na osnovu dve matrice distanci između aminokislina: *Schneider-Wrede* (Schneider i Wrede, 1994) i *Grantham* (Grantham, 1974). Kuplujući broj redosleda sekvence (engl. „*sequence-order-coupling number*“) – F9 (Chou, 2000) izračunat je na sledeći način:

$$\tau_d = \sum_{i=1}^{N-d} \left(ds_{i,i+d}\right)^2$$

Gde je $ds_{i,i+d}$ distanca na osnovu odgovarajuće matrice distanci (*Schneider-Wrede* ili *Grantham*) između aminokiselina u pozicijama *i* i $i+d$, a *d* je maksimalna vrednost laga koja je izabrana da bude 12. Broj dimenzija je jednak proizvodu broja korišćenih matrica distanci i 12 (max *lag*), dakle 24 u ovom slučaju.

- Kvazi deskriptor redosleda sekvence (engl. „*quasi sequence-order descriptors*“) – F10 (Chou, 2000) se sastoji iz dve komponente X_r i X_d. X_r se računa za svaku aminokiselinu:

$$X_r = \frac{f_r}{\sum_{r=1}^{20} f_r + \omega \sum_{d=1}^{maxlag} \tau_d}$$

Gde je r = 1, 2 … 20, $f_r\, je$ normalizovana učestalost aminokiseline *r*, τ_d je kuplujući broj redosleda sekvence (F9) definisan iznad, ω je skalirajuća konstanta koja je iznosila 0,1. Vrednost $maxlag$ je izabrana da bude 12. Broj dimenzija je jednak proizvodu broja korišćenih matrica (2) i ukupnog broja različitih aminokiselina (20), dakle 40.

X_d je izračunat na sledeći način:

$$X_d = \frac{\omega \tau_{d-20}}{\sum_{r=1}^{20} f_r + \omega \sum_{d=1}^{maxlag} \tau_d}$$

Gde je *d* = 21, 22 …20 + 12 (*maxlag*), dok su ostali parametri definisani iznad. Broj dimenzija je jednak proizvodu broja korišćenih matrica (2) i 12 (*maxlag*), dakle 24.

- Deskriptor združenih trijada (engl. *„conjoint triad descriptor“*) - F11 je inicijalno korišćen za predviđanje protein-protein interakcija na osnovu klasifikacije aminokiselina (Shen i sar., 2007). Ovaj deskriptor predstavlja „kompresiju“ frekvence svih mogućih tripeptida u proteinskoj sekvenci za čije je predstavljanje potreban vektorski prostor od 8000 dimenzija (20 × 20 × 20) u vektorski prostor od 343 (7 × 7 × 7) dimenzije. Ovo se postiže grupisanjem aminokiselina u sedam klasa prema dipolu i zapremini bočnih lanaca: klasa I: Ala, Gly, Val; klasa II: Ile, Leu, Phe, Pro; klasa III: Tyr, Met, Thr, Ser; klasa IV: His, Asn, Gln, Tpr; klasa V: Arg, Lys; klasa VI: Asp, Glu i klasa VII: Cys. Tako su na primer trijadom I II III predstavljeni tripeptidi AIT, GIT, VIT i druge kombinacije aminkiselina iz ove tri klase. Da bi se omogućilo svrsishodnije poređenje proteina različitih dužine frekvenca svih mogućih trijada je normalizovana jednačinom:

$$d_i = \frac{f_i - \min\{f_1, f_2, \dots, f_{343}\}}{\max\{f_1, f_2, \dots, f_{343}\}}$$

  Gde je f_i učestalost trijade *i* (*i* = 1, 2, ..., 343). d_i se prema tome kreće od 0 – 1.

- Pseudo-aminokiselinski sastav (PseAAC, engl. *„pseudo-amino acid composition“*) – F12 (Chou, 2001) predstavlja autokorelaciju srednje vrednosti tri aminokiselinske odlike: hidrofobnost, hidrofilnost i masa bočnog niza[5]. Dimenzije ovog deskirptora zavise od λ parametra koji definiše kolika je udaljenost aminokiselina u sekvenci za koje je definisana autkorelacija, koji je izabran da bude 20. Dimenzije ovog deskriptora su 20 fiksnih dimenzija (za svaki tip aminokiseline po jedna) + λ, što u ovom slučaju iznosi 40.
- Amfifilni pseudo-aminokiselinski sastav (APseAAC, engl. *„amphiphilic pseudo-amino Acid Composition*) – F13 (Chou, 2001) je sličan F12 sa tom razlikom da ne uzima u obzir masu bočnog niza, a autokorelacija kroz sekvencu je data posebno za hidrofobnost i hidrofilnost[6]. Dimenzije ovog deskirptora zavise od λ parametra koji definiše kolika je udaljenost aminokiselina u sekvenci za koje je definisana autokorelacija, koji je izabran da bude 20. Pri λ = 20 dimenzije ovog deskriptora su 20 fiksnih dimenzija (za svaki tip aminokiseline po jedna) + 2λ, što u ovom slučaju iznosi 60.

Kompozicioni deskriptor (engl. *„composition“*) – F14, tranzicioni deskriptor (engl. *„transition“*) – F15 i distribucioni deskriptor (engl. *„distribution“*) – F16 razvili su Dubchak i sar. (1995). Ovi deskriptori podrazumevaju grupisanje aminokiselina u po tri klase prema sedam odlika (Tabela 3).

- Kompozicioni deskriptor (F14) je definisan kao udeo svake klase (Tabela 3) u proteinskoj sekvenci. Pošto ima sedam odlika sa po tri grupe aminokiselina ovaj deskriptor ima 21 dimenziju.
- Tranzicioni deskriptor (F15) opisuje dipeptide i definisan je kao udeo dipeptida u čiji sastav ulaze različite klase aminokiselina u okviru svake odlike, pri čemu redosled klasa u dipeptidu nije od značaja. Za svaku od odlika dipeptid može biti sačinjen od klasa 12, 13 i 23 (ekvivalentno sa 21, 31 i 32 pošto redosled nije značajan). Pošto ima sedam odlika aminokiselina, a svaki dipeptid može biti kodiran sa tri kombinacije ovaj deskriptor ima 21 dimenziju.
- Distribucioni deskriptor (F16) opisuje distribuciju aminokiselina za svaku od odlika aminokiselina (Tabela 3). U okviru jedne odlike za svaku od klasa računa se relativna pozicija u sekvenci (količnik pozicije i dužine sekvence) za prvu instancu kada se javlja data klasa, 25% instancu, 50% instancu, 75% instancu i poslednju instancu date klase u sekvenci. Pošto ima sedam odlika sa po tri klase koje se računaju u pet instanci broj dimenzija ovog deskriptora je 105.

[5] Detalji nisu dati pošto bi to zahtevalo višestruke formule i objašnjenja, što nije potrebno za dalje razumevanje teksta
[6] Detalji nisu dati pošto bi to zahtevalo višestruke formule i objašnjenja, što nije potrebno za dalje razumevanje teksta

Tabela 3. Grupisanje aminokiselina u tri klase prema sedam odlika za formiranje F14-F16 deskriptora (Dubchak i sar., 1995).

odlika	klasa 1	klasa 2	klasa 3
hidrofobnost	polarne	neutralne	hidrofobne
	R, K, E, D, Q, N	G, A, S, T, P, H, Y	C, L, V, I, M, F,
normalizovani van der Waals volumen	0-2,78	2,95-4,0	4,03-8,08
	G, A, S, T, P, D, C	N, V, E, Q, I, L	M, H, K, F, R, Y,
polarnost	4,9-6,2	8,0-9,2	10,4-13,0
	L, I, F, W, C, M, V,	P, A, T, G, S	H, Q, R, K, N, E,
polarizabilnost	0-1,08	0,128-0,186	0,219-0,409
	G, A, S, D, T	C, P, N, V, E, Q, I,	K, M, H, F, R, Y,
naboj	pozitivan	neutralan	negativan
	K, R	A, N, C, Q, G, H, I, L, M, F, P, S, T, W, Y, V	D, E
sekundarna struktura	α-heliks	β-ravan	nasumično klupko
	E, A, L, M, Q, K, R,	V, I, Y, C, W, F, T	G, N, P, S, D
pristupačnost	slaba	visoka	srednja
	A, L, F, C, G, I, V,	R, K, Q, E, N, D	M, S, P, T, H, Y

Tabela 4. Skupovi atributa korišćeni za treniranje modela MU za predviđanje hidroksilacije prolina na osnovu lokalne sekvence aminokiselina.

grupa atributa	opis	lag	broj dimenzija	atributi aminokiselina	referenca
F1	Slikanje proteinske sekvence u sekvencu numeričkih vrednosti na osnovu fizičko-hemijskih osobina aminokiselina	/	120	CIDH920105, BHAR880101, CHAM820102, BIGC670101, CHAM810101, DAYM780201	Kawashima i Kanehisa (2000)
F2	Slikanje proteinske sekvence u sekvencu numeričkih vrednosti na osnovu multidimenzionalnih obrazaca aminokiselina	/	100	Factor I–Factor V	Atchley i sar. (2005)
F3	*Moreau-Broto* autokorelacioni deskriptor	12	72	CIDH920105, BHAR880101, CHAM820102, BIGC670101, CHAM810101, DAYM780201	Kawashima i Kanehisa (2000)
F4	*Moreau-Broto* autokorelacioni deskriptor	12	60	Factor I–Factor V	Atchley i sar. (2005)
F5	*Moran* autokorelacioni deskriptor	12	72	CIDH920105, BHAR880101, CHAM820102, BIGC670101, CHAM810101, DAYM780201	Kawashima i Kanehisa (2000)
F6	*Moran* autokorelacioni deskriptor	12	60	Factor I–Factor V	Atchley i sar. (2005)
F7	*Geary* autokorelacioni deskriptor	12	72	CIDH920105, BHAR880101, CHAM820102, BIGC670101, CHAM810101, DAYM780201	Kawashima i Kanehisa (2000)
F8	*Geary* autokorelacioni deskriptor	12	60	Factor I–Factor V	Atchley i sar. (2005)
F9	Kuplujući broj redosleda sekvence	12	24	*Schneider-Wrede* i *Grantham* matrice distanci između aminokiselina	Grantham (1974); Schneider i Wrede (1994); Chou (2000)
F10	Kvazi deskriptor redosleda sekvence	12	64	*Schneider-Wrede* i *Grantham* matrice distanci između aminokiselina	Grantham (1974); Schneider i Wrede (1994); Chou (2000)
F11	Deskriptor združenih trijada	/	343		Shen i sar. (2007)
F12	Pseudo-amino kiselinski sastav	/	40		Chou (2001)
F13	Amfifilni pseudo-amino kiselinski sastav	/	60		Chou (2005); Xiao i sar. (2015)
F14	Kompozicioni deskriptor	/	21		Dubchak i sar. (1995)
F15	Tranzicioni deskriptor	/	21		Dubchak i sar. (1995)
F16	Distribcioni deskriptor	/	105		Dubchak i sar. (1995)

Skupovi atributa F3-16 konstruisani su korišćenjem R paketa *protr* (Xiao i sar., 2015), dok su F1 i F2 konstruisani korišćenjem koda razvijenog u okviru ove disertacije. Pregled deskriptora je dat u tabeli 4.

Odabir atributa

Konstruisani skupovi atributa F1 - F16 (Tabela 4**)** su zbirno imali 1294 dimenzija - različitih brojčanih vrednosti koje opisuju svaku 21-mernu sekvencu. Mnogi od ovih atributa su međusobno korelisani, ili pak ne nose nikakvu upotrebljivu informaciju o hidroksilaciji prolina i samim tim ometaju pronalaženje modela sa maksimalnim performansama. U cilju odabira podskupa optimalnih atributa upoređena su tri načina njihovog odabira, dva bazirana na principu filter algoritama: udeo dobitka informacije (IGr, engl. „*Information gain ratio*“) (Quinlan, 1986) i minimalna suvišnost maksimalna značajnost (mRMR, engl. „*Minimum Redundancy Maximum Relevance*“) (Peng i sar., 2005; Zhao i sar., 2019) i jedan baziran na principu omotač algoritama za odabir atributa (engl: „*wrapper*“) – sekvencionoj pretrazi unapred (sfs, engl. „*sequential forward selection*“, Kohavi i John, 1997).

Filter odabir atributa funkcioniše tako što se atributi poređaju po redosledu njihove asocijacije sa zavisnom promenljivom koja je u ovom slučaju hidroksilacija prolina (1 ili 0) prema nekom kriterijumu (IGr i mRMR ovde korišćeni), a zatim se odabere određen broj najbolje rangiranih za dalje treniranje algoritma MU. Da bi se odabrao optimalan broj najboljih atributa pomenutim filter metodama, apsolutan broj izabranih atributa je podešavan zajedno sa drugim hiperparametrima algoritama u opsegu od 20 do 700.

Omotač algoritmi za odabir atributa funkcionišu tako što se performanse modela MU ispitaju kada se on trenira sa različitim podskupovima atributa i odabira se onaj podskup koji pokazuje najbolje performanse, najčešće u unakrsnoj validaciji. Ovi algoritmi se razlikuju prema načinu odabira podskupova atributa koji će se ispitati. Kod sekvencione pretrage unapred prvo se odrade performanse modela u unakrsnoj validaciji sa svakim pojedinačnim atributom i odabere se atribut koji daje najbolje performanse. Zatim se datom atributu doda po jedan od svih preostalih i odredi se kombinacija atributa koja daje najbolje performanse u unakrsnoj validaciji. Ovo se nastavlja dok performanse modela ne počnu da opadaju (prestanu da rastu) dodatkom novih atributa. Jasno je da ovaj proces računarski jako zahtevan pošto podrazumeva treniranje jako velikog broja modela. Da bi se ova zahtevnost smanjila i dovela na izvodljiv nivo, u ovom slučaju nije evaluiran svaki atribut posebno sekvencionom pretragom unapred već svaka grupa atributa (F1 – F16).

Treniranje modela

Nekoliko algoritama mašinskog učenja je evaluirano za tačno predviđanje hidroksilacije Pro: RF , XGB , SVM uz kernel sa radijalnom osnovom i KNN (Tabela 5).

Tabela 5. Algoritmi korišćeni za treniranje modela MU za predviđanje pozicija Hyp na osnovu lokalne sekvence.

Algoritam MU	Paket
RF	R paket *ranger* (Wright i Ziegler, 2017)
XGB	R paket *xgboost* (Chen i sar., 2019)
SVM uz kernel sa radijalnom osnovom	R paket *e1071* (Meyer i sar., 2018)
KNN	R paket *kknn* (Schliep i Hechenbichler, 2016)

Treniranje algoritama je izvedeno korišćenjem *mlr* paketa (Bischl i sar., 2016). Optimalni hiperparametri su određeni Bajesovskom optimizacijom pomoću R paketa *mlrMBO* (Bischl i sar., 2017) kroz 100 iteracija unakrsne validacije uz podrazumevana podešavanja. Za svaki algoritam optimizovan je određen skup hiperparametara dat u Tabeli 6. Kako bi se umanjio uticaj neizbalansiranosti klasa na performanse modela usled različite zastupljenosti Hyp i Pro u skupu za treniranje modela, za svaki algoritam (sem za KNN koji ne podržava ovu opciju) podešen je odnos značajnosti klasa (engl. „*class weight*“). Značajnost pozitivne klase (Hyp) je podešena da bude jednaka odnosu Pro/Hyp u skupu za treniranje (737/150 = 4.9133). Kao kriterijum za odabir hiperparametara korišćena je prosečna AUC u unakrsnoj validaciji. Detalji procene performansi modela su dati u sledećem odeljku.

Tabela 6. Hiperparametri koji su optimizovani za svaki algoritam MU.

algoritam	hiperparametar	opis hiperparametra	opseg optimizacije
RF	*ntree*	broj stabala odlučivanja koja čine ansambl	50 – 2000
	sample.fraction	udeo opservacija korišćenih za treniranje svakog pojedinačnog stabla odlučivanja	0,1 – 1
	mtry	broj atributa nasumično odabranih za svaki čvor	1 – 20
SVM	*cost*	regularizaciona konstanta C	10^{-6} – 10^{6}
	gamma	širina radijalnog kernela	10^{-6} – 10^{6}
XGB	*nrounds*	broj stabala odlučivanja u modelu	50 – 2000
	eta	skupljanje po koraku (engl. „*step-size shrinkage*“) - definiše tempo učenja ansambla	0,005 – 0,2
	max_depth	maksimalna dubina pojedinačnog stabla odlučivanja	3 – 15
	colsample_bytree	udeo atributa korišćenih za pojedinačno stablo odlučivanja	0,3 – 1
	colsample_bylevel	udeo atributa korišćenih za pojedinačnu podelu (čvor) unutar stabla odlučivanja	0,3 - 1
	subsample	udeo opservacija korišćenih za treniranje svakog pojedinačnog stabla odlučivanja	0,3 – 1
	alpha, gamma, lambda	regularizacioni parametri koji utiču na konzervativnost algoritma	0 – 3
	min_child_weight	minimalna suma Hesijan matrice potrebna za dalju podelu unutar stabla odlučivanja. Parametar koji kontroliše konzervativnost podele unutar pojedinačnog stabla odlučivanja	1 – 10
KNN	*k*	broj susednih opservacija koja se uzimaju u obzir	1 – 50
	distance	parametar koji definiše Minkowski distancu	0,5 – 8

Procena performansi modela

Performanse modela procenjene su korišćenjem unakrsne validacije u dva sloja. Unutrašnji sloj korišćen je za podešavanje hiperparametara i sastojao se od dva puta ponovljene unakrsne validacije sa tri podele, dok je spoljašnji sloj korišćen za procenu performansi i sastojao se od unakrsne validacije sa deset podela. U slučaju sekvencione pretrage unapred, spoljašnji sloj unakrsne validacije korišćen je i za evaluaciju i odabir različitih kombinacija skupova atributa (F1 – F16). Za oba sloja korišćena je blok unakrsna validacija što podrazumeva da sve 21-mere koje potiču od iste proteinske sekvence idu zajedno u bloku, odnosno da se u jednoj instanci unakrsne validacije sve 21-mere poreklom od iste proteinske sekvence koriste ili za pravljenje modela ili za procenu njegovih performansi. Performanse modela merene su korišćenjem AUC metrike.

Procena praga odluke

Pošto u skupu podataka za treniranje postoji disbalans u frekvencijama klasa (prolin je oko pet puta frekventniji od hidroksiprolina) pristupljeno je optimizaciji praga odluke. Nakon treniranja i evaluacije modela, za algoritam/metodologiju MU koji su pokazali najbolje performanse urađena je procena uticaja praga odluke na nekoliko metrika koje se koriste za kvantifikaciju performansi modela: osetljivost, specifičnost, balansirana tačnost, *Cohen*-ov *kappa* koeficijent (Cohen, 1960) i *Matthews* koeficijent korelacije (Matthews, 1975). Procena uticaja praga odluke na pomenute metrike urađena je unakrsnom validacijom u dva sloja, gde se spoljašnji sloj koji je služio za ispitivanje pomenutog uticaja sastojao od tri puta ponovljene unakrsne validacije sa deset podela, dok se unutrašnji sloj koji je služio za optimizaciju hiperparametara po ranije opisanom principu sastojao od dva puta ponovljene unakrsne validacije sa tri podele. Kao optimalni prag odluke izabrana je verovatnoća koja maksimizuje balansiranu tačnost u spoljašnjem sloju unakrsne validacije.

Algoritam/metodologija koja je pokazala najbolje performanse obučavana je na celom skupu podataka za treniranje korišćenjem hiperparametara pronađenih kroz 100 iteracija Bajesovske optimizacije upotrebom dva puta ponovljene trostruke unakrsne validacije. Tako dobijeni model je evaluiran na skupu podataka za evaluaciju korišćenjem prethodno odabranog praga odluke.

Uticaj dužine lokalne sekvence na performanse modela

Da bi se ispitao uticaj dužine lokalne sekvence na performanse modela, 21-merne sekvence su smanjenje na 19 (po 9 aminokiselina sa obe strane prolina od interesa), 17, 15 i 13 aminokiselina uklanjanjem po jedne aminokiseline sa krajeva. Smanjenje redundantnosti urađeno je kao što je opisano i za 21-merni skup podataka.

Skupovi varijabli i algoritam MU koji su imali najbolje performanse sa 21-mernim skupom podataka korišćeni su za pravljenje modela na drugim dužinama k-mera. Optimizacija hiperparametara, procena performansi modela i uticaja praga odluke, kao i finalna optimizacija i treniranje modela urađeni su kao što je prethodno opisano za 21-merni skup podataka.

Interpretacija predviđanja modela

Interpretacija predviđanja modela treniranog na osnovu 21-mernih sekvenci je urađena ispitivanjem značajnosti atributa za predviđanja modela:

- sumiranjem uticaja svakog od korišćenih atributa na povećanje homogenosti u okviru svakog čvora stabala odlučivanja unutar ansambla pojačavanja. Ovo predstavlja metriku inherentnu za sam XGB algoritam i ostale ansamble koji se sastoje od stabala odlučivanja.
- Permutacionom analizom, gde se svaki pojedinačni atribut permutuje čime se remeti njegova asocijacija za zavisnom promenljivom, nakon čega se procenjuje gubitak performansi modela na ovakvom skupu podatka. Permutacije svakog atributa pojedinačno ponovljene su 10 puta, a AUC je korišćen za evaluaciju performansi. Ovako dobijena značajnost atributa je potpuno nezavisna od modela i može se upotrebiti na praktično sve tipove modela. Za permutacionu analizu je korišćen *DALEX* R paket (Biecek, 2018).
- Permutacionom analizom analognom gore opisanoj gde su umesto pojedinačnih atributa permutovane cele grupe atributa (deskriptori) koje su korišćene za treniranje finalnog modela.

Raspodela najznačajnijih atributa prema gore opisanim kriterijumima je ispitana u zavisnosti od statusa hidroksilacije prolina u okviru svake 21-mere na skupu podataka za treniranje. Pored toga, uticaj najznačajnijih atributa na predviđanja modela je ispitan graficima koji opisuju zavisnost parcijalnog uticaja (pdp – engl. „*partial dependence plot*“) atributa na predviđanja verovatnoće pripadnosti pozitivnoj klasi (Friedman, 2001). Ovaj tip analize izoluje uticaj jednog atributa na ponašanje modela. Za dalje ispitivanja uticajnih atributa koji pripadaju klasi autokorelacionih deskriptora urađena je simulacija milion nasumičnih 21-mernih sekvenci sa P u sredini i ispitan je aminokiselinski sastav u njima u zavisnosti od vrednosti ispitivanog autokorelacionog deskriptora.

3.1.2. Pravljenje R paketa za identifikaciju i analizu HRGP sekvenci

Model za predviđanje hidroksilacije prolina na osnovu lokalne sekvence je inkorporiran u R funkciju *predict_hyp* u R paketu *ragp* (https://github.com/missuse/ragp). Uputstva za postupak rada sa paketom se nalaze na https://missuse.github.io/ragp, a web server za predviđanje hidroksilacije prolina na osnovu lokalne sekvence se nalazi na adresi https://ragp.shinyapps.io/Rapp. Osim modela za predviđanje Pro hidroksilacije u paket su inkorporirane funkcije za efikasnu komunikaciju sa Internet serverima za detekciju domena, predviđanje N-sp, GPI i neuređenih regiona u proteinima kao i funkcije za detekciju i analizu AG karakterističnih motiva. Sve funkcije paketa su dokumentovane na adresi https://missuse.github.io/ragp/reference/index.html.

Komunikacija sa eksternim web serverima

Predviđanje prisustva N-sp je u *ragp* paketu postignuto efikasnom komunikacijom sa Internet serverima: *TargetP* (http://www.cbs.dtu.dk/services/TargetP/) (Emanuelsson i sar., 2007), *SignalP* (http:/www.cbs.dtu.dk/services/SignalP-4.1/) (Petersen i sar., 2011) i *Phobius* (http://phobius.sbc.su.se/) (Käll i sar., 2007) korišćenjem funkcija *get_targetp*, *get_signalp* i *get_phobius*. Predviđanje transmembranskih regiona (TM) omogućeno je korišćenjem *Phobius* servera. Predviđanje pozicija GPI sidra (omega mesta) postignuto je komunikacijom sa Internet serverima *big Pi plant predictor* (Eisenhaber i sar., 2003), *PredGPI* (Pierleoni i sar., 2008) i *NetGPI* (Gíslason i sar., 2021) korišćenjem funkcija *get_big_pi*, *get_pred_gpi* i *get_netGPI*.

Navedene funkcije koriste R pakete *httr* (Wickham i sar., 2018a) i *xml2* (Wickham i sar., 2018b) za komunikaciju sa odgovarajućim serverima.

Anotacija domena u *ragp* paketu postignuta je korišćenjem funkcije *get_hmm* koja komunicira sa *hmmscan* Internet serverom (Finn i sar., 2011, https://www.ebi.ac.uk/Tools/hmmer/search/hmmscan) i *pfam2go* koja preslikava *Pfam* anotacije u GO (engl. *„gene onthology“*) korišćenjem http://geneontology.org/external2go/pfam2gomapping. Osim toga omogućena je i anotacija pomoću baze konzerviranih domena (CDD, engl. *„conserved domain database“*, Lu i sar., 2020) preko funkcije *get_cdd*. Za predviđanje neuređenih delova proteina *ragp* sadrži funkciju *get_espritz* koja komunicira sa *Espritz* serverom (Walsh i sar., 2012). Ove funkcije takođe koriste *httr* i *xml2* R pakete.

Analiza motiva u HRGP sekvencama

Analiza motiva karakterističnih za HRGP u proteinskim sekvencama je u *ragp* paketu postignuta kroz dve funkcije *maab* i *scan_ag*. U okviru funkcije *maab* inkorporirana je MAAB klasifikaciona šema (Johnson i sar., 2017a) koja klasifikuje HRGP sekvence u 23 deskriptivne klase. Funkcija *scan_ag* služi za detekciju lokalizovanih klastera AG motiva korišćenjem regularnih izraza. Korisnik unosi minimalan broj motiva koji su potrebni i maksimalno dozvoljeno rastojanje među njima, a funkcija konstruiše regularni izraz kojim obavlja detekciju u datim sekvencama. Ova funkcija može da uvrsti znanje o pozicijama Hyp u sekvencama korišćenjem rezultata *predict_hyp* funkcije, kao i da maskira određene tipove nepoželjnih motiva kao što su ekstenzinski.

3.1.3 Anotacija HRGP sekvenci iz 62 biljna proteoma

Sa ciljem demonstracije mogućnosti *ragp* paketa izabrana su 62 biljna proteoma iz Phytozome V12 baze (https://phytozome.jgi.doe.gov/pz/portal.html, PhytozomeV12_unrestricted) za anotaciju HRGP sekvenci. Prvi korak u *ragp* metodologiji je filtriranje sekvenci koje sadrže N-terminalni signalni peptid. Za ovo je korišćen većinski glas predviđanja *Phobius*, *SignalP4.1* i *TargetP1.1* servera kojima je pristupljeno *ragp* funkcijama. Sekvence za koje je utvrđeno da verovatno sadrže N-sp ovom metodom su u sledećem koraku evaluirane za prisustvo hidroksiprolina, i filtrirane su sekvence koje sadrže barem tri predviđena hidroksiprolina. Ove sekvence su dalje klasifikovane u MAAB klase, a AGP sekvence su identifikovane detekcijom AG motiva. Kriterijum za nalaženje AG motiva bilo je prisustvo najmanje tri dipeptida: AO, SO, TO, GO, VO, OA, OS, OT, OG i OV koji su međusobno razdvojeni sa ne više od 10 aminokiselina. Za potrebe brojanja detekcije motiva, ekstenzinski motivi (niz od najmanje 3 P ili O) su bili maskirani. Domeni prisutni u sekvencama koje sadrže AGP motive su detektovani pomoću *hmmer3 3.1b2* (Eddy, 2011) korišćenjem *Pfam* 32 baze podataka (https://ftp.ebi.ac.uk/pub/databases/Pfam/releases/Pfam32.0/). Anotacije su dostupne na Zenodo paltformi (doi: 10.5281/zenodo.2605302, url: https://zenodo.org/record/2605302).

3.2. Identifikacija AGP sekvenci kičice

Identifikacija HRGP sekvenci *C. erythraea* urađena je korišćenjem *ragp* 0.3.2 R paketa. Kao polazna tačka korišćene su *in silico* translirane proteinske sekvence iz *de novo* sastavljenog transkiptoma kičice (Ćuković i sar., 2020). Proteinske sekvence su prvo filtrirane na osnovu prisustva sekretornih signala. Kao kriterijum da li sekvenca sadrži N-sp korišćen je većinski broj glasova (dva ili više) servera za predviđanje N-sp: *TargetP1.1*(Emanuelsson i sar., 2007), *SignalP4.1* (Petersen i sar., 2011), i *Phobius1.01* (Käll i sar., 2007). Sekvence za koje je na ovaj način određeno da sadrže sekretorni signal na N-terminusu su dalje korišćene za predviđanje pozicija hidroksiprolina. Sekvence koje su sadržale tri ili više predviđenih hidroksiprolina su dalje analizirane na prisustvo signala za dodatak GPI, AGP motiva i klasifikovane prema MAAB metodi (Johnson i sar., 2017a). Prisustvo GPI signala potrebno za MAAB klasifikaciju određeno je korišćenjem servera *NetGPI1.1* (Gíslason i sar., 2021) preko *ragp* R paketa.

Za identifikaciju potencijalnih arabinogalaktanskih sekvenci urađena je relaksirana pretraga sekvenci za regionima koji sadrže najmanje tri AG motiva, sa ne više od 10 aminokiselina između njih (kao što je opisano u odeljku 3.1.3.) i stroga pretraga za regionima koji sadrže najmanje četiri AG motiva sa ne više od četiri aminokiseline između, a ekstenzinski motivi (niz od najmanje 3 P ili O) su u oba slučaja maskirani. Razmatrani su samo AG motivi sa predviđenim hidroksiprolinima. Izabrane sekvence su zatim analizirane korišćenjem *NetGPI 1.1* (Gíslason i sar., 2021) za predviđanje GPI, a anotacija domena u sekvencama urađena je sa *hmmscan 3.3.2* (Eddy, 2011) korišćenjem *Pfam*33 baze podataka (https://ftp.ebi.ac.uk/pub/databases/Pfam/releases/Pfam33.0). Svim pomenutim Internet serverima pristupano je korišćenjem *ragp* paketa.

Nakon identifikacije AGP sekvenci izabrano je 18 predstavnika za dalju procenu ekspresije. Sekvence su odabrane na osnovu primarne strukture i mogućnosti da se za njih naprave odgovarajući prajmeri: od toga 6 sekvenci sa fasciklinskim domenima (FLA), 6 sa protein kinaznim domenima (engl. *„kinase like AGPs*“ – KLA) i 6 kratkih sekvenci – AG-peptida (AGp).

3.3. Filogenetska analiza AGP sekvenci kičice

Filogenija odabranih AGP sekvenci *C. erythraea* procenjena je poređenjem sa homolozima iz 18 biljnih vrsta: *Arabidopsis thaliana, Beta vulgaris, Capsicum annuum, Coffea canephora, Cynara cardunculus, Daucus carota, Glycine max, Helianthus annuus, Hordeum vulgare, Ipomoea triloba, Nicotiana attenuata, Olea europaea var. sylvestris, Oryza sativa Japonica, Populus trichocarpa, Prunus avium, Solanum lycopersicum, Triticum aestivum* i *Zea mays.* Ovaj skup čine biljne vrste koje su dobro proučeni model organizmi, zatim komercijalno značajne vrste kao i vrste taksonomski bliske kičici. Proteomi izabranih biljnih vrsta preuzeti su iz Ensembl baze verzija 49 (https://plants.ensembl.org/, http://ftp.ensemblgenomes.org/pub/plants/release-49/fasta/, pristupljeno: 12.02. 2021).

Za određivanje filogenije FLA i KLA sekvenci, prvo su korišćenjem *biomartr* paketa (Drost i Paszkowski, 2017) izdvojene sekvence sa *Pfam* domenima PF02469 (fasciklinski domen) i PF07714 (tirozin i serin/treonin kinazni domen) iz proteoma izabranih biljnih vrsta, a zatim je

urađen proteinski *blast* sa izabranim FLA i KLA sekvencama kičice. Za svaku sekvencu kičice izdvojeno je pet najsličnijih homologa (prema *raw bitscore blast* metrici) iz različitih biljaka i ove sekvence su korišćene za filogenetsku analizu.

Višestruko poravnavanje sekvenci urađeno je korišćenjem *DECIPHER* 2.18.1 R paketa (Wright, 2015). Filogenetska stabla su određena procenom maksimalne verovatnoće (engl. „*maximum likelihood*“) preko *phangorn* R paketa 2.5.5 (Schliep, 2010). Stabilnost klastera je procenjena neparametrijskim *bootstrap*-om od 100 iteracija.

Filogenija AG peptida određena je korišćenjem pristupa bez poravnjanja sekvenci pošto su ove sekvence kratke i imaju nizak stepen međusobne sličnosti. Za filogenetsku analizu iskorišćene su proteinske sekvence dužine 30-100 aminokiselina iz odabranih 18 biljnih vrsta. AG peptidi u pomenutoj grupi su identifikovani kao što je opisano u odeljku 3.1.3. Identifikovane sekvence AG peptida iz odabranih biljnih vrsta, zajedno sa AGp sekvencama kičice su iskorišćene za procenu matrice udaljenosti sekvenci korišćenjem *kmer* R paketa (Wilkinson, 2018), a korišćena je dužina k-mera 3. Za svaku AGp sekvencu kičice filtrirano je pet najsličnijih homologa (najmanja distanca u matrici) iz različitih biljaka i ove sekvence su iskorišćene za konstrukciju filogenetskog stabla metodom pridruživanja suseda (NJ, engl. „*neighbour joining*“) na osnovu matrice udaljenosti dobijene *kmer* R paketom sa dužinom k-mera 3. Neparametrijski *bootstrap* od 100 iteracija je primenjen za procenu stabilnosti klastera.

Filogenetski klasteri sa manje od 50/100 *bootstrap* podrške su spojeni u politomije.

3.4. Uslovi gajenja biljaka za ispitivanje ekspresije odabranih *AGP* kičice

3.4.1. Uspostavljanje *in vitro* kulture korenova i izdanaka kičice (*Centaurium erythraea*)

Kao početni materijal za uspostavljanje *in vitro* kultura korišćena su komercijalna semena *Centaurim erythraea* Rafn. (Jelitto Staudensamen GmbH, Schwarmstedt, Germany).

Semena su površinski sterilisana 20% vodenim rastvorom komercijalnog preparata natrijum hipohlorita sa 4% aktivnog hlora u aseptičnim uslovima u trajanju od 10 minuta, ispirana 3 puta po 5 minuta sterilnom dejonizovanom vodom, a zatim postavljana na hranljivu podlogu za isklijavanje u petri kutije. Podloga (Tabela 7) se sastojala se od ½MS (Murashige i Skoog, 1962) sa dodatkom saharoze u koncentraciji 30 gL^{-1}, mioinozitola (Sigma) 100 mgL^{-1} i agara (Torlak) 7gL^{-1}. pH vrednost hranljive podloge podešena je na 5,8 pre sterilizacije pomoću 1N NaOH i/ili HCl. Isklijavanje se odigravalo u klimatizovanoj prostoriji u kontrolisanim uslovima, na temperaturi od 25 ± 2 °C, u uslovima dugog dana (fotoperiod od 16 h svetlosti i 8 h mraka). Kao izvor svetosti korišćene su bele flourescentne lampe (Tesla, Pančevo) jačine 65 W. Gustina svetlosnog fluksa iznosila je 47 $\mu mols^{-1}m^{-2}$ i izmerena je instrumentom LI-1400 DataLogger sa LI-190SA quantum senzorom (*LI-COR Biosciences*, Bad Homburg Germany).

Klijanci razvijeni iz proklijalih semena stari tri nedelje su prebacivani na svežu hranljivu podlogu istog sastava kao podloga za isklijavanje i dalje gajeni na njoj tokom tri meseca u istim uslovima kao i tokom isklijavanja.

Tabela 7. Sastav upola razblažene hranljive MS podloge (½MS) (Murashige i Skoog, 1962).

Makro mineralne soli	**mgL^{-1}**
NH_4NO_3	825
KNO_3	950
KH_2PO_4	220
$MgSO_4$ x 7 H_2O	185
KH_2PO_4	85
Mikro mineralne soli	**mgL^{-1}**
$MnSO_4$ x 4 H_2O	11,15
$ZnSO_4$ x 7 H_2O	4,3
H_3BO_3	3,1
KJ	0,415
$NaMoO_4$ x 2 H_2O	0,125
$CuSO_4$ x 5 H_2O	0,125
$CoCl_2$ x 6 H_2O	0,125
Kompleks gvožđa	**mgL^{-1}**
NaEDTA x 2 H_2O	18,6
$FeSO_4$ x 7 H_2O	13,9
Vitaminski kompleks	**mgL^{-1}**
Vitamin B_1	0,1
Vitamin B_6	0,5
Nikotinska kiselna	0,5
Glicin	2,0

3.4.2. Ispitivanje uticaja mehaničke povrede na ekspresiju *AGP* gena kičice gajene u uslovima *in vitro*

Biljke, na kojima je rađeno ispitivanje uticaja mehaničke povrede na ekspresiju *AGP* gena, gajene su na ½MS hranljivoj podlozi (Tabela 7).

Nakon isklijavanja, po osam klijanaca dužine 15-20 mm, prebačeno je na svežu ½ MS hranljivu podlogu, na kojoj su gajeni naredna tri meseca u istim uslovima kao tokom isklijavanja (odeljak 3.4.1.). Nakon tri meseca sa biljaka su uzimani isečci vrhova listova u fazi rozete (isečci dužine oko 5 mm sa po dva para listova najbližih apeksu) i isečci vrhova korenova sa apikalnim meristemom, dužine oko 10 mm. Eksplantati su prebačeni u petri kutije na svežu ½MS hranljivu podlogu na kojoj su održavani tokom različitih vremenskih perioda (30 min, 3 h, 6 h, 12 h, 24 h i 48 h) nakon povrede. Prilikom uzorkovanja za svako ponavljanje sakupljano po ~1,5 g tkiva koje je zatim zamrznuto u tečnom azotu i čuvano na -80 °C do kvantifikacije genske ekspresije. U eksperimentu je korišćeno po 30 eksplantata za svaku vremensku tačku, koji su bili podeljeni u po tri petri kutije sa po deset eksplantata. Jedna petri kutija je predstavljala jedno biološko ponavljanje. Kao kontrola (0 min) su korišćeni isečci koji su odmah nakon uzorkovanja zamrzavani u tečnom azotu. Uzimanje isečaka listova i korenova predstavlja mehaničku povredu koja dovodi do izazivanja stresa, a vreme koje eksplantati provode na hranljivoj podlozi između

mehaničke povrede i zamrzavanja uzoraka predstavlja vreme oporavka od stresa (engl. „*recovery time*“).

3.4.3. Ispitivanje efekta βGlcY na ekspresiju *AGP* gena kičice gajene u uslovima *in vitro*

Za ispitivanje efekata βGlcY na ekspresiju *AGP* gena kičice segmenti listova i korenova su gajeni na ½MS podlozi (Tabela 7), sa dodatkom βGlcY u različitim koncentracijama (0, 5, 10, 15, 25, 50, 75, 100 ili 150 μM). Korišćen je βGlcY dobijen sintezom iz prekursora p-aminofenil-β-D-glukoze i fluoroglucinola uz prisustvo $NaNO_2$ prema protokolu Yariv i sar. (1962). Kao eksplantati su korišćeni isečci listova i korenova dobijeni na isti način kao u eksperimentu sa ispitivanjem efekta mehaničkih povreda na ekspresiju *AGP* gena kičice (odeljak 3.4.2.). U hranljive podloge za ispitivanje uticaja βGlcY na ekspresiju *AGP* gena u listovima kičice dodati su 0,2 mgL^{-1} N-(2-hloro-4-piridil)-N'-fenilurea (CPPU) i 2,4-dihlorfenoksisirćetna kiselina (2,4-D) pošto je pokazana uspešna regeneracija izdanaka kičice na hranjivim podlogama sa pomenutim regulatorima rastenja, putem indukovane SE (Filipović i sar., 2014), za razliku od korenova kod kojih se direktna SE odigrava spontano. U eksperimentu je korišćeno po 30 eksplantata za svaku koncentraciju βGlcY reagensa. Deset eksplantata je činilo jedno biološko ponavljanje. Eksplantati su gajeni na podlogama sa βGlcY tokom četiri nedelje nakon čega su zamrznuti u tečnom azotu i čuvani na -80 °C do analize ekspresije.

3.4.4. Ispitivanje ekspresije *AGP* gena kičice u različitim tkivima biljaka gajenih *in vitro* i delovima biljaka iz prirode

Za ispitivanje ekspresije *AGP* gena u različitim tkivima kičice, korišćeni su uzorci tkiva biljaka gajenih u uslovima *in vitro* bez dodatih regulatora rastenja, uzorci tkiva iz različitih faza organogeneze i somatske embriogeneze, kao i uzorci tkiva biljaka iz prirode (Tabela 8, Ćuković i sar., 2020). Uzorci su pripremljeni tokom izrade doktorata koleginice Katarine Ćuković sa Odeljenja za fiziologiju biljaka, Instituta za biološka istraživanja „Siniša Stanković“ koja je velikodušno podelila cDNA radi ispitanja ekspresije *AGP* gena. Među šesnaest uzoraka razlikuju se četiri grupe:

- uzorci koji obuhvataju organe *in vitro* gajenih biljaka: celi klijanci, listovi i korenovi biljaka u fazi rozete i korenovi iz kulture korenova. Uzorci su uzimani sa biljaka koje su gajene na čvrstoj ½MS hranljivoj podlozi bez dodatih regulatora rastenja u uslovima dugog dana. Tri nedelje stari klijanci su ili uzorkovani ili su prebacivani na svežu ½MS podlogu. Nakon tri meseca gajenja u istim uslovima kao i u toku isklijavanja uzorkovani su uzorci listova i korenova rozetaste forme biljke, koji su poslužili kao osnova za ostatak eksperimenta. Korenovi rozetaste forme biljaka su korišćeni za uspostavljanje kulture korenova koja je gajena na ½MS hranljivoj podlozi.
- uzorci tkiva iz različitih faza organogeneze: organogeni kalus i adventivni pupoljci formirani na eksplantatima listova koji su gajeni u uslovima dugog dana, na hranljivoj podlozi sa regulatorima rastenja (0,2 mgL^{-1} 2,4-D i 0,5 mgL^{-1} CPPU), kao i adventivni

pupoljci formirani na eksplantatima listova ili korenova koji su gajeni u uslovima dugog dana, bez regulatora rastenja. Uzorci su indukovani na hranljivim podlogama i svetlosnom režimu koji su naznačeni u Tabeli 8, a sakupljeni su tri nedelje nakon indukcije.

- uzorci tkiva iz različitih faza somatske embriogeneze: embriogeni kalus, globularni i kotiledonarni embrion. Uzorci su indukovani na hranljivim podlogama uz dodatak 0,2 mgL^{-1} 2,4-D i 0,5 mgL^{-1} CPPU i gajeni u uslovima kontinualnog mraka (Tabela 8). Sakupljeni su tri nedelje nakon indukcije.
- Uzorci organa sa biljaka u cvetu sakupljenih u prirodi: list, koren, stablo, zreli i nezreli cvetovi (Tabela 8).

Tabela 8. Uzorci tkiva *C. erythraea* korišćeni za analizu ekspresije *AGP* gena.

poreklo biljnog materijala	uzorak	hranljiva podloga	fotoperiod	oznaka uzorka
biljke gajene u uslovima *in vitro*	listovi rozete	čvrsta hranljiva podloga bez regulatora rastenja	16 h svetlost/8 h mrak	rl
	korenovi rozete			rr
	klijanci			sd
	korenovi iz kulture korenova			rc
biljke iz prirode	list	bez tretmana	prirodno osvetljenje	ln
	koren			rn
	stablo			st
	zreli (otvoreni) cvetovi			mf
	nezreli (zatvoreni) cvetovi			imf
organogeneza	organogeni kalus	0,2 mgL^{-1} 2,4-D 0,5 mgL^{-1} CPPU	16 h svetlost/8 h mrak	oc
	adventivni pupoljci formirani na eksplantatima listova			ablh
	adventivni pupoljci formirani na eksplantatima listova	čvrsta hranljiva podloga bez regulatora rastenja		abl
	adventivni pupoljci formirani na eksplantatima korena			abr
somatska embriogeneza	embriogeni kalus	0,2 mgL^{-1} 2,4-D 0,5 mgL^{-1} CPPU	kontinualni mrak	ec
	globularni embrion			gse
	kotiledonarni embrion			cse

3.5. Molekularno biološke metode za ispitivanje ekspresije *AGP*

3.5.1. Izolacija RNK iz listova i korenova kičice

Izolacija ukupne RNK iz biljnih uzoraka urađena po modifikovanom protokolu koji su koristili Gašić i sar. (2004). Pre početka izolacije, svi korišćeni sudovi su tretirani 0,1% DEPC vodom (vodom u koju je dodat inhibitor ribonukleaza dietilpirokarbonat) i sterilisani autoklaviranjem. Oko 1,5 g biljnog tkiva po biološkom ponavljanju koje je prethodno uzorkovano i zamrznuto na -80 °C je samleveno u avanima u tečnom azotu. Zatim je dodato po 650 µl ekstrakcionog pufera (100 mM Tris-HCl pH 8,0; 25 mM Na-EDTA; 2 M NaCl; 0,5 gL^{-1} spermidin;

2% (*w/v*) polivinilpirolidon; 2% (*w/v*) etiltrimetilamonijum-bromid и 15 µl β-merkaptoetanola) u svaki uzorak i homogenat je prebačen u sterilne tubice od 2 ml na ledu. Uzorci su mućkani na vorteksu 2 min i zatim inkubirani 15 min u vodenom kupatilu na 60 °C. U svaki uzorak dodato je po 650 µl smeše hloroform:izoamilalkohol (24:1) nakon čega su uzorci mućkani na vorteksu 2 min i centrifugirani 10 min na 10000 g na 4 °C. Supernatant je prebačen u nove tubice i postupak ekstrakcije smešom hloroform:izoamilalkohol ponovljen je još jedan put. U supernatant je potom dodato 166 µl 7,5 M LiCl i uzorci su ostavljeni da se inkubiraju tokom noći na +4 °C. Uzorci su potom centrifugirani 45 minuta na 12000 g na 4 °C, nakon čega je supernatant uklonjen. Svaki uzorak je ispiran sa po 1 ml 70% DEPC etanola i centrifugiran na 12000 g na +4 °C 10 minuta. Supernatant je odstranjen, nakon čega su tubice sa uzorcima ostavljene u komori sa laminarnim protokom vazduha 10 minuta radi sušenja. Uzorci su potom rastvoreni u 100 µl sterilne vode i dodato im je 10 µl 3 M natrijum acetata (pH 5,5) i 275 µl 70% etanola. Uzorci su promešani okretanjem tubica, zatim inkubirani 2 h na -80 °C i potom centrifugirani na 12000 g na +4 °C 45 minuta. Nakon odlivanja supernatanta, precipitat je ispran sa 1 ml 70% DEPC etanola i centrifugiran na 12000 g na +4 °C 10 minuta. Supernatant je odstranjen, a precipitat je osušen u komori sa laminarnim protokom vazduha i rastvoren u 50 µl DEPC vode.

Koncentracija izolovane RNK je određena spektrofotometrijski (*N60 Nano-Photometer®, Implen* GmbH, Nemačka), merenjem apsorbanci na 260 i 280 nm, pošto odnos A_{260}/A_{280} ukazuje na potencijalnu kontaminaciju. Ukoliko je taj odnos 1,9-2,1 smatra se da je RNK zadovoljavajućeg kvaliteta. Kvalitet izolovane RNK proveren je na agaroznom gelu.

3.5.2. Elektroforeza na agaroznom gelu

Kvalitet izolovane RNK i razdvajanje proizvoda reakcije lančanog umnožavanja (PCR, engl. *„polymerase chain reaction“*) je urađeno horizontalnom elektroforezom na 1,2 - 2% agaroznom gelu. Za vizuelizaciju je korišćen etidijum bromid (0,5 µg ml^{-1}) u TBE puferu (89 mM Tris, 89 mM borna kiselina, 2 mM EDTA). U svaki uzorak je dodata boja (*6x DNA Loading Dye, Thermo Scientific*, SAD) pre nanošenja na gel. Na gel je nanošeno 12 µl smeše uzorka i boje. Za elektroforezu su korišćene kadice *BlueMarine™ 200* ili *BlueMarine™ 100* (*SERVA Electrophoresis GmbH*, Heidelberg, Nemačka) sa izvorom napajanja *Standard Power Pack P25, Biometra®,* (*Goettingen*, Nemačka) koje su pre upotrebe, zajedno sa češljićima, tretirane rastvorom vodonik peroksida (3%) u trajanju od 60 minuta za potrebe provere kvaliteta izolovane RNK. RNK na gelu je detektovana pomoću UV transiluminatora (*ST4 3026-WL/26M, Vilber Lourmat, Torcy,* Francuska). Za utvrđivanje veličine fragmenata korišćeni su DNK markeri *O'GeneRuler 100 bp Plus DNA Ladder* (*#SM1153, Fermentas, Waltham,* SAD) ili *GeneRuler 50 bp DNA Ladder* (*#SM0371, Fermentas, Waltham,* SAD).

3.5.3. Tretman dezoksiribonukleazom (DNK-azom)

Izolovana RNK je tretirana dezoksiribonukleazom I (DNase I, *Thermo Scientific, Waltham,* SAD) da bi se eliminisali tragovi genomske DNK. Sastav reakcione smeše je dat u Tabeli 9.

Tabela 9. Sastav reakcione smeše za tretman DNK-azom

komponenta	zapremina (µl)
RNK u vodi (1µg)	7,75
10x reakcioni pufer za DNK-azu I	1
DNK-aza I (1U/ µl)	1
inhibitor RNK-aza (40 U/µl)	0,25

Tretman je trajao 30 minuta na 37 °C u aparatu *Mastercycler® nexus Gradient* (*Eppendorf*, Hamburg, Nemačka), nakon čega je reakcija zaustavljena dodavanjem 1 µl 25 mM EDTA u svaki uzorak i inkubacijom 10 min na 65 °C. Dobijena RNK korišćena je dalje u reakciji reverzne transkripcije.

3.5.4. Reverzna transkripcija

Reakcija reverzne transkripcije (RT), odnosno prevođenje RNK u komplementarne jednolančane molekule DNK (cDNK) izvedena je korišćenjem *RevertAid First Strand cDNA Synthesis Kit* kompleta (*Thermo Scientific*, *Waltham*, SAD) po uputstvu proizvođača. Sastav reakcione smeše prikazan je u Tabeli 10.

Tabela 10. Sastav reakcione smeše za reakciju RT

komponenta	zapremina (µl)
RNK (1µg)	11
oligo (*dT*) prajmer	1
5x reakcioni pufer	4
inhibitor RNK-aza (20 U μl^{-1})	1
10 mM *dNTP* mix	2
RT (*RevertAid™, M-MuLV* (*200 U* μl^{-1}))	1

Uzorci su inkubirani 60 min na 42 °C, a zatim je reakcija zaustavljena inkubacijom na 70 °C u trajanju od 5 minuta. Dobijena cDNK čuvana je na - 80 ºC do kvantifikacije ekspresije.

3.5.5. Dizajniranje prajmera za odabrane *AGP* gene

Za odabrane sekvence *AGP* gena dizajnirani su specifični prajmeri korišćenjem *Primer-BLAST* softverskog alata (www.ncbi.nlm.nih.gov/tools/primer-blast) (Tabele 11, 12 i 13). Specifičnost prajmera tokom konstrukcije je proveravana *in silico* korišćenjem *blastn* paketa i transkriptoma kičice. Eksperimentalna provera specifičnosti prajmera urađena je gel elektroforezom RT-PCR proizvoda.

Tabela 11. Sekvence prajmera za amplifikaciju izabranih sekvenci FLA.

naziv gena	sekvenca prajmera 5'→ 3'	temperatura vezivanja	dužina amplikona
CeFLA1	F: CGGAGTTGATACGACCACGG	56,9	105
	R: GCAGGACCAAACATACTAAGAGG		
CeFLA3	F: CTTGTCGCTTCTCGTTCTGC	57,5	161
	R: ACCCTACCTTTCCGCCTTTC		
CeFLA4	F: CTCCACCACCTGTGATTGCC	56,2	188
	R: TCAACACTCTTTCACTCACATCC		
CeFLA5	F: TAGCCCGAAAACTACTCCTGC	58,1	148
	R: TAGCACTCCTCCACACACTG		
CeFLA6	F: GTGCTTTTGCCTGTGGAGTTT	56,9	159
	R: CAGCCCCACAATCACTCTCC		
CeFLA7	F: TACCACTTACACTCCCAGCAC	55,1	141
	R: TCGGCAAGGAAGAGTGAGGT		

Tabela 12. Sekvence prajmera za amplifikaciju izabranih sekvenci KLA.

naziv gena	sekvenca prajmera 5'→ 3'	temperatura vezivanja	dužina amplikona
CeKLA1	F: GTTCTTGGTTCGGCGTTGAG	59,1	198
	R: CAAAGTCAAGCACCTCCAGC		
CeKLA2	F: CGATCATGGCAGCATCTCCT	56,7	213
	R: GAGCATCTGGAGGAGCACTG		
CeKLA3	F: TAGGGGAACAGAGATCACACA	54	102
	R: ACATCTCCAAATCCGCCTC		
CeKLA5	F: TTGCCAGATGAGTTGCCTGT	57,3	178
	R: AGAAGGAAAGCACTCGCACA		
CeKLA6	F: TGGGGTGGTTATGTTGGAGC	55,4	174
	R: CGGGAGAGAGACTTGACTGG		
CeKLA7	F: AGATGAAAGCCAGCCAGAGG	57,9	187
	R: CGGTCTTTGATGTGGTTCCC		

Tabela 13. Sekvence prajmera za amplifikaciju izabranih sekvenci AGp.

naziv gena	sekvenca prajmera 5'→ 3'	temperatura vezivanja	dužina amplikona
CeAGp3	F: AGGGCGTCGTTCGGAGTAG	59,1	159
	R: AGCCAAAGCCACCAGCATTAG		
CeAGp6	F: TCTCCTACATTCCTCCACCTG	57,3	141
	R: CCACAACGATAAACGAAAGGC		
CeAGp7	F: CTTTCCTCATCTTTTCTGCTCAC	56,1	152
	R: GTGCTTGTGCCTGTGCG		
CeAGp8	F: GCCTTTCTTCCACCTTCTCC	56,7	147
	R: TGCTGGAGTCGGAGTTTTAGC		
CeAGp9	F: GCAGAGGCACAAGCAGTTTC	55,4	132
	R: AGTGAGAACCAAAGCCAAAACC		
CeAGp10	F: ACACTTTTCTTCGCTTTCTTGC	56,1	135
	R: CCCCTCCAGTCTTTCCCG		

3.5.6. PCR amplifikacija

Za proveru kvaliteta izolovane RNK, za optimizaciju temperature vezivanja prajmera, kao i za kontrolu dizajniranih prajmera korišćena je reakcija lančanog umnožavanja (PCR, engl. „*polymerase chain reaction*“). Reakcija je izvedena u *peqSTAR 96 Universal Gradient* (*Peqlab, Biotechnologie GmbH, Erlangen*, Nemačka), korišćenjem AmpliTaq Gold® PCR Kit (*Applied Biosystems, Waltham*, SAD) u zapremini reakcije od 25 µl. Sastav reakcione smeše za PCR amplifikaciju prikazan je u Tabeli 14.

Tabela 14. Sastav reakcione smeše za PCR amplifikaciju

komponente	zapremina (µl)
10x reakcioni pufer sa $MgCl_2$	2,5
dNTP(10mM*)*	1,5
Taq polimeraza (5 U/µl)	0,5
Forward (F) + *Reverse* (R) prajmer (10 µM)	1,0
H_2O	18,5
cDNK	1

Program za PCR amplifikaciju se sastojao od inicijalne denaturacije na 95 °C u trajanju od 10 min, zatim 40 ciklusa denaturacije (95 °C, 30 s), vezivanja prajmera (za temperaturu videti Tabele 11 - 13, 30 s) i elongacije (72 °C, 1 min), kao i finalne elongacije na 72 °C tokom 10 min. Temperatura vezivanja korišćenih prajmera je bila različita, za svaki par prajmera specifična. PCR reakcije su se odvijale u mašini *Mastercycler® nexus Gradient* (*Eppendorf*, Hamburg, Germany). PCR proizvodi su proveravani na agaroznom gelu kao što je opisano u odeljku 3.5.2.

3.5.7. Prečišćavanje PCR proizvoda i priprema standarda za qRT-PCR reakcije

Proizvodi PCR reakcije su nakon elektroforeze na agaroznom gelu prečišćeni korišćenjem *GeneJet extraction* kita (*Thermo Scientific, Waltham,* SAD) po protokolu proizvođača. Nakon detektovanja proizvoda PCR amplifikacije pod UV svetlom, delovi gela sa amplikonima DNK su isecani i stavljani u tubice od 1,5 ml čija je masa prethodno izmerena, da bi se utvrdila masa isečenog gela. U svaki uzorak je dodat vezujući pufer (engl. „*binding buffer*“) u odnosu 1:1 (*v:v*) i uzorci su inkubirani na temperaturi od 60 ºC u trajanju od 10 minuta ili dok se celokupan gel ne rastvori. Svakom uzorku je zatim dodat 100% izopropanol u odnosu 1:1 (*v:v*). Zatim su uzorci prebacivani u kolonu za prečišćavanje i centrifugirani u trajanju od 1 minuta na 12000 g. U svaku kolonu je zatim dodato 700 µl pufera za ispiranje (engl. „*wash buffer*“) i uzorci su opet centrifugirani 1 minut nakon čega je odbačena tečnost i prazna kolona centrifugirana još jednom da bi se uklonio etanol iz prečišćene DNK, koji može da inhibira enzimske reakcije. Kolone su zatim prebačene u čiste tube od 1,5 ml i isprane sa 50 µl elucionog pufera (engl. „*elution buffer*) centrifugiranjem tokom 1 minuta na 12000 g. Koncentracija prečišćene DNK je određena spektrofotometrijski merenjem apsorbance na talasnoj dužini od 260 nm, a DNK je skladištena na -20 ºC i korišćena za pravljenje odgovarajućih razblaženja DNK za standardnu krivu (razblaživanjem se dobijaju rastvori sa 10^2-10^9 kopija).

3.5.8. Kvantitativni RT-PCR

Metoda kvantitativnog RT-PCR (qRT-PCR) korišćena je za određivanje nivoa ekspresije odabranih *AGP* gena upotrebom *Maxima SYBR Green/Rox qPCR Master Mix* (*Thermo Scientific, Waltham,* SAD) u mašini *QuantStudio™ 3 Real-Time PCR* (*Thermo Fisher Scientific*, *Waltham,* SAD) prema uputstvu proizvođača.

Za svaki gen postavljana je kontrola bez cDNK (engl. „*non template control*“-NTC), koja je umesto cDNK sadržala 1 µl H_2O. Komponente reakcije su prikazane u Tabeli 15.

Tabela 15. Sastav reakcione smeše za qPCR reakcije

komponenta	**zapremina (µl)**
H_2O	2,8
SYBR Master Mix	5
F prajmer (10 µM)	0,6
R prajmer (10 µM)	0,6
cDNK(100 ng / µl) ili standard	1

Program za qPCR reakciju se sastojao od inicijalne denaturacije na 95°C u trajanju od 5 min, a zatim 40 ciklusa koji se sastoje od: denaturacije (95 °C, 15 s), vezivanja prajmera na temperaturi specifičnoj za dati prajmer (Tabele 11 - 13, 30 s) i elongacije (72 °C, 30 s). Na kraju sledi finalna ekstenzija na 72 °C tokom 10 min i analiza krive topljenja. Relativan nivo ekspresije gena određen je prema $2^{-\Delta\Delta Ct}$ metodi (Livak i Schmittgen, 2001) uz upotrebu odgovarajućih referentnih gena.

3.5.9. Izbor referentnih gena za ispitivanje ekspresije *AGP* gena

Radi odabira optimalnih referentnih gena za eksperiment sa mehaničkim povredama upoređena je stabilnost ekspresije jedanaest gena koji su odabrani na osnovu funkcije i literaturnih podataka: gen za histone (*H3*), gvožđe-superoksid dismutaze (*SOD1* i *SOD2*), RAS-povezani nuklearni protein (*RAN*), β-tubulin (*TUB*), adenozin kinazu (*AK*), elongacioni faktor 2 (*EF2*), α-tubulin (*TUA*), TATA-vezujući protein 1 (*TBP1*), Ribozomalni protein L2 (*RPL2*) i 18s ribozomalnu RNK (*18s*). Sekvence prajmera za odabrane referentne gene su date u Tabeli 16.

Tabela 16. Sekvence prajmera za amplifikaciju odabranih referentnih gena.

naziv gena	sekvenca prajmera 5'→ 3'	temperatura vezivanja	dužina amplikona
TBP1	F: CATCACACGAGCCCACAATG	56,9	176
	R: AGAGCCAAAAACGACAGCAC		
AK	F: CTTTAGCCTCACCAGCCTCA	56,9	167
	R: CCTCGGCGTTCCTCACATT		
RAN	F: ACCTTGTTTCCACATAGCAC	53	167
	R: CAGGAGAAGTTTGGCGG		
RPL2	F: AGGGTCGTGAATGATGTCGG	53	137
	R: CACAGCGTAAGGGAGCAGG		
H3	F: AGTAATCGGACTCTTCGTCGC	56,9	122
	R: TCTCTGCCGTTCACCTCAAT		
SOD1	F: CTTCTTGATGCCAATGCCTCG	55,1	198
	R: TACTTCGGTTGCTGGGTTCG		
SOD2	F: GTCTTGTGATGCTTTCCCCA	58	171
	R: GACTCGGTGCTTTTCCAGAC		
TUB	F: GACTCGGTGCTTTTCCAGAC	59	156
	R: GGCACACCTGAAATCCTTGG		
TUA	F: GCGAGCGAAGTTGTTAGCG	58,8	192
	R: GTCGGAGGCGGAGATGATG		
EF2	F: CAGACCAACCATAGCAACTG	53	119
	R: CGGGAGAGAAGAAGGACC		
18S	F: ATGCCCTTAGATGTTCTGGGC	55,1	193
	R: AAGGGCAGGGACGTAGTCAA		

Stabilnost genske ekspresije ovih jedanaest gena evaluirana ja sa *geNorm* (Vandesompele i sar., 2002) i *NormFinder* (Andersen i sar., 2004) algoritmima uz upotrebu R paketa *NormqPCR* (Perkins i sar., 2012). Algoritam *GeNorm* procenjuje stabilnost ekspresije gena (M) za svaki gen kao prosečnu varijaciju odnosa ekspresija između određenog gena i svih ostali testiranih kandidata (Vandesompele i sar., 2002). Najmanje stabilan gen se eliminiše i računa se nova vrednost M za preostale gene. Niže vrednosti M ukazuju na veću stabilnost genske ekspresije. Algoritam *NormFinder* upoređuje varijacije genskih ekspresija unutar jednog skupa uzoraka i između različitih skupova uzoraka i kombinuje oba rezultata u vrednost stabilnosti za svaki gen (Andersen i sar., 2004).

Za određivanje adekvatnih referentnih gena u eksperimentu sa βGlcY reagensom, evaluirana su četiri gena (*TBP1, AK, RAN* i *RPL2,* Tabela 16) koji su imali najstabilniju ekspresiju u različitim tkivima kičice u radu Ćuković i sar. (2020). Urađen je qRT-PCR ovih gena sa uzorcima tkiva gajenih na različitim koncentracijama βGlcY i rezultati su zatim analizirani algoritmima *GeNorm* i *NormFinder* kako je već opisano.

Kao referentni geni za eksperimente sa različitim tkivima biljaka kičice gajenih *in vitro* i delovima biljaka iz prirode korišćeni su *TBP1* i *RPL2* geni koji su se pokazali kao najstabilniji (Ćuković i sar., 2020).

3.6. Statistička obrada podataka o genskoj ekspresiji

Statistička analiza podataka urađena je korišćenjem R programskog jezika (R Core Team, 2020). Relativna ekspresija gena (na logaritamskoj skali) iz eksperimenata sa povredama biljnog tkiva, kao i sa različitim koncentracijama Yariv reagensa analizirana je korišćenjem Studentovog t-testa (Student, 1908), gde je svaka vremenska tačka/koncentracija βGlcY upoređivana sa kontrolom (0 min ili 0 μM βGlcY). Relativna ekspresija gena iz eksperimenata sa različitim tkivima kičice analizirana je korišćenjem *Welch*-ovog t-testa (Welch, 1947) u kome je svaki uzorak upoređivan sa kontrolom (rl – list rozete). P-vrednosti dobijene pomenutim poređenjima za oba eksperimenta podešene su za višestruka poređenja korišćenjem BH metode (Benjamini i Hochberg, 1995). Linearni odnosi između parova genskih ekspresija procenjeni su korišćenjem *Pearson*-ove korelacije (cor_{Pear}) (Pearson, 1895), a korelaciona matrica prezentovana je kao toplotna mapa korišćenjem *gplots* R paketa (Warnes i sar., 2020). Redovi i kolone toplotne mape poređani su prema hijerarhijskoj analizi klastera urađenoj na osnovu matrice korelacionih distanci (1 - cor_{Pear}, maksimalna pozitivna korelacija od 1 postaje distanca 0, a maksimalna negativna korelacija od -1 postaje distanca 2). Aglomeracija klastera izvedena je korišćenjem metode potpunog povezivanja (engl. „*complete linkage*“). Dodatno su međusobne zavisnosti ekspresije gena procenjene korišćenjem korigovanog za subjektivnost uzajamnog sadržaja informacije (engl. „*jackknife bias corrected mutual information pairwise matrix*“), koji je određen za sve kombinacije parova gena korišćenjem R paketa *mpmi* (Pardy, 2020).

4. Rezultati

4.1. Identifikacija i analiza HRGP sekvenci

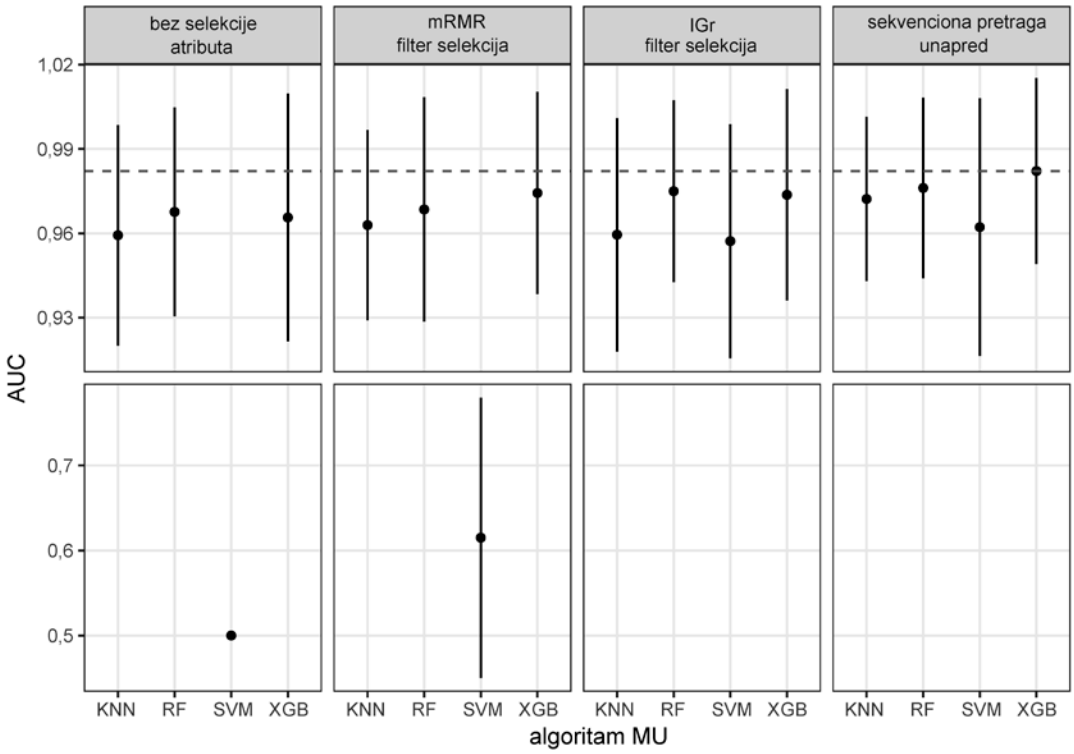
4.1.1. Procena performansi modela za predviđanje verovatnoće hidroksilacije prolina na osnovu lokalne sekvence

Centralni deo metodologije za identifikaciju i analizu HRGP sekvenci implementirane u okviru *ragp* R paketa razvijenog u okviru ove disertacije je predviđanje verovatnoće hidroksilacije prolina na osnovu lokalne sekvence. Da bi se odabrao pristup MU koji bi imao zadovoljavajuće performanse, upoređena su četiri algoritma MU: KNN, SVM, RF i XGB. Algoritmi su trenirani na proteinskim sekvencama biljaka sa eksperimentalno određenim Hyp, odnosno na lokalnim 21-mernim sekvencama u čijem centru se nalaze ciljni Pro/Hyp. Pošto je skup atributa napravljen za opisivanje ovih lokalnih 21-mernih sekvenci sadržao 1294 jedinstvena atributa podeljena u 16 grupa (Tabela 4), pokušano je nekoliko pristupa za njihov odabir koji su za cilj imali smanjenje dimenzionalnosti problema: dva tipa filter algoritama IGr i mRMR, kao i sekvenciona potraga unapred. Dakle ukupno je upoređeno šesnaest različitih kombinacija: (tri načina odabira atributa uz dodatan pristup bez odabira atributa) × četiri algoritma MU.

Da bi se smanjila pristrasnost evaluacije modela, korišćena je unakrsna validacija u dva sloja gde je spoljašnji sloj korišćen za procenu performansi modela, a unutrašnji za odabir hiperparametara Bajesovskom optimizacijom. Performanse modela su merene korišćenjem vrednosti površine ispod *ROC* krive (AUC).

Performanse RF i XGB algoritama su bile primetno više u poređenju sa KNN i SVM nezavisno od korišćenog pristupa za izbor atributa. SVM je pokazao najslabije performanse u svim slučajevima, a posebno kada je korišćen mRMR algoritam za odabir atributa, odnosno u slučaju kada su svi atributi korišćeni za treniranje. U slučaju RF i XGB algoritama su izmerene relativno male razlike u performansama između različitih metoda za odabir atributa (Slika 5).

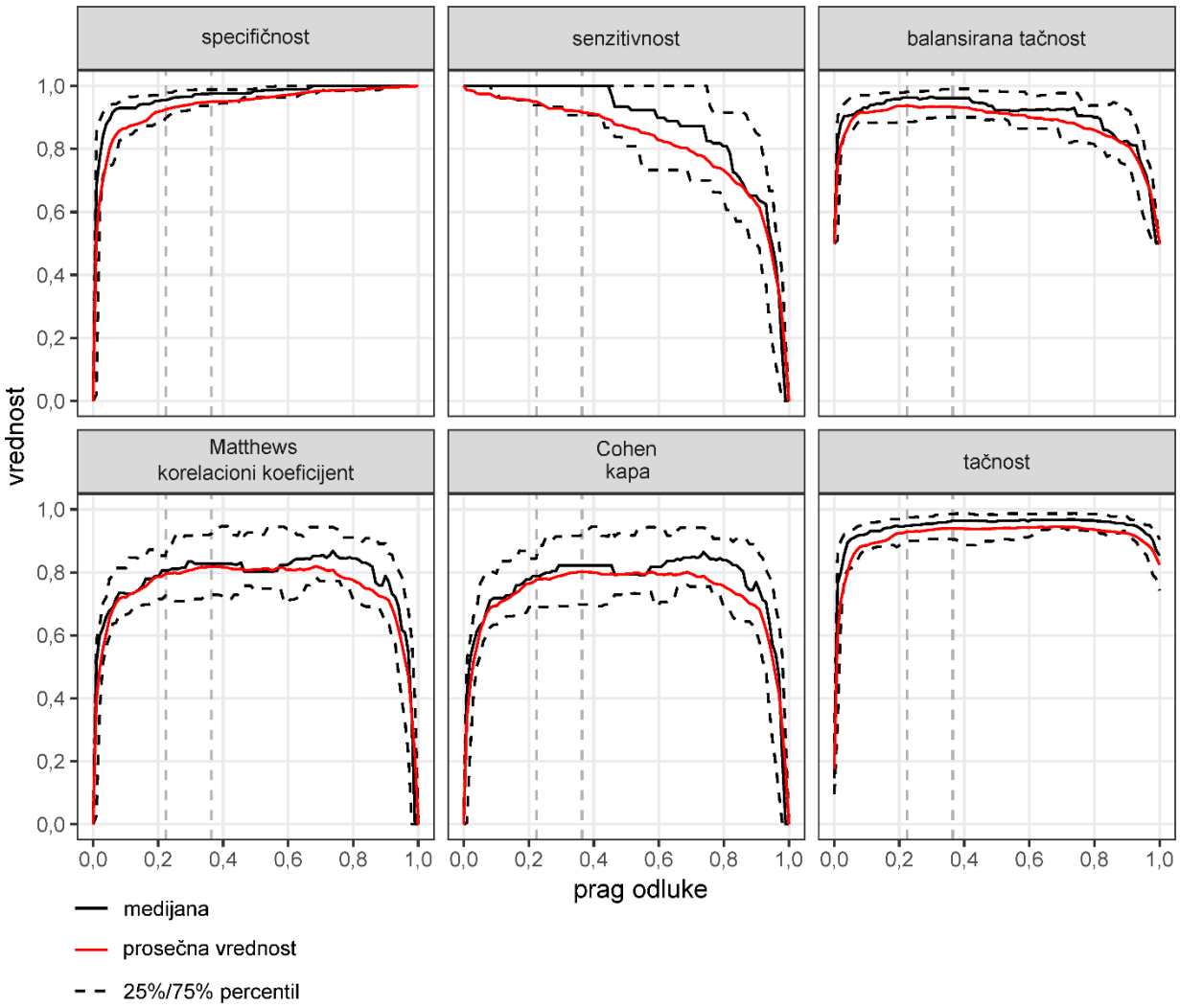
Za sva četiri algoritma MU, najviše performanse su dobijene korišćenjem sfs odabira atributa, pri čemu je XGB imao neznatno bolje performanse (srednja vrednost ± SD: 0,982 ± 0,033 AUC) u poređenju sa RF (0,976 ± 0,039 AUC), zatim slede KNN (0,972 ± 0,029) i SVM (0,965 ± 0,043 AUC). Prema tome, za treniranje modela koji je inkorporiran u *ragp* R paket izabran je XGB algoritam zajedno sa sfs metodom za odabir atributa kroz 16 grupa atributa. Grupe atributa koji su sfs metodom odabrane su F1, F3, F4 i F10 (Tabela 4) što je rezultiralo u 316 jedinstvenih atributa na kojima je model treniran.



Slika 5. Poređenje performansi predviđanja Hyp u unakrsnoj validaciji u dva sloja. Za procenu performansi je korišćena AUC metrika. Prikazana je srednja vrednost ± SD performansi u spoljašnjem sloju unakrsne validacije koji se sastojao od unakrsne validacije sa 10 podela. Unutrašnji sloj unakrsne validacije koji je korišćen za odabir hiperparametara kroz 100 iteracija Bajesovske optimizacije sastojao se od dva puta ponovljene unakrsne validacije sa tri podele. KNN, RF, SVM i XGB - različiti algorimi MU; mRMR, IGr i sfs - različite metode odabira atributa.

4.1.2. Optimizacija praga odluke

Optimizacija praga odluke urađena je dodatnom unakrsnom validacijom u dva sloja, korišćenjem F1, F3, F4 i F10 skupova atributa za treniranje, koji su se prethodno pokazali kao optimalni sfs algoritmom, uz XGB algoritam MU. Spoljašnji sloj unakrsne validacije se u ovom slučaju sastojao od tri puta ponovljene unakrsne validacije sa deset podela, dok je unutrašnji sloj, koji je služio za odabir hiperparametara, bio istovetan kao što je opisano u odeljku 4.1.1. Na osnovu predviđanja modela u spoljašnjem sloju unakrsne validacije ispitana je zavisnost praga odluke i nekoliko metrika za evaluaciju performansi modela: osetljivost, specifičnost, tačnost, balansirana tačnost, *Matthews*-ov korelacioni koeficijent i *Cohen*-ov kapa koeficijent (Slika 6).



Slika 6. Efekat praga odluke na performanse modela treniranog na 21-mernim lokalnim sekvencama u unakrsnoj validaciji u dva sloja. Modeli su trenirani sa kombinacijom F1, F3, F4 i F10 skupova atributa (Tabela 4). Za svaku metriku su prikazane srednje vrednosti, medijane, 25% i 75% percentil na osnovu predviđanja u spoljašnjem sloju unakrsne validacije koji se sastojao od tri puta ponovljene unakrsne validacije sa deset podela. Unutrašnji sloj unakrsne validacije koji je korišćen za odabir hiperparametara kroz 100 iteracija Bajesovske optimizacije sastojao se od dva puta ponovljene unakrsne validacije sa tri podele. Horizontalne isprekidane linije odgovaraju pragu odluke od 0,224 koji maksimizuje srednju vrednost balansirane tačnosti i pragu odluke od 0,364 koji odgovara specifičnosti od 0,95 (na osnovu srednje vrednosti).

Sve ispitivane metrike su bile stabilne na širokom opsegu verovatnoća (Slika 6) što ukazuje na dobro razdvajanje klasa. U većini slučajeva je negativnim instancama (Pro) dodeljena (predviđena) niska verovatnoća hidroksilacije, dok je pozitivnim instancama (Hyp) dodeljena visoka verovatnoća hidroksilacije. Odlučeno je da podrazumevani prag odluke za predviđanje hidroksilacije prolina bude baziran na balansiranoj tačnosti, odnosno da ima vrednost 0,224 koja maksimizuje balansiranu tačnost (Slika 6). Na osnovu izvedene unakrsne validacije, srednja osetljivost na ovom pragu je 0,951, dok je srednja specifičnost iznosila 0,925. Treba napomenuti da u *ragp* paketu, korisnici mogu da podešavaju prag odluke prilikom definisanja predviđanja.

4.1.3. Treniranje i evaluacija finalnog modela za predviđanje na 21-mernim lokalnim sekvencama

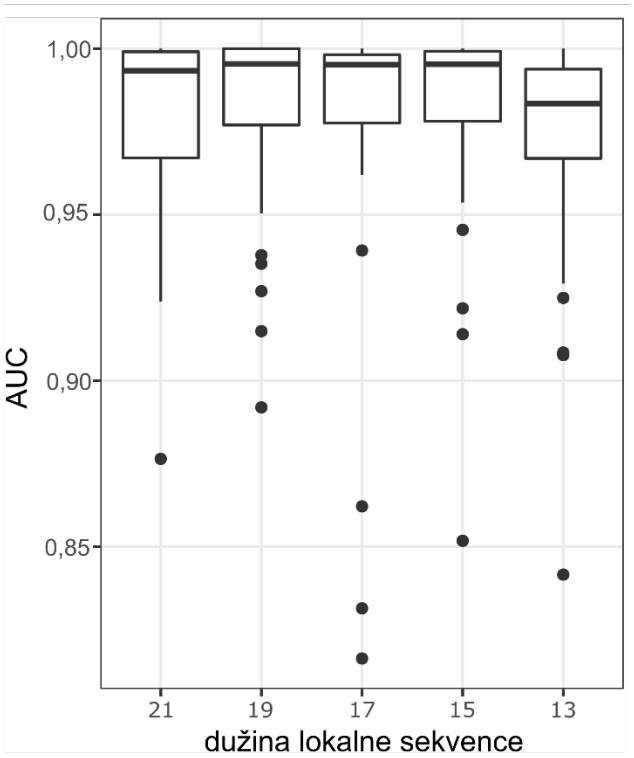
Hiperparametri finalnog modela su određeni kroz sto iteracija Bajesovske optimizacije korišćenjem dva puta ponovljene unakrsne validacije sa tri podele na celom skupu podataka za treniranje. Predloženi hiperparametri: *nrounds* = 758; *min_child_weight* = 1,028; *max_depth* = 15; *eta* = 0,005; *gamma* = 0,527; *colsample_bytree* = 0,801; *subsample* = 0,825; *alpha* = 0,707; *lambda* = 1,61 i *colsample_bylevel* = 0,986 su korišćeni za treniranje modela na celom skupu za treniranje. Performanse dobijenog modela su potom procenjene korišćenjem skupa za evaluaciju - nezavisnih podataka (sekvence proteina) koji nisu ni na koji način učestvovali u procesu pravljenja modela. Korišćenjem podrazumevanog praga odluke (0,224) dobijeni model je imao osetljivost 0,938 (30 TP od 32 pozitivnih instanci) i specifičnost 0,971 (169 TN od 174 negativnih instanci) uz AUC od 0,986 kada je primenjen na skupu podataka za evaluaciju.

4.1.4. Uticaj dužine lokalne sekvence na performanse predviđanja hidroksilacije prolina

Ograničenje modela treniranog na 21-mernim lokalnim sekvencama je nemogućnost predviđanja hidroksilacije prolina koji se nalazi među deset N- ili C-terminalnih aminokiselina. Hidroksilacija N-terminalnih prolina nije od interesa pošto HRGP sadrže N-sp koji je u većini slučajeva duži od deset aminokiselina. Samim tim, ne postoje eksperimentalno dokazani Hyp u okviru deset N-terminalnih aminokiselina kod sekretovanih biljnih proteina. Sa druge strane nemogućnost predviđanja Hyp u okviru deset C-terminalnih aminokiselina dovodi do gubitka potencijalno značajne informacije. Kako bi se ovaj nedostatak delimično otklonio, inicijalno je ispitano predviđanje Hyp samo na osnovu lokalne sekvence koja mu prethodi. Međutim, ovaj pristup je rezultovao niskim (nezadovoljavajućim) performansama (nije prikazano). Sa druge strane predviđanje Hyp na osnovu lokalne sekvence simetrične dužine je pokazalo solidne performanse i kada je lokalna sekvenca bila kraća od 21 aminokiseline.

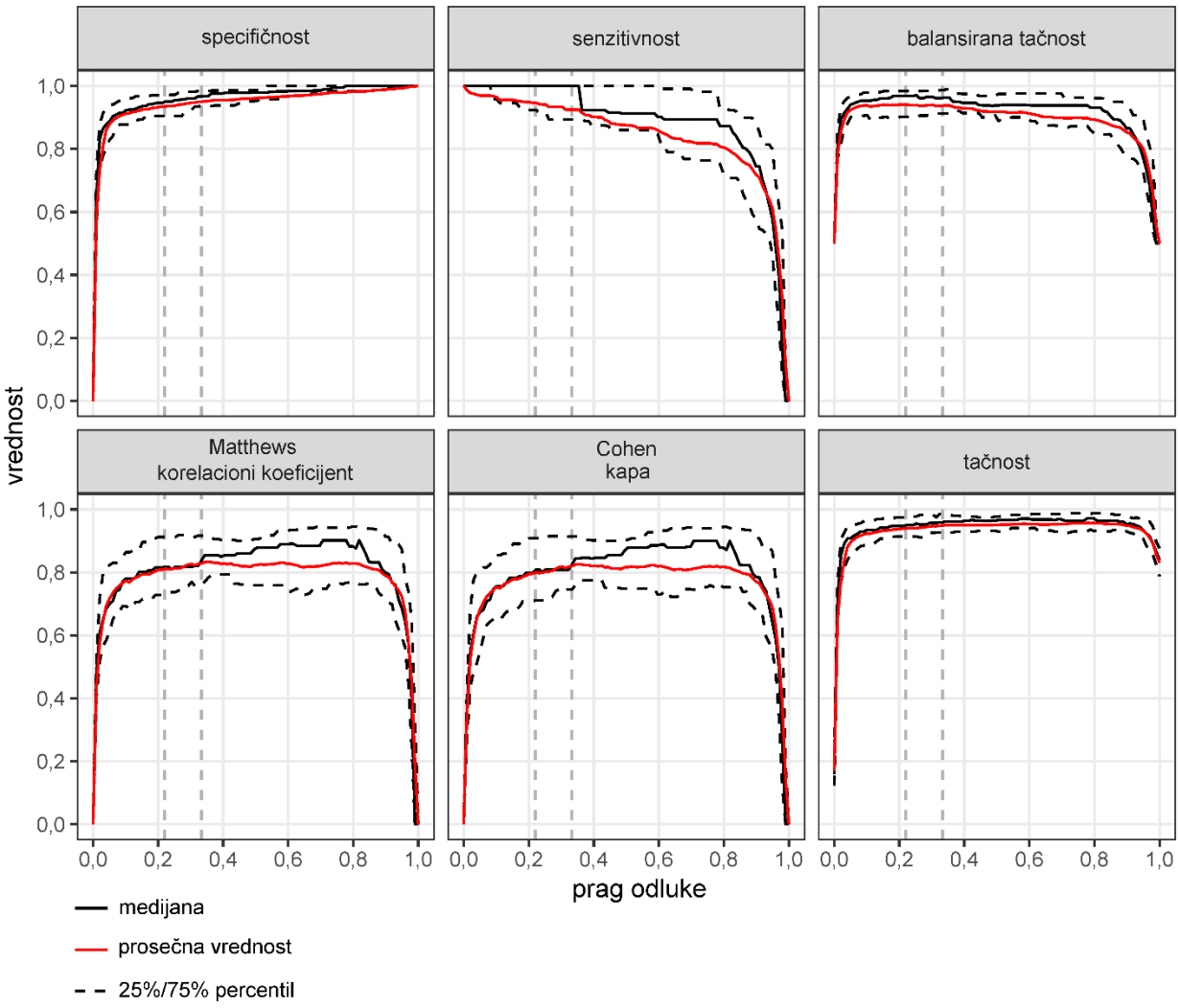
Uticaj dužine lokalne sekvence na performanse modela ispitana je korišćenjem skupova za treniranje i evaluaciju kreiranih na osnovu lokalnih sekvenci dužina 19, 17, 15 i 13 aminokiselina u čijem centru su se nalazili ciljni Pro/Hyp. Modeli su trenirani i evaluirani kao i u slučaju 21-mernih sekvenci korišćenjem unakrsne validacije u dva sloja. Spoljašnji sloj je korišćen za procenu performansi modela, a unutrašnji za odabir hiperparametara Bajesovskom optimizacijom. Modeli su trenirani sa kombinacijom F1, F3, F4 i F10 skupova atributa (Tabela 4) koji su se pokazali optimalni u slučaju 21-mernih lokalnih sekvenci bez daljeg odabira atributa. Za procenu performansi je korišćena AUC metrika. Performanse modela treniranih na skupovima atributa napravljenim od 19, 17 i 15-mernih lokalnih sekvenci su bile slične performansama modela treniranog na skupovima atributa konstruisanih na osnovu 21-mernih sekvenci. Primetan pad performansi je zabeležen kod modela treniranog na 13-mernim sekvencama (Slika 7). Zbog toga je odlučeno da se kao dodatni model za predviđanje verovatnoće hidroksilacije u okviru C-terminalnih prolina koristi model treniran na osnovu 15-mernih lokalnih sekvenci. Ovaj model nije

u mogućnosti da predviđa hidroksilaciju prolina koji se nalaze u okviru sedam C- ili N- terminalnih aminokiselina.



Slika 7. Procena performansi modela za predviđanje Hyp treniranog korišćenjem lokalnih sekvenci različite dužine. Modeli su trenirani sa kombinacijom F1, F3, F4 i F10 skupova atributa (Tabela 4) konstruisanih na osnovu lokalne sekvence odgovarajuće dužine. Prikazana je raspodela AUC vrednosti dobijenih na osnovu predviđanja u spoljašnjem sloju unakrsne validacije koji se sastojao od tri puta ponovljene unakrsne validacije sa deset podela. Unutrašnji sloj unakrsne validacije koji je korišćen za odabir hiperparametara kroz sto iteracije Bajesovske optimizacije sastojao se od dva puta ponovljene unakrsne validacije sa tri podele.

Kao kod modela treniranog na 21-mernim sekvencama, ispitana je zavisnost praga odluke i nekoliko metrika za evaluaciju performansi modela korišćenjem unakrsne validacije u dva sloja (Slika 8). Model treniran na 15-mernim sekvencama pokazao je sličnu zavisnost praga odluke i različitih metrika za evaluaciju performansi predviđanja kao i model treniran na 21-mernim sekvencama. Za predviđanje je odabran prag odluke od 0,22 koji maksimizuje srednju balansiranu tačnost. Korišćenjem ovog praga odluke srednja osetljivost u spoljašnjem sloju unakrsne validacije je iznosila 0,946, dok je specifičnost iznosila 0,936.



Slika 8. Efekat praga odluke na performanse modela treniranog na kombinovanim F1, F3, F4 i F10 skupovima atributa konstruisanim na osnovu 15-mernih lokalnih sekvenci. Za svaku metriku su prikazane srednje vrednosti, medijane, 25% i 75% percentil na osnovu predviđanja u spoljašnjem sloju unakrsne validacije koji se sastojao od tri puta ponovljene unakrsne validacije sa deset podela. Unutrašnji sloj unakrsne validacije koji je korišćen za odabir hiperparametara kroz sto iteracija Bajesovske optimizacije sastojao se od dva puta ponovljene unakrsne validacije sa tri podele. Horizontalne isprekidane linije odgovaraju pragu odluke od 0,22 koji maksimizuje srednju vrednost balansirane tačnosti i pragu odluke od 0,333 koji odgovara specifičnosti od 0,95 (na osnovu srednje vrednosti).

Hiperparametri finalnog modela treniranog na lokalnim 15-mernim sekvencama odabrani su kroz sto iteracija Bajesovske optimizacije korišćenjem dva puta ponovljene unakrsne validacije sa tri podele. Na ovaj način odabrani su sledeći hiperparametri: *nrounds* = 379; *min_child_weight* = 1,357; *max_depth* = 5; *eta* = 0,031; *gamma* = 1,882; *colsample_bytree* = 0,749; *subsample* = 0,589; *alpha* = 0,457; *lambda* = 2,329 i *colsample_bylevel* = 0,98. Ovi hiperparametri su iskorišćeni za treniranje modela korišćenjem celog skupa za treniranje konstruisanog na osnovu 15-mernih lokalnih sekvenci. Dobijeni model je potom evaluiran korišćenjem skupa za evaluaciju - nezavisnih podataka koji nisu ni na koji način učestvovali u procesu pravljenja modela.

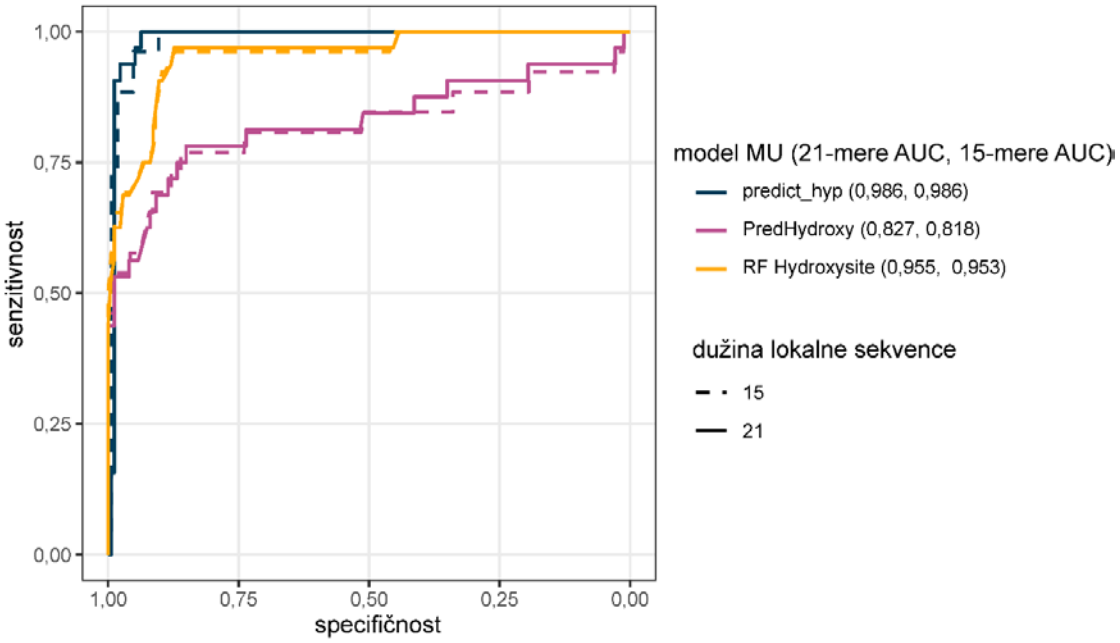
Korišćenjem podrazumevanog praga odluke (0,22) model je ostvario osetljivost od 0,962 (25 TP od 26 pozitivnih instanci) i specifičnost od 0,952 (157 TN od 165 negativnih instanci) uz AUC od 0,986 na skupu podataka za evaluaciju.

4.1.5. Poređenje performansi modela sa uspostavljenim serverima za predviđanje Hyp

Performanse modela treniranih na skupovima podataka konstruisanih na osnovu lokalnih sekvenci dužina 21 i 15 aminokiselina upoređene su sa nekoliko prethodno uspostavljenih servera za predviđanje Hyp: *RF- Hydroxysite* (Ismail i sar., 2016), *PredHydroxy* (Shi i sar., 2015), *iHyd-PseCp* (Qiu i sar., 2016) i *iHyd-PseAAC* (Xu i sar., 2014). Za poređenje performansi korišćen je skup sekvenci za evaluaciju. Prilikom ovog poređenja pokazalo se da serveri *PredHydroxy*, *iHyd-PseAAC* i *iHyd-PseCp* imaju relativno nisku osetljivost (0,4 - 0,47) i visoku specifičnost (0,84 - 0,99) predviđanja. Moguće objašnjenje je nedovoljna zastupljenosti biljnih proteinskih sekvenci u skupovima koji su korišćeni za treniranje ovih algoritama, tako da su njihova predviđanja vođena obrascima hidroksilacije Pro u životinjskim proteinskim sekvencama. Nasuprot pomenutim serverima, *RF-Hydroxysite* je postigao značajno bolje performanse sa AUC vrednošću od 0,955 (Tabela 17, Slika 9) uz osetljivost 0,969 i specifičnost 0,828 korišćenjem podrazumevanog praga odluke. Glavna mana *RF-Hydroxysite* servera je to što dozvoljava istovremenu analizu samo jedne proteinske sekvence, odnosno on nije napravljen za analize visoke propusnosti kada je potrebno analizirati cele proteome. Na osnovu skupa sekvenci za evaluaciju, prethodno opisani modeli koji su inkorporirani u *ragp* R paket imaju nešto nižu osetljivost u poređenju sa *RF-Hydroxysite* serverom i značajno veću specifičnost.

Tabela 17. Poređenje performansi nekoliko metoda za predviđanje pozicija hidroksilacije prolina na proteinskim sekvencama za evaluaciju. Podvučene su maksimalne vrednosti ostvarene za svaku metriku.

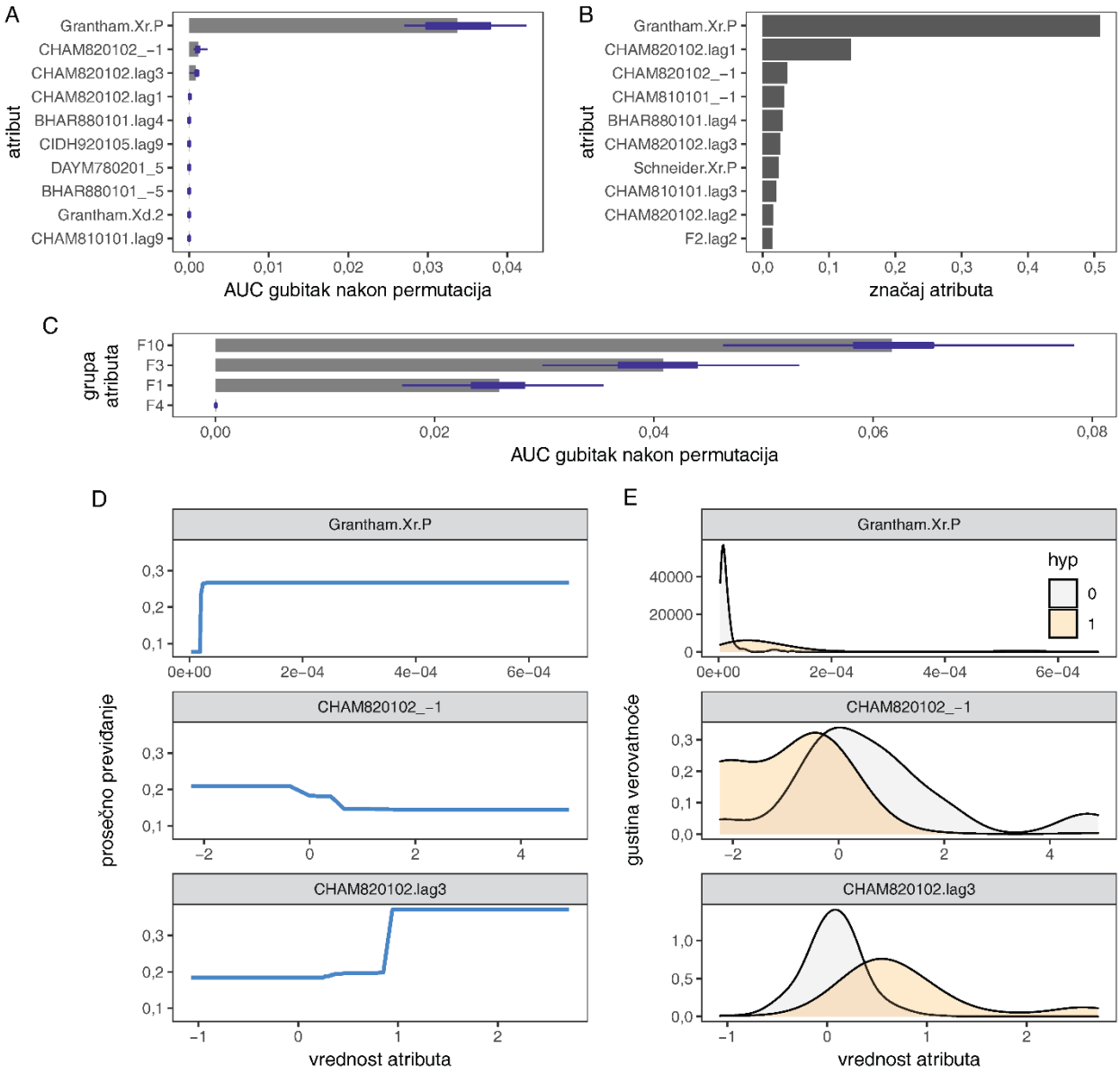
model	dužina lokalne sekvence	tačnost	balansirana tačnost	osetljivost	specifičnost	*Cohen* kappa	MCC	AUC
RF Hydroxysite	21	0,850	0,898	**0,969**	0,828	0,581	0,632	0,955
PredHydroxy		0,908	0,729	0,469	**0,989**	0,565	0,602	0,827
iHyd PseAAC		0,777	0,638	0,438	0,839	0,245	0,249	-
iHyd PseCp		0,859	0,674	0,406	0,943	0,394	0,401	-
ragp		**0,966**	**0,954**	0,938	0,971	**0,875**	**0,877**	**0,986**
RF Hydroxysite	15	0,843	0,893	**0,962**	0,824	0,541	0,598	0,954
PredHydroxy		0,916	0,725	0,462	**0,987**	0,558	0,591	0,818
iHyd PseAAC		0,785	0,649	0,462	0,836	0,246	0,253	-
iHyd PseCp		0,869	0,681	0,423	0,939	0,394	0,397	-
ragp		**0,953**	**0,957**	**0,962**	0,952	**0,820**	**0,828**	**0,986**



Slika 9. Poređenje nekoliko algoritama za predviđanje pozicija hidroksilacije prolina. ROC krive dobijene su na osnovu predviđanja na skupu proteinskih sekvenci za evaluaciju modela korišćenjem *RF Hydroxysite*, *PredHydroxy* i *ragp* modela. AUC vrednosti za svaki model su označene u legendi.

4.1.6. Interpretacija predviđanja modela

Značajnost i uticaj različitih atributa koji su korišćeni za opisivanje 21-mernih sekvenci na predviđanja modela su ispitani na nekoliko načina. Permutaciona analiza je ukazala da su samo tri atributa značajna, odnosno da njihovo isključivanje iz modela dovodi do pada performansi (Slika 10A). Grantham.Xr.P je atribut sa najvećim značajem, on pripada F10 grupi atributa i to X_r komponenti (odeljak Transformacija proteinskih sekvenci u numerički oblik 3.1.1.). On predstavlja količnik normalizovane učestalosti prolina u sekvenci i skalirane sume F9 deskriptora. Povećanje vrednosti ovog atributa model tumači kroz povećanje verovatnoće za hidroksilaciju prolina (Slika 10D). Drugim rečima što je više prolina u lokalnoj sekvenci oko posmatranog, to je veća verovatnoća da na posmatranom prolinu dođe do hidroksilacije. Na osnovu distribucije ovog atributa jasno je da su retke 21-merne sekvence gde ovaj atribut ima nešto višu vrednost, a u čijoj sredini P nije hidroksilovan (Slika 10E).



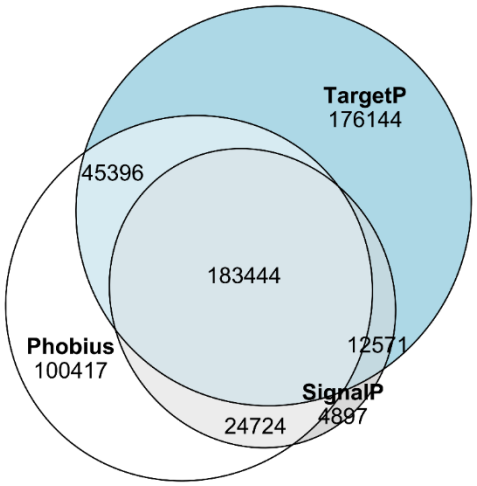
Slika 10. Interpretacija predviđanja modela treniranog na 21-mernim sekvencama. **A**. Značajnost atributa u permutacionoj analizi. Prikazano je 10 najznačajnih atributa čija permutacija dovodi do najvećeg gubitka performansi. Dužina sive kolone predstavlja srednju vrednost gubitka performansi, a varijabilnost usled permutacija je opisana grafikom raspodele podataka plave boje. **B**. Značaj atributa evaluiran sumiranjem povećanja homogenosti koji oni indukuju u svakom čvoru svakog stabla odlučivanja koje je sastavni deo XGB ansambla. **C**. Značajnost grupa atributa F1, F3, F4 i F10 u permutacionoj analizi. Dužina sive kolone predstavlja srednju vrednost, a varijabilnost usled permutacija je opisana grafikom raspodele podataka plave boje. **D**. Zavisnosti parcijalnog uticaja atributa na predviđanja verovatnoće pripadnosti pozitivnoj klasi za tri najznačajnija atributa na osnovu permutacione analize. **E**. Distribucija najznačajnijih atributa u zavisnosti od statusa hidroksilacije prolina u okviru 21-mernih sekvenci u skupu podataka za treniranje.

Druga po važnosti je vrednost CHAM820102 (pripada F1 grupi atributa) fizičko-hemijske odlike aminokiselina (predstavlja slobodnu energiju aminokiselina u rastvoru) u aminokiselini koja prethodi posmatranom prolinu u lokalnoj sekvenci (-1 koordinata). Vrednosti manje od 0 ovog atributa model tumači kao povećanje šanse za hidroksilaciju prolina (Slika 10D). Aminokiseline koje su asocirane sa pomenutim vrednostima su P, R, G, S i A; dakle kada one prethode posmatranom prolinu verovatnoća da bude hidroksilovan raste. CHAM820102.lag3 atribut, koji je treći po značaju, pripada Moreau-Broto autokorelacionom deskritporu proteina (F3) koji je izračunat na osnovu standardizovane CHAM820102 fizičko-hemijske odlike aminokiselina. Vrednosti ovog atributa koje su više od 0,8 u lokalnoj 21-mernoj sekvenci model tumači kroz povećanje verovatnoće hidroksilacije posmatranog prolina (Slika 10D). Radi lakšeg razumevanja uticaja ovog atributa generisano je million nasumičnih proteinskih sekvenci dužine 21 aminokiselinu koje su imale P u sredini u kojima je ispitana distribucija pomenutog atributa. Samo je ~ 0,08% generisanih sekvenci imalo vrednost ovog atributa višu od 0,8, a one su bogate P, Y i C aminokiselinama, dok se među deset 10 najfrekventnijih dipeptida u ovim sekvencama nalaze PP (EXT motiv), SP, PG i PS (AG motivi). Prema permutacionoj analizi (Slika 10A) izgleda kao da ostali atributi nisu bili ni potrebni za treniranje modela, pošto njihova permutacija vrlo malo ili ni malo ne utiče na performanse modela. Ovakvo tumačenje ne bi bilo korektno, već je verovatnije da među korišćenim atributima ima nekoliko grupa visoko korelisanih tako da permutovanje samo jednog od njih ne narušava značajno performanse modela. Ovo je jasno na osnovu permutacione analize grupa atributa (Slika 10C) gde se vidi da permutacija F1, F3 i F10 grupa atributa dovodi do pada performansi modela koji je veći od gubitka opaženog samo kada se najznačajniji atributi iz ovih grupa permutuju (Slika 10A). Takođe se na osnovu važnosti atributa koja je inherentna za XGB algoritam (Slika 10B) može primetiti da veći broj atributa ima značaj za predviđanja modela.

4.1.7. Anotacija HRGP sekvenci iz 62 biljna proteoma

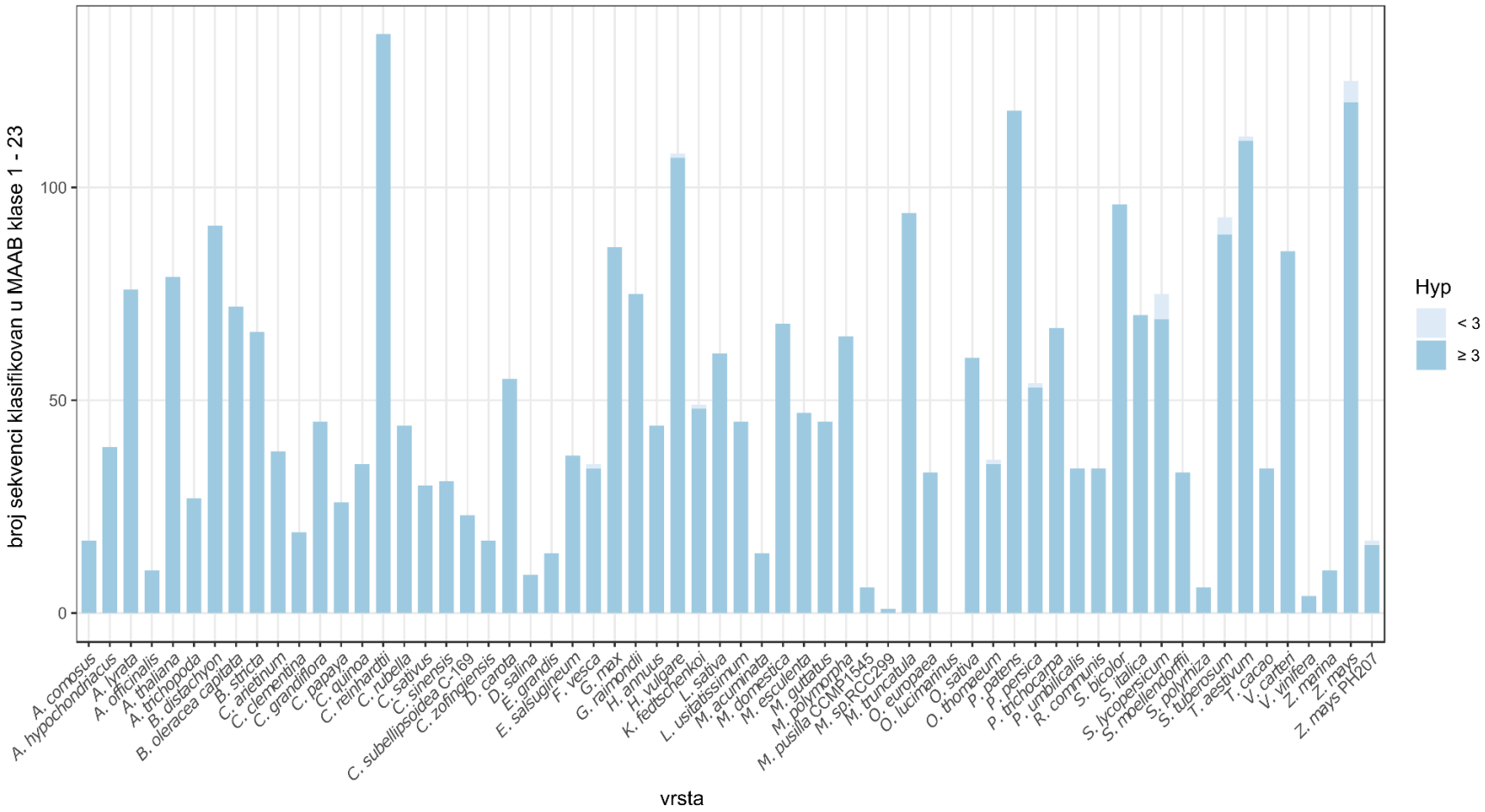
Korišćenjem *ragp* paketa analizirane su predviđene proteinske sekvence 62 biljna proteoma preuzeta iz Phytozome V12 baze. Ukupno je analizirano nešto manje od tri miliona biljnih proteinskih sekvenci (2797062). Kompletna anotacija je dostupna na *Zenodo* platformi (doi: 10.5281/zenodo.2605302, url: https:// zenodo.org/record/2605302) dok će u narednim redovima postupak i rezultati anotacije biti sumirani.

Prvi korak u analizi HRGP sekvenci implementiranoj u okviru *ragp* R paketa je identifikacija sekvenci koji sadrže N-sp pošto su HRGP lokalizovani u ćelijskom zidu. Ovo je omogućeno kroz komunikaciju sa Internet serverima za ovu namenu: *Phobius* (Käll i sar., 2007), *SignalP 4.1* (Petersen i sar., 2011) i *TargetP 1.1* (Emanuelsson i sar., 2007). Rezultati predviđanja servera (Slika 11) ukazuju da različite metode nisu u potpunosti usaglašene. *SignalP 4.1* je najkonzervativniji algoritam, dok je *TargetP 1.1* sa podrazumevanim podešavanjima najmanje striktan (Slika 11). Zbog relativno velikog neslaganja između predviđanja ova tri algoritma kao kriterijum za određivanje da li proteinska sekvenca sadrži N-sp ili ne, korišćen je većinski glas predviđanja ova tri servera što je rezultovalo sa 266135 sekvenci od 2797062 analiziranih sekvenci proteina (9,5%) za koje je predviđeno da sadrže N-sp . U ovim sekvencama je dalje predviđano prisustvo i lokacija potencijalnih hidroksiprolina i 69982 proteinske sekvence (26,3% od proteina koji se sekretuju) su imale tri ili više predviđenih hidroksiprolina.



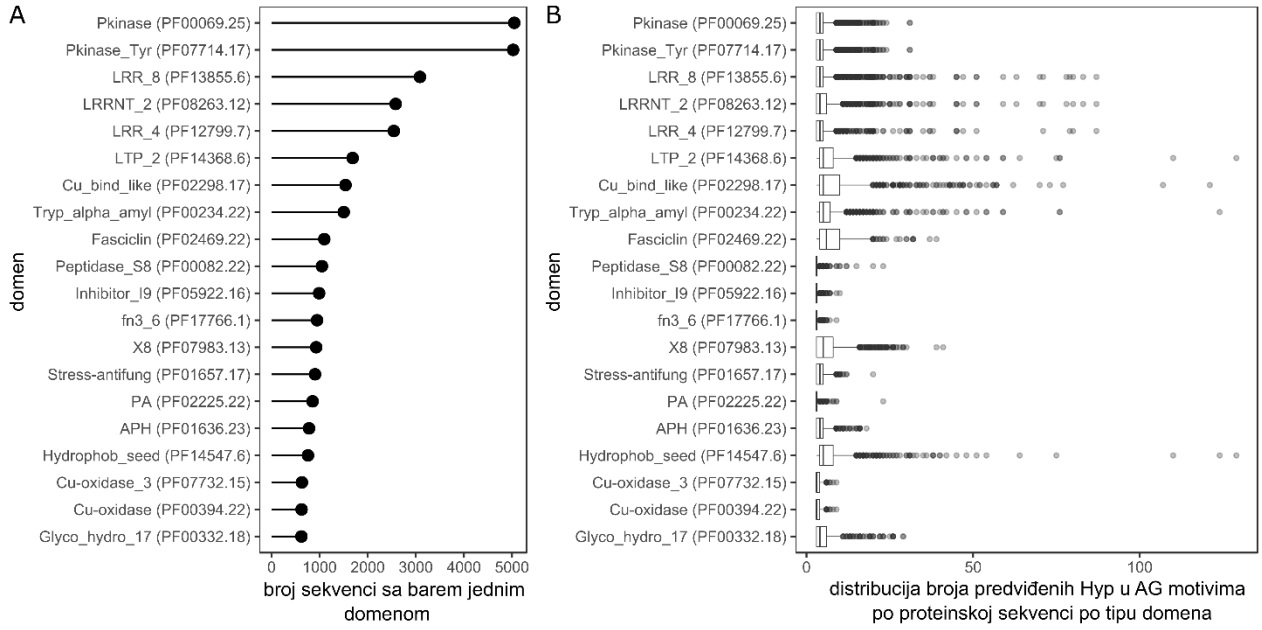
Slika 11. Saglasnost predviđanja N-sp. Ojlerov (*Euler*) dijagram predviđanja N-sp od strane *Phobius*, *SignalP 4.1* i *TargetP 1.1* servera. Ukupno je korišćeno 2797062 proteinskih sekvenci iz 62 biljne vrste (Phytozome V12). Za 266135 proteinskih sekvenci je predviđeno da imaju N-sp sa barem dva Internet servera.

MAAB klasifikacija HRGP sekvenci je određena i na skupu svih sekvenci za koje je bilo predviđeno da sadrže N-sp (266135 proteinskih sekvenci) i na skupu sekvenci koji sadrže tri ili više predviđenih hidroksiprolina (69982 proteinskih sekvenci), pa su rezultati upoređeni (Slika 12). Ukupno je 3075 sekvenci klasifikovano u MAAB klase 1-23 (prototipski HRGP, Tabela 1). Zelena alga *Chlamydomonas reinhardtii* imala je najveći broj ovakvih sekvenci (136), dok zelena alga *Ostreococcus lucimarinus* nije imala nijednu sekvencu klasifikovanu u MAAB klase 1-23. Sve sekvence klasifikovane u MAAB klase 1-23 iz 51 analiziranog organizma sadržale su barem tri predviđena hidroksiprolina, dok kod preostalih 10 organizama za samo nekoliko (ukupno 22) MAAB klasifikovanih sekvenci nije predviđeno da sadrže barem tri hidroksiprolina. Zanimljivo je da nijedna sekvenca nije klasifikovana u MAAB klase 13, 14 i 17 (Tabela 1).



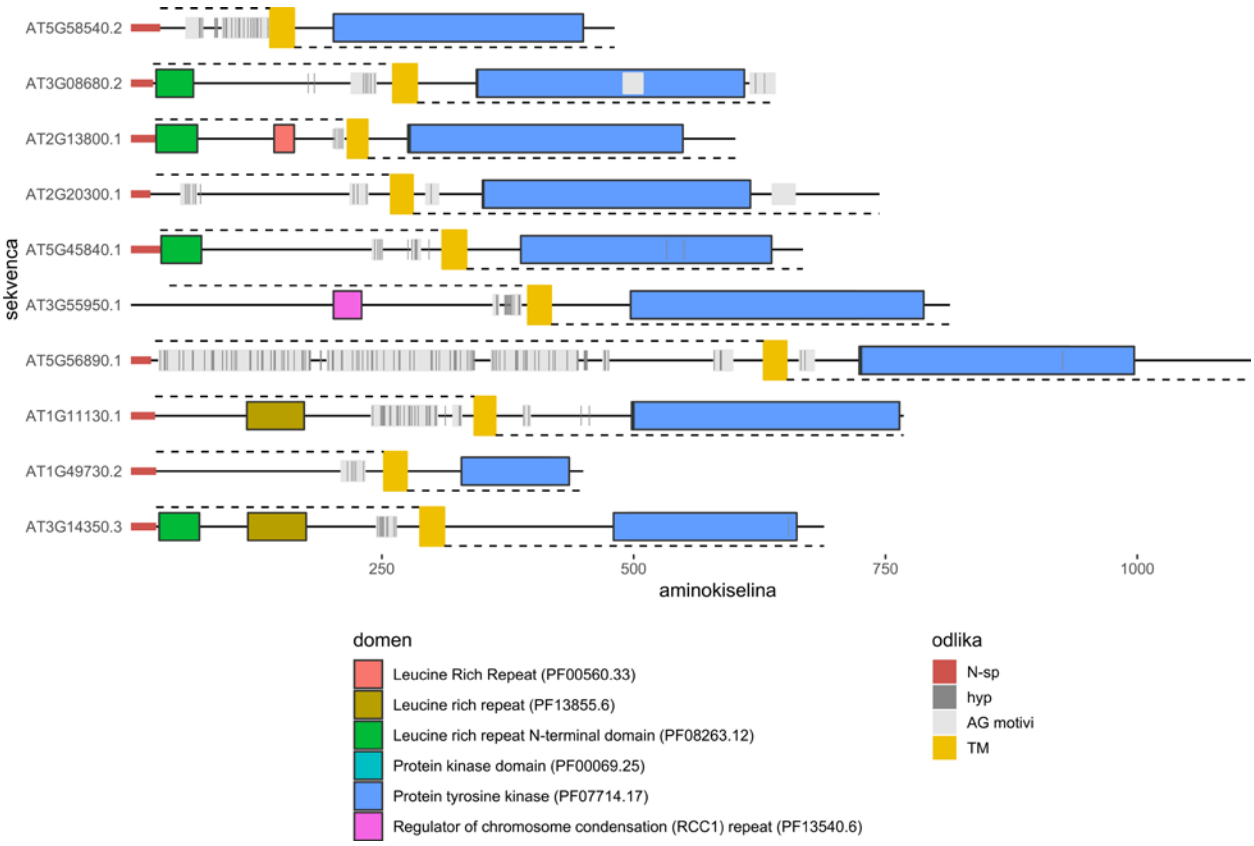
Slika 12. MAAB klasifikacija HRGP sekvenci u 62 biljna proteoma. Prikazan je broj sekvenci koje su klasifikovane kao prototipske HRGP (MAAB klase 1-23) u 62 analizirana biljna proteoma. Sekvence su grupisane po broju predviđenih hidroksiprolina kao što je prikazano u legendi.

Pored MAAB klasifikacije prototipskih HRGP sekvenci, urađena je i pretraga AG motiva na svih 266135 sekvenci za koje je predviđeno da imaju N-sp signalnu sekvencu. Samo predviđene pozicije Hyp su uzimane u razmatranje prilikom pretrage AG motiva, koji je definisan kao najmanje tri dipeptida (AO, TO, SO, GO, VO, OA, OT, OS, OG i OV) sa razmakom od najviše deset aminokiselina između bilo koja dva dipeptida. Ukupno je 36732 proteinske sekvence iz 62 biljne vrste posedovalo barem jedan AG motiv definisan na ovaj način. Za identifikaciju domena u ovim sekvencama (identifikacija himernih AGP), korišćen je *hmmer3* sa *Pfam32* bazom (Slika 13). Najčešće identifikovani domeni u sekvencama koje sadrže AG motive bili su protein kinazni (PK, PF00069) i protein tirozin kinazni (PTK, PF07714), koji se zajednički identifikuju i preklapaju u istim sekvencama (Slika 13A). Treba napomenuti da je u novijim verzijama *Pfam* baze (od verzije 33 *Pfam* baze, odnosno od PF07714.18) anotacija PF07714 domena promenjena iz protein tirozin kinazni domen u tirozin i serin/treonin receptor kinazni domen. Posle receptor kinaza po učestalosti detekcije slede leucinom bogati regioni (LRR_8, LRRNT_2 i LRR_4), koji se često mogu naći zajedno sa PK/PTK domenima u istim sekvencama.



Slika 13. Raspodela domena i predviđenih hidroksiprolina u sekvencama sa AG motivima. **A**. Najfrekventnijih 20 domena identifikovanih u sekvencama sa AG motivima. Domeni su identifikovani sa *hmmer3* korišćenjem *Pfam 32* baze podataka. U razmatranje su uzimani domeni sa nezavisnom e-vrednosti < 0,01 i svaki domen je brojan jedanput po sekvenci. **B**. Raspodela broja predviđenih hidroksiprolina u AG motivima po sekvenci među sekvencama koje sadrže barem jedan od najfrekventnijih 20 domena.

S obzirom da nema puno podataka u literaturi o vezi PK/PTK, kao najčešće identifikovanog domena prema analizi urađenoj ovde, sa arabinogalaktanskim regionima, ispitana je strukturna organizacija nekoliko identifikovanih himernih AGP nalik receptornim kinazama (engl. „*kinase like AGP*" - KLA) *A. thaliana* (Slika 14). AG motivi sa predviđenim hidroksiprolinima su lokalizovani na ekstracelularnoj strani pomenutih receptora, dok su kinazni domeni na intracelularnoj strani (Slika 14). Identifikovane KLA često sadrže i leucin bogate regione (LRR) takođe na ekstracelularnoj strani.

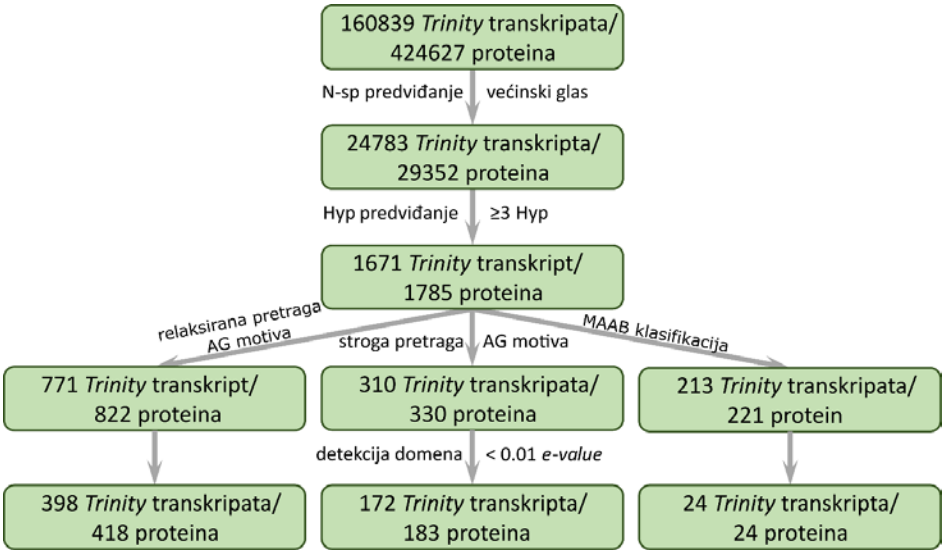


Slika 14. Šematski prikaz strukture nekoliko KLA *A. thaliana* koje sadrže AG motive. Struktura proteina je vizuelizovana korišćenjem *ragp* funkcije *plot_prot*. Transmembranski (TM) regioni su prikazani žutom bojom; ekstracelularni regioni su označeni isprekidanom linijom iznad sekvence, dok su intracelularni regioni označeni isprekidanom linijom ispod sekvence (predviđeni od strane *Phobius* pomoću *ragp* funkcije *get_phobius*); signalni peptidi (predviđeni od strane *SignalP* pomoću *ragp* funkcije *get_signalp*) označeni su zadebljalom crvenom linijom na N-terminalnom kraju sekvence; Hyp (predviđeni *predict_hyp ragp* funkcijom) su označeni vertikalnim, tamno sivim linijama; nizovi AG glikomodula (predviđeni *scan_ag ragp* funkcijom) označeni su svetlo sivom pozadinom; domeni (identifikovani *get_hmm ragp* funkcijom) su označeni pravougaonicima sa odgovarajućom bojom kao što je naznačeno u legendi.

Treba dodati da su neki od identifikovanih himernih AGP potencijalno lažni. Naime kod serinskih proteaza subtilaza koje sadrže domene subtilizin-nalik (engl. *„Peptidase_S8“* PF00082), inhibitor peptidaza I9 (engl. *„Inhibitor_I9“* PF05922), fibronektin tip III (*fn3_6*, PF17766) i domen asociran sa proteazama (*PA*, PF02225) je u većini slučajeva predviđen mali broj Hyp (Slika 13). Osim toga veliki broj ovih sekvenci je uklonjen neznatno strožim uslovima detekcije AG motiva (povećanje broja AG motiva na četiri umesto tri). Slična je situacija i sa bakar oksidazama koje sadrže domene *Cu-oxidase* (PF00394) i *Cu-oxidase 3* (PF07732).

4.2. Identifikacija i analiza *HRGP* gena *C. erythraea*

Kao polazni materijal za identifikaciju *HRGP* gena *C. erythraea* korišćen je *de novo* sastavljen transkriptom sa 160839 *Trinity* transkripata grupisanih u 105726 *Trinity* gena (Ćuković i sar., 2020). Referentni transkriptom je analiziran korišćenjem *ragp* paketa. Kao prvi korak u filtriranju određeno je prisustvo N-sp sekretornih signalnih sekvenci. Ovo je urađeno većinskim glasom između predviđanja *Phobius 1.01*, *SignalP 4.1* i *TargetP 1.1* servera kao što je opisano u odeljku 4.1.7. Na ovaj način su odabrane proteinske sekvence kodirane od strane 24783 *Trinity* transkripta. Predviđanjem verovatnoće hidroksilacije prolina u ovim proteinskim sekvencama identifikovano je 1785 proteinskih sekvenci poreklom od 1671 *Trinity* transkripta (1063 *Trinity* gena) koji sadrže barem tri predviđena hidroksiprolina (Slika 15).

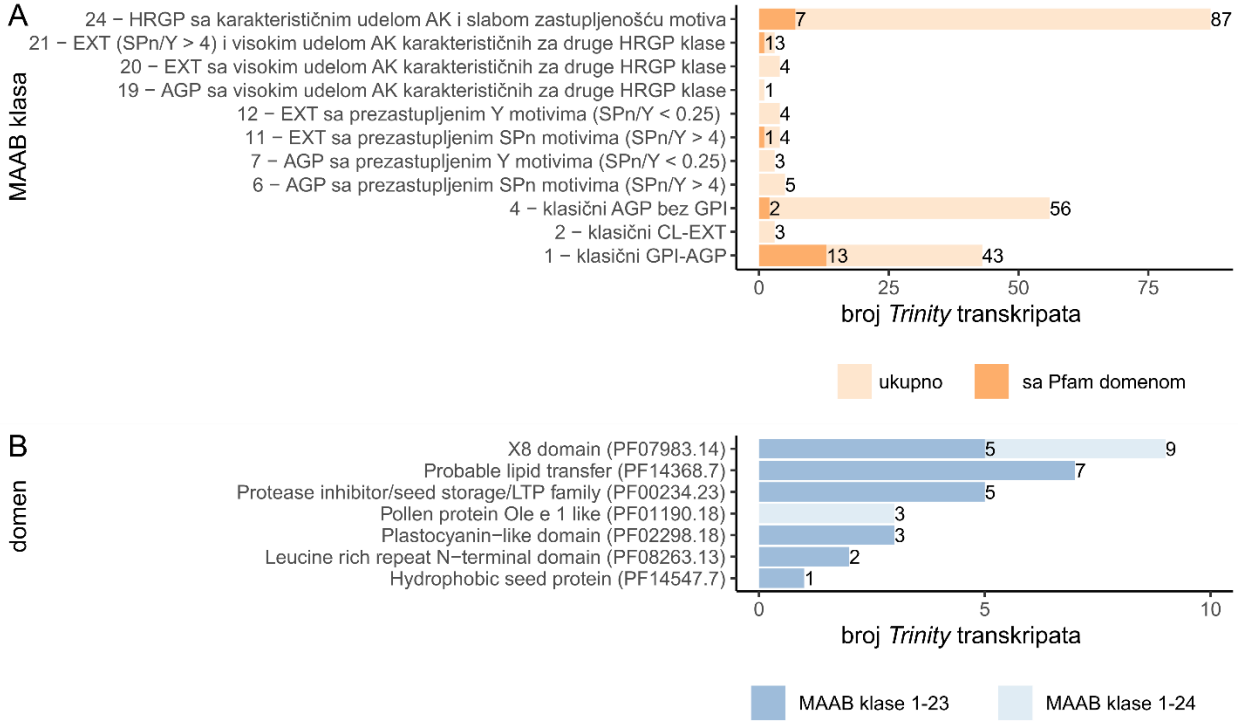


Slika 15. Broj *Trinity* transkripata i predviđenih sekvenci proteina *C. erythraea* koje su prošle različite stupnjeve filtriranja metodologijom inkorporiranom u *ragp* R paket.

4.2.1. MAAB klasifikacija HRGP sekvenci *C. erythraea*

MAAB klasifikacija je izvedena pomoću implementacije u *ragp* paketu, a za predviđanje C-terminalnog signalnog peptida za dodatak GPI, što je neophodan korak za razlikovanje pojedinih MAAB klasa (Tabela 1), je korišćen *NetGPI1.1* server. MAAB klasifikacijom identifikovana je 221 proteinska sekvenca (poreklom od 213 *Trinity* transkripata) koje pripadaju MAAB klasama od 1 do 24 (Slika 15 i 16A). Najveći broj ovih sekvenci, 87 *Trinity* transkripta, pripada MAAB klasi 24 koja obuhvata sekvence sa udelom aminokiselina karakterističnim za HRGP ali i niskim procentom HRGP motiva (pokrivenost sekvence HRGP motivima je manja od 15%). Od preostalih 126 *Trinity* transkripata koje pripadaju MAAB klasama 1 - 23, 99 *Trinity* transkripta je klasifikovano kao klasični AGP (MAAB klase 1 i 4), od toga 43 *Trinity* transkripta kao MAAB klasa 1 koja obuhvata klasične AGP sa predviđenim GPI signalnim peptidom i 56 *Trinity* transkripta kao MAAB klasa 4 koja obuhvata klasične AGP bez predviđenog C-terminalnog signalnog peptida za dodatak GPI (Slika 16A). Preostale MAAB klase su ili izostale ili su bile

prisutne tek sa nekoliko sekvenci. Identifikacija domena u ovim sekvencama urađena je sa *hmmer3* korišćenjem *Pfam33* baze. Najfrekventniji domeni su *Probable lipid transfer* (*Pfam*: PF14368.7), *Protease inhibitor/seed storage/LTP family* (PF00234.23) i *X8* (PF07983.14) (Slika 16B). Većina pomenutih domena asocirana je sa MAAB klasom 1 (klasični AGP sa GPI signalnim peptidom, Slika 16B). Domen *Pollen protein Ole e 1* (PF01190.18) identifikovan je samo u sekvencama koje su klasifikovane kao MAAB klasa 24 (Slika 16B).

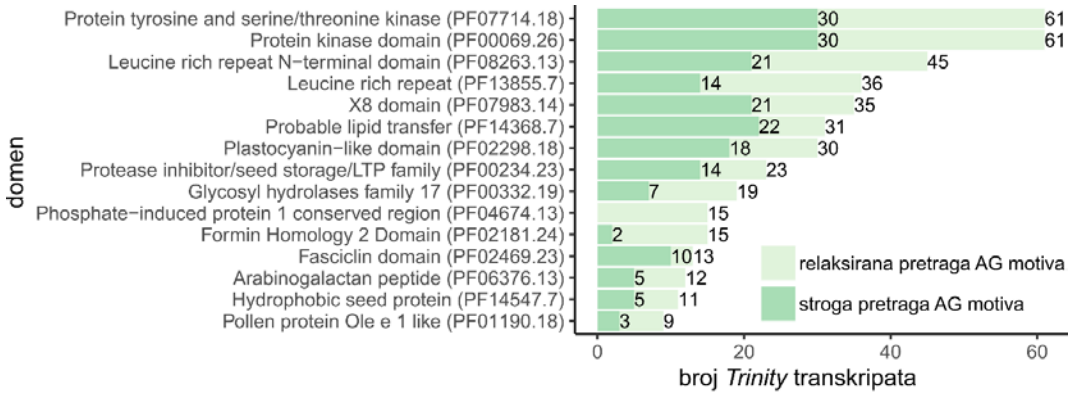


Slika 16. Raspodela identifikovanih HRGP sekvenci u transkiptomu *C. erythraea*. **A**. Raspodela identifikovanih HRGP MAAB klasa. **B**. Raspodela *Pfam* domena u sekvencama klasifikovanim u neku od MAAB klasa. Domeni su identifikovani sa *hmmer3* uz *Pfam* 33 bazu i nezavisnu e-vrednost od < 0,01.

4.2.2. Identifikacija AGP sekvenci *C. erythraea*

Za identifikaciju AGP sekvenci *C. erythraea* korišćena su dva tipa pretrage AG motiva: relaksirana pretraga, kojom su identifikovane sekvence sa najmanje tri dipeptida karakteristična za AGP (AO, TO, SO, GO, VO, OA, OT, OS, OG i OV) sa ne više od deset aminokiselina između kao što je opisano u odeljku 4.1.7. i stroga pretraga, kojom su identifikovane sekvence sa najmanje četiri prethodno pomenuta dipeptida razdvojena sa ne više od četiri aminokiseline (Slika 17). U obzir su uzimani samo karakteristični dipeptidi sa predviđenim hidroksiprolinima, a motivi sa tri ili više kontinualna prolina/hidroksiprolina (npr. AOOO ili OOOOS) koji su karakteristični za ekstenzine su maskirani (isključeni iz pretrage). Relaksiranom pretragom identifikovane su 822 AGP proteinske sekvence poreklom od 771 *Trinity* transkripta. Približno 40% ovih sekvenci (330 proteinskih sekvenci poreklom od 310 *Trinity* transkripata) identifikovano je i strogom pretragom. Najčešće identifikovani domeni u obe grupe sekvenci su bili receptor kinazni domeni (PF07714.18

i PF00069.26), zatim slede leucinom bogati ponovci (LRR, PF08263.13 i PF13855.7). Treba napomenuti da strogom pretragom nisu identifikovane sekvence koje imaju domen karakterističan za proteine indukovane fosfatima (engl. „*Phosphate−induced protein 1 conserved region*“, PF04674.13), a identifikovano je malo sekvenci sa forminskim domenom (engl. „*Formin Homology 2 Domain*“, PF02181.24).



Slika 17. Raspodela najfrekventnijih 15 *Pfam* domena u sekvencama *C. erythraea* identifikovanim relaksiranom detekcijom AG motiva (tri karakteristična AG dipeptida razdvojena sa ne više od 10 aminokiselina između) i sekvencama identifikovanim strogom detekcijom AG motiva (četiri karakteristična AG dipeptida razdvojena sa ne više od četiri aminokiseline između). Brojani su samo jedinstveni domeni po sekvenci, a za detekciju AG motiva razmatrani su samo karakteristični dipeptidi sa predviđenim hidroksiprolinima. Motivi sa tri ili više kontinualna prolina/hidroksiprolina (npr. AOOO ili SOOOO) karakteristčni za ekstenzine su maskirani prilikom detekcije AG motiva. Domeni su identifikovani sa *hmmer3* uz *Pfam 33* bazu i nezavisnu e-vrednost od < 0,01.

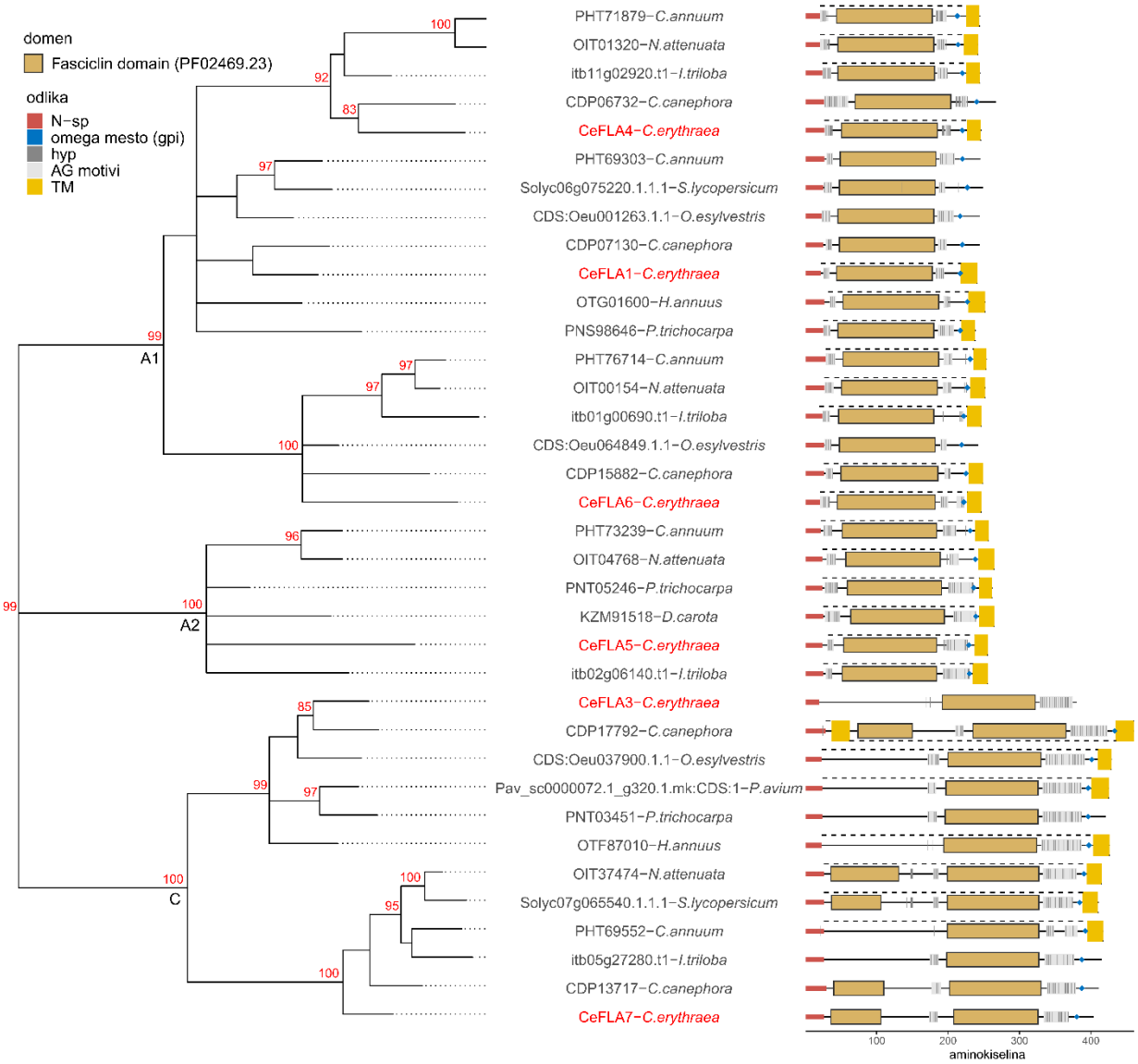
4.2.3. Strukturne odlike odabranih AGP sekvenci

Od identifikovanih sekvenci sa AG motivima, izabrano je 18 transkripata sa dobro definisanim regionima sa AG motivima, za koje je bilo moguće konstruisati specifične prajmere za kvantifikaciju samo konkretnog *Trinity* transkripta (Tabela 18). Pomenutih 18 transkripata činilo je po šest predstavnika KLA, FLA i AGp klasa AGP. Tri odabrane AGp sekvence su sadržale nedavno opisani arabinogalaktanski peptidni domen PF063765 (Simonović i sar., 2016).

Tabela 18. Osamnaest odabranih AGP sekvenci sa AG motivima koji su u njima detektovani.

ncbi pristup	naziv	AG motiv
MW792522	CeAGp3	30-SOAOAO-35
MW792523	CeAGp6	27-AOfeafAOAOAOT-39
MW792524	CeAGp7	34-AOAOAOA-40
MW792525	CeAGp8	24-SOqfSOeaAOTOepaflpOSOSOSlvaVOAOAOATOTOAOTOAOATOTO AOAOAO-78
MW792526	CeAGp9	40-SOAOAOAOA-48
MW792527	CeAGp10	58-OVhqiSOrAO-67
MW792515	CeFLA1	26-SOGOSGOT-33; 183-GOAVOAOAOSOA-194
MW792516	CeFLA3	329-AOSOAOGOAOetSOSOSOlGOSOAeSOlSOSOSlSOOAOOTaSOA-373
MW792517	CeFLA4	28-AOTOAOG-34; 197-AOAOAO-202
MW792518	CeFLA5	32-AOAOAOAO-39; 199-SOVAOSOktTOAadAOAgsassknSOOSO-227
MW792519	CeFLA6	22-TOAAOAOAOAGO-33; 190-AOAOSOOAOS-199; 213-VOSseGOSOS-222
MW792520	CeFLA7	174-SOeasAOTOSOS-185; 335-AOTOSOAOAOGOAahkkhkSOOAOOSad AOAdGO-368
MW792530	CeKLA1	243-SOSOSOSOqsaSOlfssaeAOfaspklSOlestsSOVeAOSOShpglktttSO-295
MW792533	CeKLA2	40-SOeaallqpoqsSOoinsVOSO-61; 73-SOTnlllpOVavTOOVnOTgnitalTOSAOodAOAhsehpipdTOAOlp SOSgvSOnnOAvSOGlSOOSOlOGOhgnvsteoOAipaoooaAOVgiAOVlp OThsvepeipOSsfpVOAvlpOSmSO-199; 224-OAAOOVOSrngvgVOTaAOoheiinhhSOknOShgkglphlSOSmSOV-271; 297-SOAooooSOeqGOksnssfhAOSOSksvsliAOSvSOOSOSmghnssA
MW792534	CeKLA3	253-OSOStnvssdoloOAOSOG-271
MW792528	CeKLA5	51-VOOSOlrVOOSOSrfamSO-69
MW792532	CeKLA6	243-GOAooooogTOOVrOGgnOS-262
MW792531	CeKLA7	450-SOGnonigndhOSOGTOGSOodTOG-474

Filogenetska analiza odabranih sekvenci urađena je poređenjem sa sličnim sekvencama iz osamnaest biljnih vrsta (Slika 18, 19, 20). U cilju boljeg razumevanja dobijenih filogenetskih modela (dendograma), napravljeni su šematski strukturni dijagrami proteina za sekvence iz svakog dendograma.

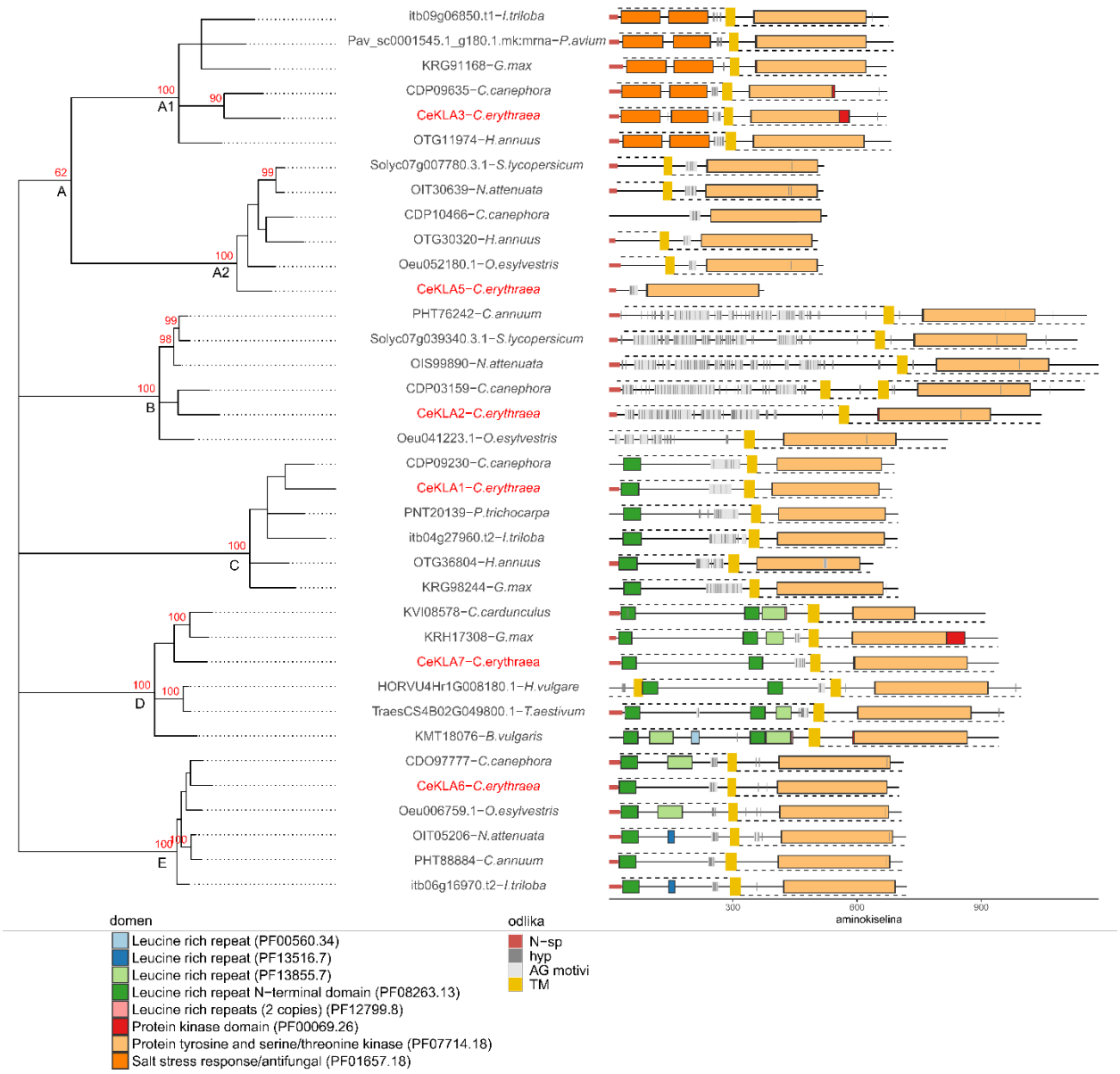


Slika 18. Poređenje filogenetskih odnosa FLA proteinskih sekvenci *C. erythraea* sa homolognim sekvencama iz drugih biljnih vrsta. Filogenetsko stablo predstavlja neukorenjeno stablo najveće verovatnoće (engl. "*unrooted maximum likelihood tree*") konstruisano korišćenjem *WAG* (Whelan i Goldman, 2001) modela izmene aminokiselina. Stabilnost klastera je procenjena upotrebom 100 ponavljanja neparametrijskog *bootstrap*-a. Klasteri sa ≥ 80/100 *bootstrap* podrškom označeni su crvenim brojevima. Klasteri sa ≤ 50/100 *bootstrap* podrškom su spojeni u politomije. Šematski dijagrami proteina su konstruisani korišćenjem *ragp* R paketa: N-sp sekvence su predviđene sa *Signalp4.1* i označene sa crvenim zadebljalim linijama na N-terminusu; mesta dodavanja GPI (ω-mesta) predviđena korišćenjem *NetGPI1.1* predstavljena su plavim romboidima; domeni identifikovani sa *hmmscan 3.3.2* korišćenjem *Pfam* 33 baze predstavljeni su kao što je prikazano na legendi pored slike; transmembranski regioni (TM) predviđeni korišćenjem *Phobius1.01* predstavljeni su žutim pravougaonicima; ekstracelularni regioni označeni su isprekidanim linijama iznad dijagrama sekvenci; predviđeni Hyp predstavljeni su vertikalnim crnim linijama, a regioni sa AG motivima su označeni kao svetlo sivi pravougaonici. Klade su označene prema He i sar. (2019).

Filogenetsko stablo FLA sastoji se od tri glavne klade obeležene sa A1, A2 i C (Slika 18). Sekvence iz klada A1 i A2 pripadaju FLA grupi A, dok sekvence iz klade C pripadaju FLA grupi C prema He i sar. (2019). Pripadnost kladama je određena na osnovu sličnosti celih

sekvenci i izolovanih FAS domena sa odgovarajućim FLA sekvencama *Arabidopsis thaliana*. Klaster A1, koji obuhvata CeFLA4, CeFLA1 i CeFLA6 sekvence, sastoji se od sekvenci sa jednim fasciklinskim (FAS) domenom i dva AGP regiona koji ga okružuju (Slika 18). Klaster A2 obuhvata sekvence strukturno slične sekvencama u A1, dok se klaster C sastoji od dužih FLA sekvenci koje sadrže jedan FAS domen bliže C-terminusu ili dva FAS domena (Slika 18). Sekvence iz C klastera sa dva FAS domena (kao CeFLA7) sadrže dva AG regiona i to kraći između dva FAS domena i duži blizu C-terminusa. Sekvence iz C klastera sa jednim FAS domenom (kao CeFLA3) sadrže kraći AG region pre FAS domena i duži posle FAS domena. Od analiziranih *C. erythraea* FLA sekvenci, četiri (CeFLA4, CeFLA1, CeFLA3 i CeFLA6) su filogenetski najbliže odgovarajućim FLA sekvencama *Coffea canephora* što je posledica taksonomske bliskosti ove dve vrste koje pripadaju istom rodu - Gentianales. Svi razmatrani FLA (sem CeFLA3) sadrže predviđeno GPI ω mesto (mesto pričvršćivanja GPI sidra). Ovo je verovatno posledica toga što je CeFLA3 delimična sekvenca bez kompletnog C-terminusa.

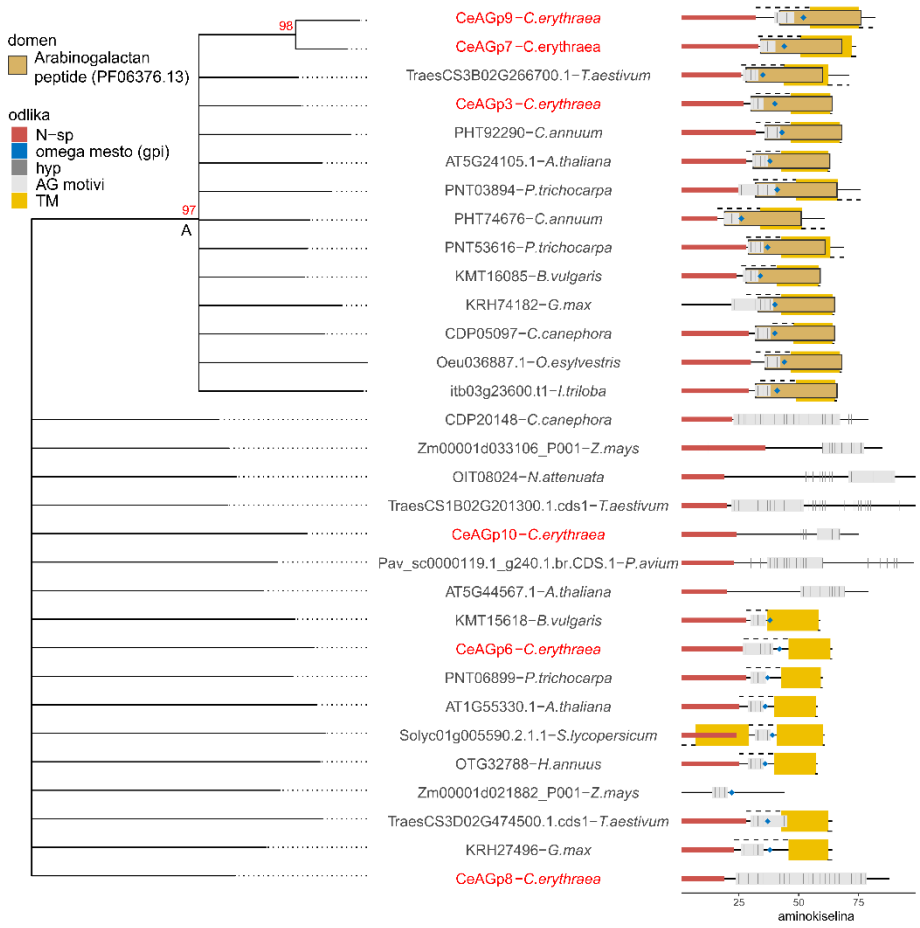
Filogenetsko stablo KLA sekvenci (Slika 19) sastoji se od pet klastera obeleženih sa A-E. Većinu KLA sekvenci odlikuje prisustvo transmembranskog regiona blizu sredine sekvenci, posle koga sledi intracelularni kinazni region. Klaster A sa sastoji se od dva stabilna podklastera obeležena sa A1 i A2. Podklaster A1 je sastavljen od sekvenci koje sadrže po dva domena odgovorna za odgovor na stres izazvan solima/antifungalni odgovor (PF01657) na ekstracelularnoj strani nakon čega sledi kratak region bogat predviđenim hidroksiprolinima neposredno pre transmembranskog dela sekvence. Kod tri sekvence, CeKLA3 - *C. erythraea*, CDP09635 - *C. canepohora* i OTG11974 - *H. annuus,* predviđeni Hyp se nalaze u sklopu AG motiva. Podklaster A2 se sastoji od sekvenci koje nemaju druge domene osim kinaznog. Arhitektura sekvenci u A2 odudara od ostalih analiziranih: dve sekvence nemaju predviđen TM region (CeKLA5 - *C. erythraea* i CDP10466 - *C. canephora*), dok četiri imaju predviđeni TM, ali su detektovani AG motivi i predviđeni Hyp u intracelularnim regionima. Predviđanja Hyp u intracelularnim regionima sekvenci treba uzeti sa rezervom pošto su modeli za predviđanje Hyp inkorporirani u *ragp* trenirani predominantno na ekstracelularnim delovima sekvenci. Klaster B, čiji je predstavnik CeKLA2 sekvenca iz kičice, se sastoji od prolinom-bogatih receptor kinaza sa dugim ekstracelularnim regionom koji podseća na hibridne HRGP pošto sadrži AG i ekstenzinske motive (hibridni PERK). Klasteri C, D i E sastoje se od KLA sa leucinom bogatim regionima u ekstracelularnom delu (kao kod CeKLA1, CeKLA6 i CeKLA7 sekvenci). Klaster C (predstavnik iz kičice je CeKLA1 sekvenca) je sastavljen je sekvenci sa jednim N-terminalnim leucinom bogatim regionom i dugim AG regionom koji prethodi TM segmentu. Klaster D (CeKLA7) čine duge KLA sekvence sa nekoliko LRR regiona i slabo izraženim AG karakteristikama (nekoliko sekvenci ima kratke AG motive sa nekoliko predviđenih Hyp, dok KVI08578 - *C. cardunculus* i KMT18076 - *B. vulgaris* nemaju detektovane AG motive). Klaster E (CeAKLA6) čine sekvence sa jednim ili dva LRR regiona i kratkim AG motivima u ekstracelularnom regionu, dok jedna od sekvenci u ovom klasteru (Oeu006759.1 - *O. esylvestris*) poseduje samo ekstenzinski region. Slično filogenetskom stablu FLA, četiri od šest KLA proteinskih sekvenci kičice CeKLA6, CeKLA1, CeKLA3 i CeKLA2 su u bliskim filogenetskim odnosima sa KLA sekvencama *C. canephora*.



Slika 19. Filogenetski odnosi šest izabranih KLA proteinskih sekvenci *C. erythraea* sa homolognim sekvencama iz drugih biljnih vrsta. Filogenetsko stablo predstavlja neukorenjeno stablo najveće verovatnoće (engl. "*unrooted maximum likelihood tree*") konstruisano korišćenjem *JTT* (Jones i sar., 1992) modela izmene aminokiselina. Stabilnost klastera je procenjena upotrebom 100 ponavljanja neparametrijskog *bootstrap*-a. Klasteri sa ≥ 80/100 *bootstrap* podrškom označeni su crvenim brojevima. Klasteri sa ≤ 50/100 *bootstrap* podrškom su spojeni u politomije. Vrednosti *bootstrap* podrške od 50 do 80/100 označene su na glavnim klasterima. Šematski dijagrami proteina su konstruisani korišćenjem *ragp* R paketa: N-sp sekvence predviđene korišćenjem *Signalp4.1* su označene kao crveni segmenti na N-terminusu; mesta dodavanja GPI (ω) predviđena korišćenjem *NetGPI1.1* predstavljena su plavim romboidima na grafiku; domeni detektovani korišćenjem *hmmscan 3.3.2* i *Pfam*33 baze podataka predstavljeni su kao što je prikazano u legendi pored slike; transmembranski regioni (TM) predviđeni korišćenjem *Phobius1.01* predstavljeni su žutim pravougaonicima; ekstracelularni regioni su označeni isprekidanim linijama iznad dijagrama sekvenci, a intracelularni regioni su označeni isprekidanim linijama ispod dijagrama sekvenci; predviđeni

Hyp predstavljeni su vertikalnim crnim linijama, dok su AG regioni označeni kao svetlo sivi pravougaonici.

Filogenetsko stablo za kratke sekvence arabinogalaktanskih proteina, AG peptide (Slika 20), konstruisano je pristupom koji se nije bazirao na poravnanju sekvenci, već na učestalosti tripeptida (3-mernih sekvenci). Ovakav pristup je izabran pošto nije bilo moguće *blastp* programom identifikovati homologe svih odabranih AGp sekvenci, niti konstruisati smisleno višestruko poravnanje. Filogenetsko stablo (Slika 20) ukazuje na izraženu divergenciju između AGp sekvenci i samo je klaster A podržan visokom *bootstrap* vrednošću. Ovo je posledica prisustva domena karakterističnog za podskup arabinogalaktanskih peptida (PF06376.13) u sekvencama koje čine pomenuti klaster. Osim toga, sve sekvence iz klastera A imaju predviđeno mesto vezivanja GPI i kratak AG region između N-sp i omega mesta. Preostale sekvence se sastoje ili od AGp sa kratkim AG regionima i predviđenim mestom za vezivanje GPI-sidra (predstavnik je CeAGp6) i AGp sa nešto dužim AG regionima koji su u nekim slučajevima kombinovani sa ekstenzinskim motivima (predstavnici su CeAGp10 i CeAGp8), za koje nije predviđeno mesto vezivanja GPI sidra. Pošto su CeAGp8 i CeAGp10 parcijalne sekvence kojima nedostaje C-terminus moguće je da, u zavisnosti od dužine nepoznatog regiona, one nisu AG peptidi već da su duži AGP.



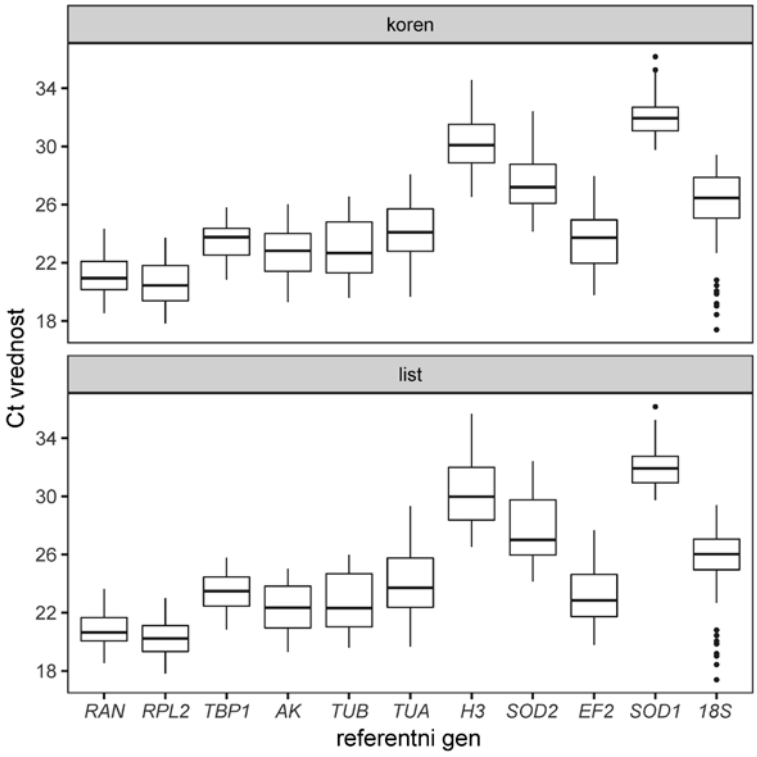
Slika 20. Filogenetski odnosi izabranih AGp proteinskih sekvenci *C. erythraea* sa homolognim sekvencama iz drugih biljnih vrsta. Filogenetsko stablo predstavlja neukorenjeno NJ (engl. „*Neighbour Joining*“) stablo zasnovano na matricama rastojanja bez poravnanja sekvenci (brojanje 3-mera kao što je implementirano u *kmer* R paketu). Stabilnost klastera procenjena je korišćenjem 100 ponavljanja neparametarskog *bootstrap*-a. Klasteri sa ≥ 80/100 *bootstrap*

podrške označeni su crvenim brojem. Klasteri sa ≤ 50/100 *bootstrap* podrške su spojeni u politomije. Šematski dijagram proteinskih sekvenci konstruisan je pomoću *ragp* R paketa kao što je opisano u legendama slika 18 i 19.

4.3. Analiza ekspresije odabranih *AGP* gena kičice

4.3.1. Odabir referentnih gena za ispitivanje ekspresije *AGP* gena

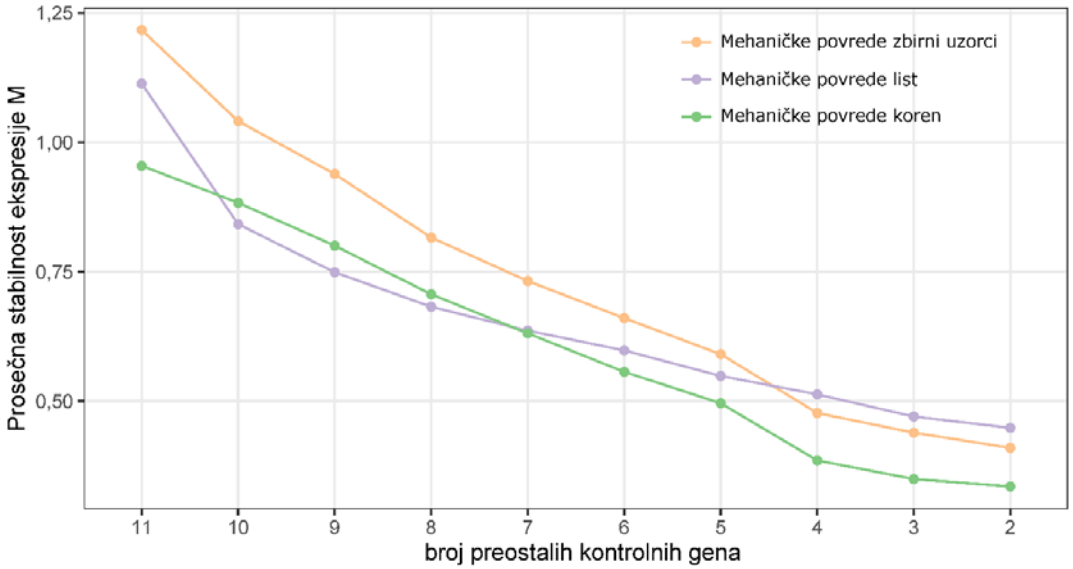
Za robusnu i pre svega tačnu evaluaciju ekspresije gena potrebno je odabrati odgovarajuće referentne gene sa stabilnom ekspresijom u ispitivanim uslovima. U cilju pronalaženja gena sa stabilnom ekspresijom (ekspresijom koja malo varira) koji bi se koristili prilikom analize ekspresije odabranih *AGP* gena evaluirana je stabilnost jedanaest gena (odeljak 3.5.9.) u odgovoru na mehaničke povrede korena i lista biljaka gajenih u uslovima *in vitro*. Na osnovu distribucija ekspresije (Ct vrednosti) (engl. „*cycle threshold*“) ispitivanih gena u uzorcima tkiva nakon povreda listova i korena (Slika 21) može se primetiti da su Ct vrednosti za devet gena u oba skupa uzoraka u preporučenom opsegu 15 - 30, dok *H3* i *Fe-SOD1* imaju nešto više Ct vrednosti, odnosno niži nivo ekspresije od ostalih gena. Četiri gena, *RAN, RPL2, TBP1* i *AK* imaju nisku varijaciju Ct vrednosti, uz uzak interkvartalni opseg bez opservacija koje vidno odskaču od distribucije, što je indikator stabilnosti ekspresije ovih gena (Slika 21).



Slika 21. Distribucija Ct vrednosti jedanaest gena u eksperimentu gde je praćena ekspresija nakon povreda biljnog tkiva. Detalji o ispitivanim genima su dati u odeljku 3.5.9.

Stabilnost ekspresije jedanaest ispitivanih gena procenjena je algoritmima *geNorm* (Vandesompele i sar., 2002) i *NormFinder* (Andersen i sar., 2004). Stabilnost ekspresije M *geNorm* algoritmom je određena za tri tipa uzoraka: uzorke povreda lista, uzorke povrede

korena i zbirno oba organa (Slika 22). Ako posmatramo dva ili tri gena sa najnižom M vrednošću onda skup uzoraka mehaničkih povreda korena ima najveću stabilnost ekspresije (najnižu M vrednost), dok skup uzoraka mehaničkih povreda lista ima najnižu stabilnost (najvišu M vrednost). Za sve ispitivane kombinacije uzoraka, skupovi od dva ili tri gena sa najstabilnijom ekspresijom imaju M vrednost nižu od 0,5 (Slika 22).



Slika 22. Prosečna stabilnost ekspresije gena (M) izračunate *geNorm* algoritmom za jedanaest ispitivanih gena u tri skupa uzoraka: mehaničke povrede lista, mehaničke povrede korena i zbirno, mehaničke povrede korena i lista.

Redosled odabranih gena prema stabilnosti ekspresije algoritmima *geNorm* i *NormFinder* prikazan je u Tabeli 19.

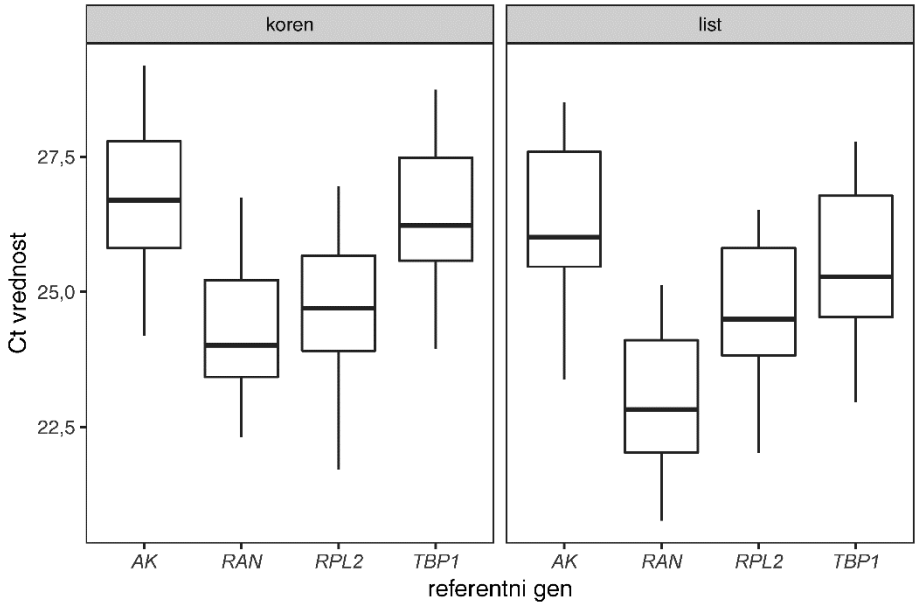
Tabela 19. Rangiranje jedanaest ispitivanih gena prema stabilnosti ekspresije evaluirane pomoću *geNorm* (GN) i *NormFinder* (NF) algoritama u tri skupa uzoraka. Podvučeni su geni koji su odabrani za normalizaciju.

skup / rang	povrede list		povrede koren		zbirno uzorci povrede	
	GN	*NF*	*GN*	*NF*	*GN*	*NF*
11	*EF2*	*EF2*	*SOD2*	*18S*	*EF2*	*SOD1*
10	*18S*	*H3*	*H3*	*SOD2*	*18S*	*EF2*
9	*H3*	*RAN*	*TUA*	*TUB*	*H3*	*H3*
8	*SOD1*	*18S*	*18S*	*EF2*	*SOD2*	*18S*
7	*TUB*	*SOD2*	*EF2*	*H3*	*TUA*	*SOD2*
6	*SOD2*	*TUA*	*TUB*	*TUA*	*SOD1*	*AK*
5	*TUA*	*AK*	*SOD1*	*TBP1*	*TUB*	*TUB*
4	*TBP1*	*TUB*	*TBP1*	*RPL2*	*TBP1*	*RAN*
3	*RPL2*	*SOD1*	*RPL2*	*SOD1*	*RPL2*	*TUA*
2		*RPL2*		*AK*		*RPL2*
1	*AK/RAN*	*TBP1*	*AK/RAN*	*RAN*	*AK/RAN*	*TBP1*

Na osnovu procene stabilnosti ekspresije gena kandidata za referentne gene u eksperimentima sa povredama lista odabrani su geni *TBP1* i *RPL2* koji su se pokazali

najstabilniji prema *NormFinder* algoritmu, a bili su jako stabilni (među najbolje rangirana četiri gena) i prema *geNorm* algoritmu. Kao referentni geni za skup uzoraka povreda korena odabrani su geni *AK* i *RAN* koji su imali najveću stabilnost ekspresije prema oba testirana algoritma (Tabela 19).

Za određivanje adekvatnih referentnih gena u eksperimentu sa Yariv reagensom, evaluirana su četiri gena *TBP1*, *AK*, *RAN* i *RPL2* koja su se pokazali kao stabilni prema oba testirana algoritma u uzorcima povreda lista i korena (Tabela 19) kao i u svim grupama uzoraka testiranih u radu Ćukovic i sar. (2020). Distribucija Ct vrednosti ova četiri gena ukazuje da imaju malu varijaciju ekspresije u dva skupa uzoraka tretirana Yariv reagensom (Slika 23). Kao referentni geni u eksperimentu gde se ispituje uticaj različitih koncentracija Yariv reagensa na eksplantate lista u uslovima *in vitro* odabrani su *TBP1* i *RPL2*, a za korenove su odabrani *AK* i *RAN* (Tabela 20).



Slika 23. Distribucija Ct vrednosti četiri referentna gena *TBP1*, *AK*, *RAN* i *RPL2* u eksperimentu gde su biljke gajene na različitim koncentracijam βGlcY reagensa.

Tabela 20. Rangiranje četiri ispitivana gena prema stabilnosti ekspresije evaluirane pomoću *geNorm* (GN) i *NormFinder* (NF) algoritama u dva skupa uzoraka. Podvučeni su geni koji su odabrani za normalizaciju.

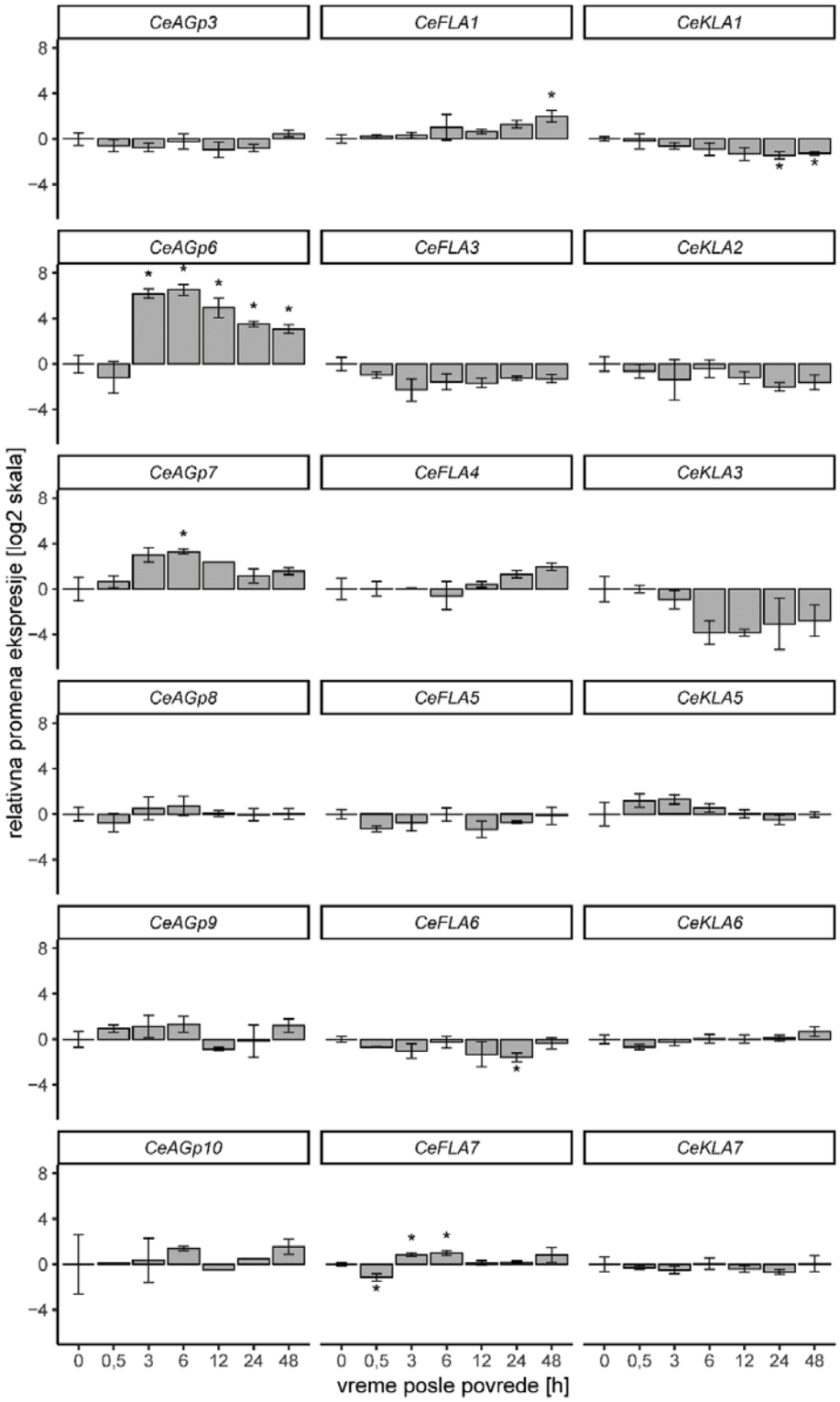
skup / **rang**	**Yariv list**		**Yariv koren**	
	GN	*NF*	*GN*	*NF*
4	*RAN*	*AK*	*RPL2*	*TBP1*
3	*RPL2*	*TBP1*	*RAN*	*RAN*
2		*RAN*		*RPL2*
1	*AK/TBP1*	*RPL2*	*AK/TBP1*	*AK*

4.3.2. Ekspresija odabranih *AGP* gena nakon mehaničke povrede lista i korena kičice gajene u uslovima *in vitro*

U cilju ispitivanja da li je neki od analiziranih gena uključen u odgovor na stres izazvan povredama, ispitana je ekspresija odabranih *AGP* gena u periodu do 48 h nakon sečenja eksplantata listova i korenova. Predstavljena je relativna genska ekspresija (na log skali), a za normalizaciju je korišćena aritmetička sredina dva gena (na log skali) sa najstabilnijom ekspresijom (odeljak 4.3.1.) u svakom skupu uzoraka (Slike 24, 25).

U uzorcima lista šest proučavanih gena, *CeAGp6*, *CeAGp7*, *CeFLA1*, *CeFLA6*, *CeFLA7* i *CeKLA1* su imali statistički značajnu promenu ekspresije u nekim vremenskim tačkama u poređenju sa kontrolnim tkivima koja su zamrznuta odmah nakon povrede (0 min, Slika 24). Većina ispitivanih gena je imala malu ili umerenu promenu ekspresije u odnosu na kontrolni uzorak i samo su se *CeAGp6* i *CeAGp7* isticali na osnovu profila ekspresije (Slika 24). Oba gena su indukovana povredama, sa maksimumom ekspresije 6 h nakon povrede: *CeAGp6* je imao skoro 100 puta povećanu ekspresiju (relativna promena od log_2 6,5 puta), dok je *CeAGp7* je imao oko 10 puta povećanu ekspresiju (relativna promena od log_2 3,3 puta). Oba gena su imala sličan profil ekspresije sa umerenim promenom 30 min nakon povrede, uz naglo povećanje ekspresije 3 h nakon povrede i vrhuncem 6 h nakon povrede, nakon čega je ekspresija polako opadala (Slika 24).

Od ostalih ispitivanih gena, profil ekspresije *CeKLA3* je zanimljiv, iako promene u njegovoj ekspresiji nisu značajne ni u jednoj vremenskoj tački. Nasuprot *CeAGp6* i *CeAGp7*, ekspresija *CeKLA3* je opadala nakon povrede, dostižući minimum sa oko 14 puta manjom ekspresijom u odnosu na kontrolno tkivo 6 h nakon povrede (relativna promena od log_2 -3,8 puta). Ipak, zbog velike varijanse u ekspresiji bioloških ponavljanja i korekciji za broj statističkih poređenja, ove promena nije imala statistički značaj. Prema trendu ekspresije, značajni su, takođe, *CeFLA1* i *CeKLA1* koji pokazuju stabilnu, ali skromnu promenu ekspresije nakon povreda, sa suprotnim trendom i vrhuncem ekspresije pri samom kraju posmatranog perioda. Ekspresija *CeFLA1* se monotono povećavala dostižući vrhunac 48 h nakon povrede (relativna promena od log_2 2 puta), dok je ekspresija *CeKLA1* monotono opadala dostižući minimum 24 h nakon povrede (relativna promena od log_2 -1,5 puta).

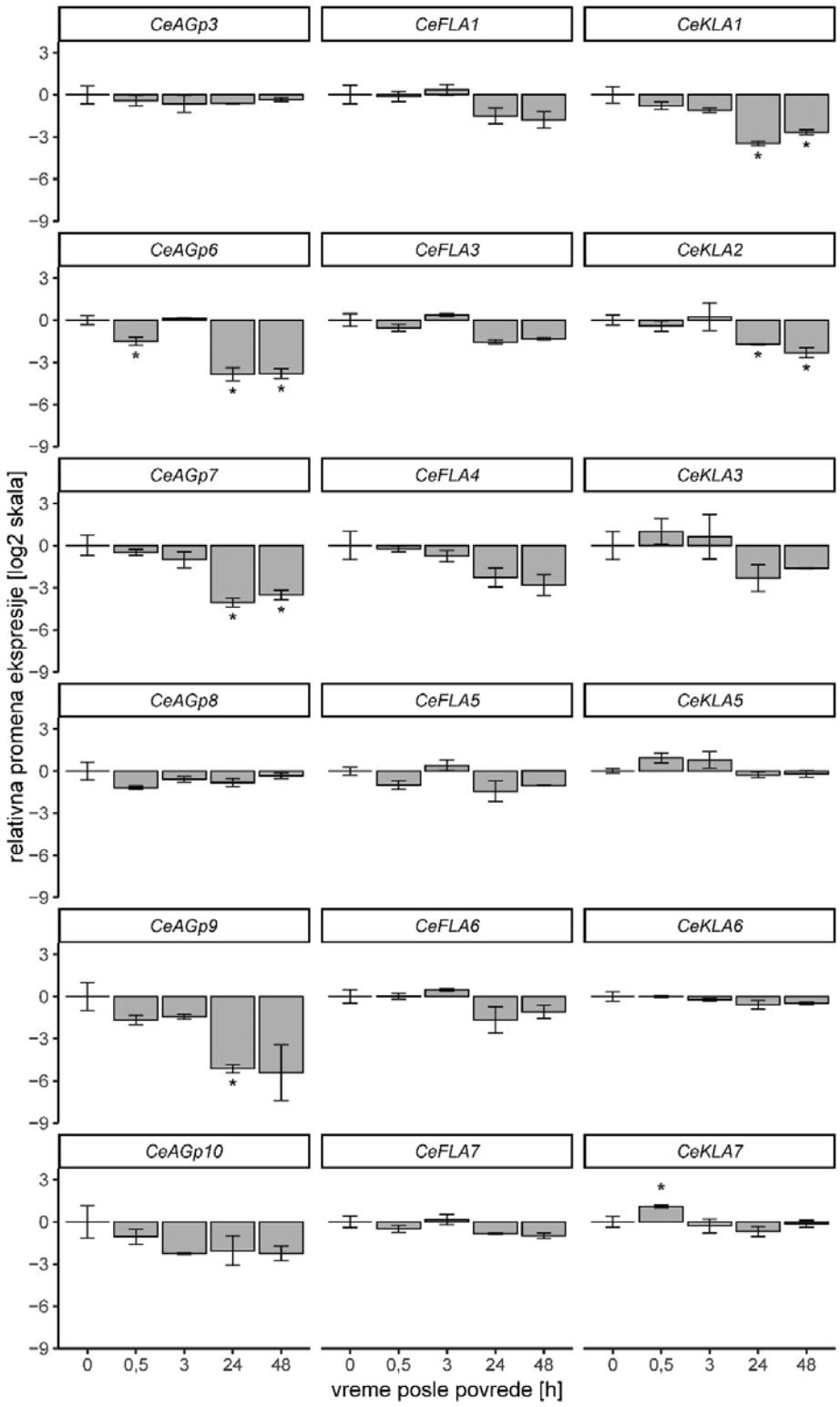


Slika 24. Profili ekspresije za 18 odabranih *AGp*, *FLA* i *KLA* gena nakon povređivanja, odnosno isecanja eksplantata listova. Srednja vrednost i standardna devijacija relativne genske ekspresije (log_2) su prikazane za tri biološka ponavljanja. Kao kontrolni uzorak korišćeno je tkivo zamrznuto odmah nakon povrede (0 min). Zvezdice označavaju statistički značajne p-vrednosti ($p < 0,05$) korigovane zbog višestrukih poređenja.

Za praćenje profila ekspresije odabranih gena u eksplantatima korenova nakon povrede postavljene su ekvivalentne vremenske tačke kao u eksperimentu sa listovima. Uzorci

korenova su se pokazali dosta izazovniji za izolaciju kvalitetne RNK čak i nakon optimizacije protokola. Pošto su imali jako niske koncentracije RNK uzorci korenova 6 h i 12 h nakon povrede nisu uključeni u analizu ekspresije, odnosno, ekspresija je izmerena i u ovim uzorcima, međutim, detektovana je samo za mali broj gena, u ponekim biološkim ponavljanjima. Dakle za uzorke korenova praćena je promena ekspresije u brzom odgovoru nakon povrede (do 3 h) i nešto sporijem odgovoru (nakon 24 h i 48 h) (Slika 25).

Šest ispitivanih gena, *CeAGp6, CeAGp7, CeAGp9, CeKLA1, CeKLA2* i *CeKLA7* su imali statistički značajnu promenu ekspresije nakon povrede korena u poređenju sa kontrolnim tkivom (Slika 25). Nijedan od FLA gena nije pokazao statistički značajnu promenu ekspresije. Kod većine gena, najveće promene zabeležene su 24 h i 48 h nakon povrede korena, i uglavnom je zabeležena smanjena ekspresija gena u ovim vremenskim tačkama. Geni *CeAGp6, CeAGp7, CeAGp9, CeKLA1* i *CeKLA2* su imali slične profile ekspresije sa relativno malim smanjenjem tokom brzog odgovora na povrede (do 3 h) i drastičnijim smanjenjem ekspresije 24 h i 48 h nakon povrede korena. Jedino je *CeKLA7* pokazao statistički značajno povećanje ekspresije i to pola sata nakon povrede, međutim ovaj rezultat treba uzeti sa rezervom pošto je izmereno povećanje bilo malog intenziteta i detektovano samo u jednoj vremenskoj tački.

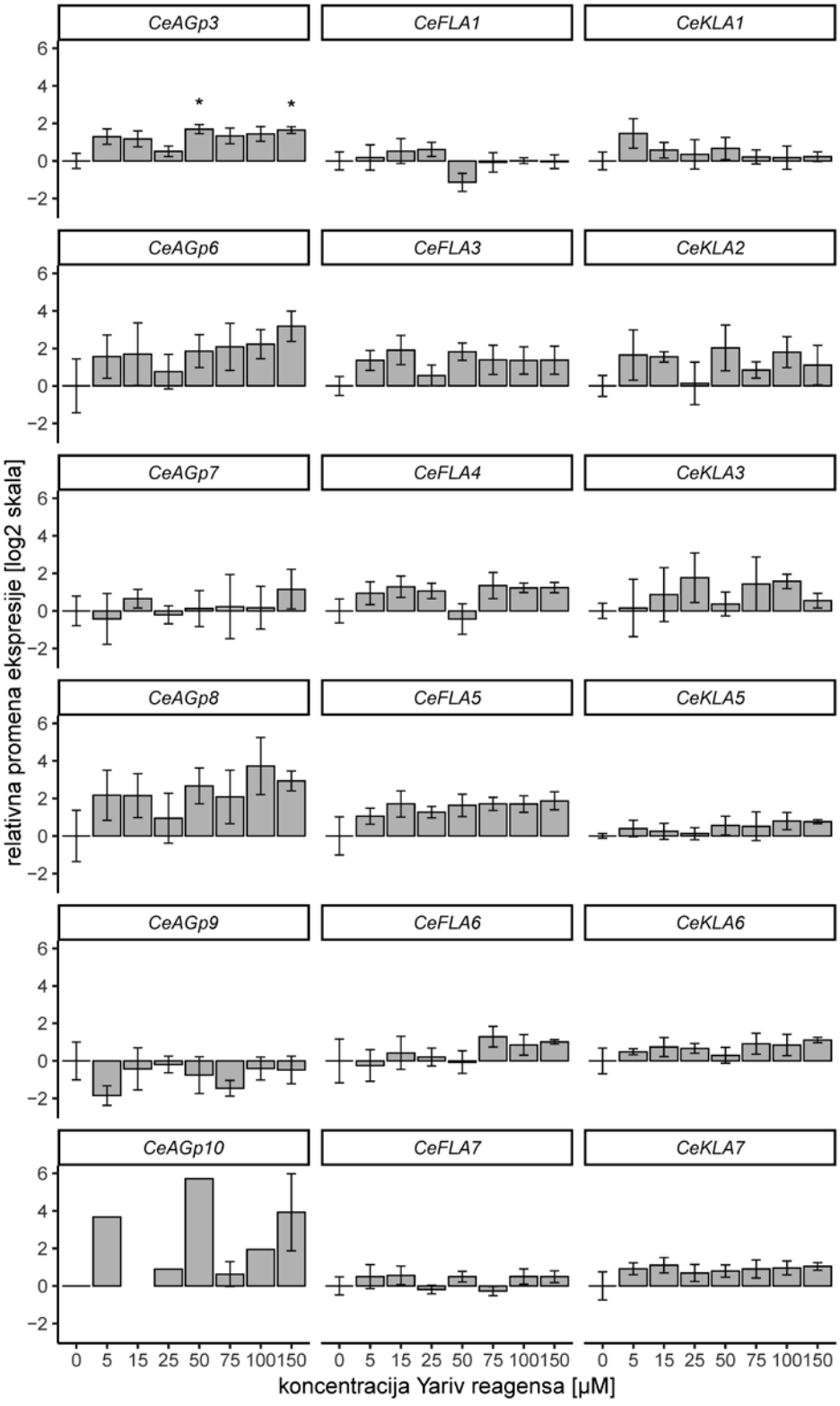


Slika 25. Profili ekspresije 18 odabranih *AGp*, *FLA* i *KLA* gena nakon nakon povređivanja, odnosno isecanja eksplantata korenova. Srednja vrednost i standardna devijacija relativne genske ekspresije (log_2) su prikazane za tri biološka ponavljanja. Srednja vrednost i standardna devijacija relativne genske ekspresije (log_2) su prikazane za tri biološka ponavljanja. Kao kontrolni uzorak korišćeno je tkivo zamrznuto odmah nakon povrede (0 min). Zvezdice označavaju statistički značajne p-vrednosti ($p < 0,05$) korigovane zbog višestrukih poređenja.

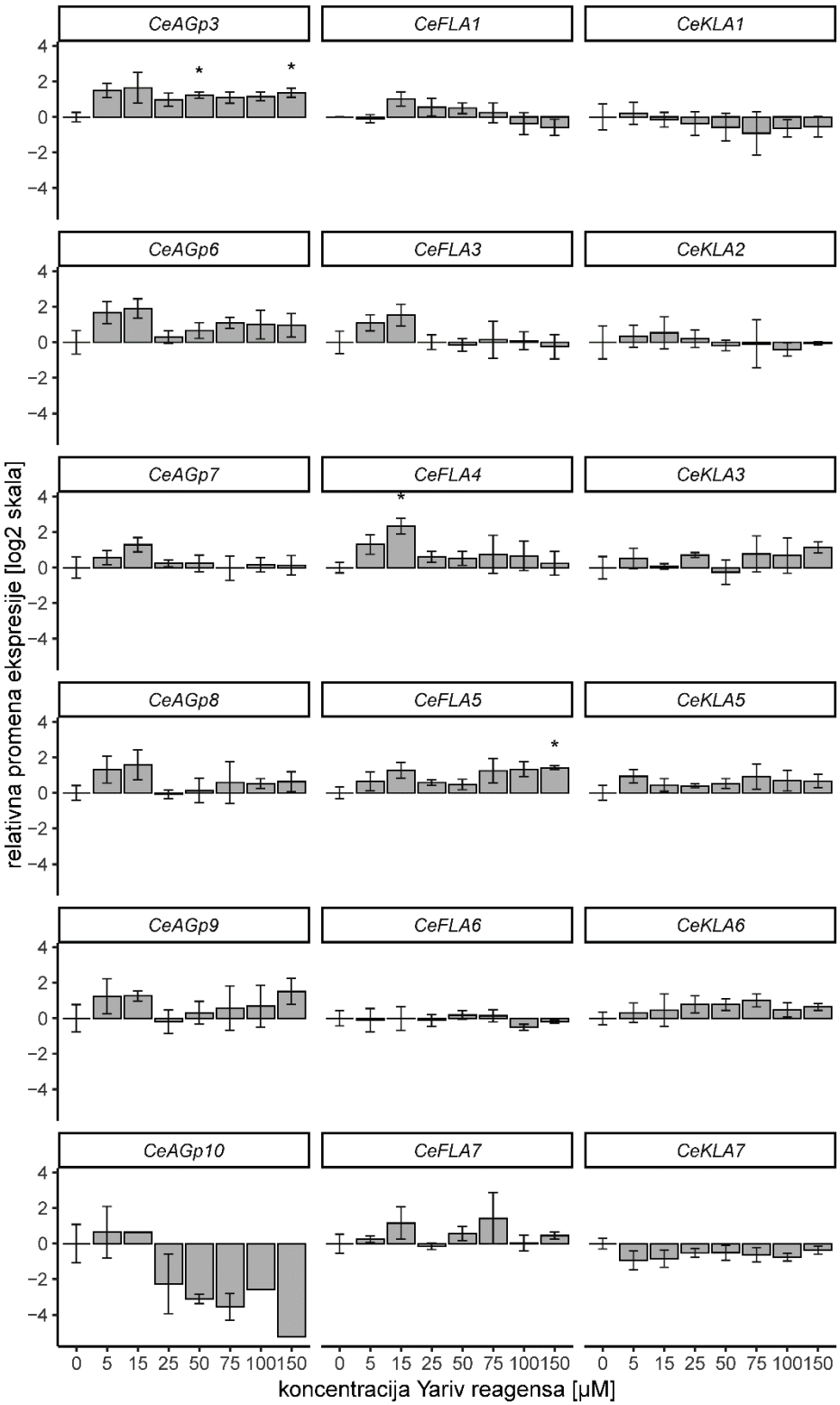
4.3.3. Ekspresija odabranih *AGP* gena nakon dugotrajnog izlaganja eksplantata lista i korena kičice gajenih u uslovima *in vitro* različitim koncentracijama βGlcY

Često korišćeni pristup proučavanju uloge AGP tokom različitih fizioloških procesa biljaka je upotreba βGlcY, koji specifično interaguje sa AGP i koristi se za njihovu lokalizaciju, izolaciju i kvantifikaciju. Ekspresija osamnaest odabranih *AGP* gena analizirana je u odgovoru na rastuće koncentracije βGlcY u hranljivim podlogama na kojima su gajeni eksplantati listova i korenova u uslovima *in vitro* tokom četiri nedelje. Analizirana je relativna promena ekspresije gena u odnosu na tkivo koje je raslo na podlogama bez dodatog βGlcY. Za normalizaciju ekspresije je korišćena aritmetička sredina Ct vrednosti dva gena koji su se pokazali najstabilniji u odgovarajućim skupovima uzoraka (odeljak 4.3.1.) i Tabela 20.

Zabeležene su male promene ekspresije svih ispitivanih gena u uzorcima eksplantata listova i korenova u odgovoru na rastuću koncentraciju βGlcY u hranljivoj podlozi (Slike 26 i 27). U eksplantatima listova samo je *CeAGp3* imao statistički značajnu promenu ekspresije pri koncentracijama od 50 i 150 μM βGlcY u podlogama u poređenju sa kontrolnim uzorkom (Slika 26). U uzorcima korena su, pored *CeAGp3*, statistički značajno povećanje ekspresije imali i *CeFLA4* i *CeFLA*5 pri pojedinim koncentracijama βGlcY (Slika 27). Treba napomenuti da je kod nekoliko gena (*CeAGp*3, *CeAGp6, CeAGp8, CeFLA3, CeFLA4* i *CeAFLA*5) primećeno povećanje ekspresije u eksplantatima listova pri odgovoru na povećane koncentracije βGlcY u hranljivoj podlozi. Te promene su u većini slučajeva bile diskretne (malog intenziteta), ili su imale relativno veliku varijansu među biološkim ponavljanjima, pa zahvaljujući relativno strogoj statističkoj obradi nisu bile značajno različite u odnosu na odgovarajući kontrolni uzorak (Slika 26). Velike promene u ekspresiji primećene su kod *CeAGp10* u eksplantatima listova i korenova. Relativna ekspresija ovog gena je u eksplantatima listova rasla sa povećanjem koncentracije βGlcY u podlogama, dok je u eksplantatima korenova opadala. Ove promene treba posmatrati pre svega kroz izrazitu nisku ekspresiju pomenutog gena u svim uzorcima uključujući i kontrole sa velikom varijansom. U mnogim biološkim ponavljanjima ekspresija *CeAGp10* nije ni detektovana, pa za slučajeve kada u dva od tri biološka ponavljanja ekspresija nije detektovana statistička analiza nije rađena. Upravo zbog ovih razloga relativno velike promene u ekspresiji *CeAGp10* nisu bile statistički značajne.



Slika 26. Relativna ekspresija proučavanih *AGp*, *FLA* i *KLA* gena u uzorcima listova kičice gajene u uslovima *in vitro* na hranljivim podlogama sa različitim koncentracijama βGlcY reagensa tokom četiri nedelje. Srednja vrednost i standardna devijacija relativne genske ekspresije (log_2) su prikazane za tri biološka ponavljanja. Kao kontrolni uzorak korišćeno je tkivo gajeno na hranljivoj podlozi bez βGlcY (0 μM βGlcY). Zvezdice označavaju statistički značajne p-vrednosti ($p < 0,05$) korigovane zbog višestrukih poređenja.

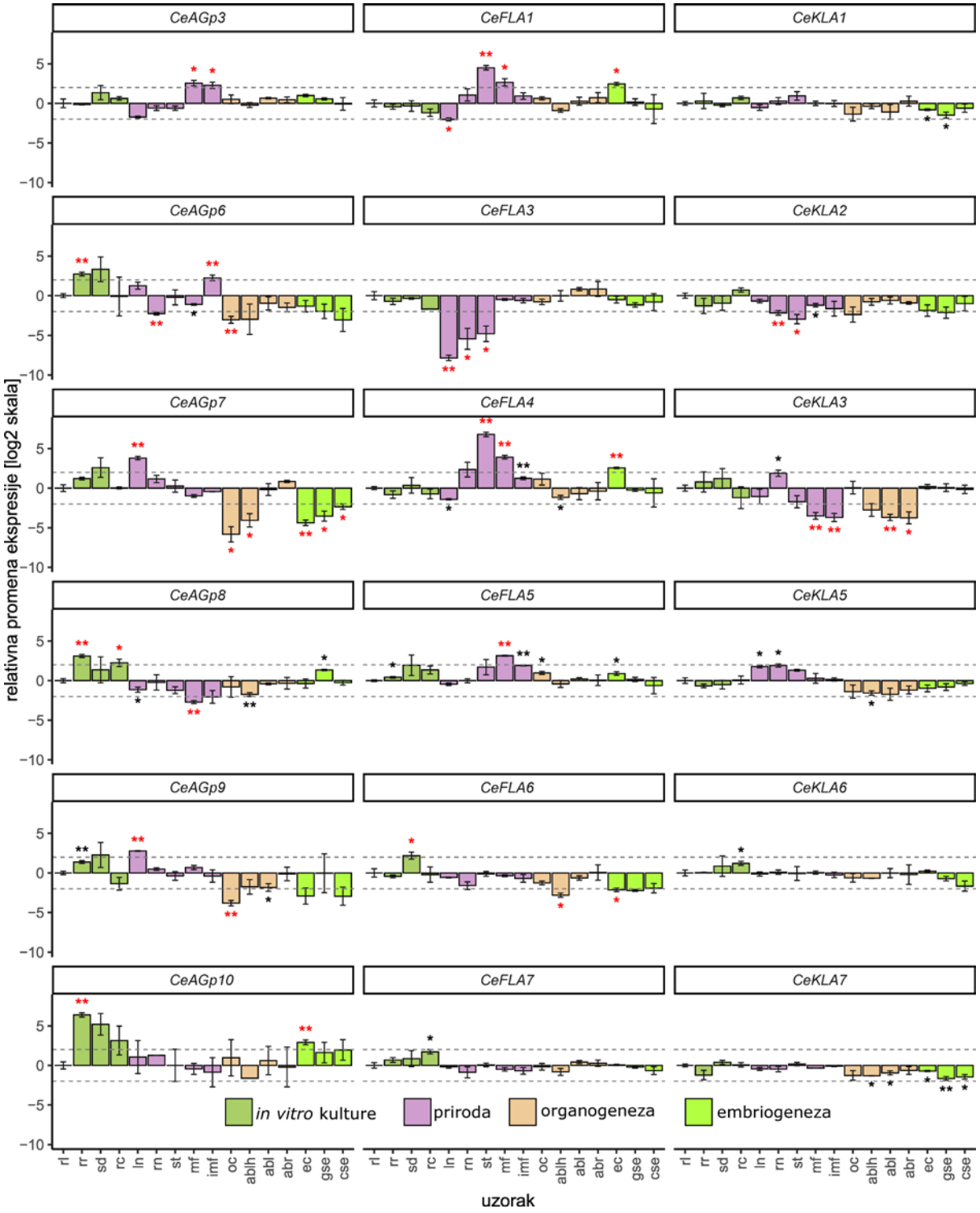


Slika 27. Relativna ekspresija proučavanih *AGp*, *FLA* i *KLA* gena u uzorcima korena kičice gajene u uslovima *in vitro* na hranljivim podlogama sa različitim koncentracijama βGlcY tokom četiri nedelje. Srednja vrednost i standardna devijacija relativne genske ekspresije (log_2) su prikazane za tri biološka ponavljanja. Kao kontrolni uzorak korišćeno je tkivo gajeno na u hranljivoj podlozi bez βGlcY (0 μM βGlcY). Zvezdice označavaju statistički značajne p-vrednosti ($p < 0,05$) korigovane zbog višestrukih poređenja.

4.3.4. Ekspresija odabranih *AGP* gena u različitim tkivima biljaka gajenih *in vitro* i delovima biljaka iz prirode

Značaj AGP tokom SE kod kičice je do sada potvrđen u više istraživanja (Simonović i sar., 2015; Filipović i sar., 2021). Upravo zbog toga je proverena ekspresija odabranih *AGP* gena u uzorcima tkiva organa *in vitro* gajenih biljaka, tkiva iz različitih faza organogeneze i SE kao i tkiva iz prirode. Uzorci pripadaju eksperimentalnom sistemu koji je detaljno objašnjen u Ćukovic i sar. (2020). Detalji o uzorcima i oznake su dati u Tabeli 8.

Relativna ekspresija svih proučavanih gena značajno je bila drugačija, barem u nekim uzorcima, u poređenju sa listom rozete (rl) koji je korišćen kao kontrolni uzorak (Slika 28). Treba napomenuti da je na dobijene statističke značajnosti u mnogim poređenjima uticala pre svega niska varijansa ekspresije, a ne velike promene u intenzitetu ekspresije (*CeKLA, CeFLA6, CeFLA7* i *CeAGp3*, Slika 28). Zbog toga je razmatranje značajnih razlika u ekspresiji gena ograničeno na ona poređenja koja su imala relativnu promenu srednje ekspresije od barem log_2 *2* puta u odnosu na rl. Od 18 analiziranih gena, 13 je zadovoljavalo pomenuti uslov. Tih 13 gena obuhvata sve analizirane *AGp* gene, sve *FLA* gene sem *CeFLA7* i dva *KLA* gena, *CeKLA2* i *CeKLA3*. Zanimljivo je da su *CeAGp6* i *CeAGp7*, koji su bili indukovani nakon povreda listova, a reprimirani nakon povreda korenova imali smanjenu ekspresiju u većini embriogenih i organogenih uzoraka (Slika 28). Ovo smanjenje je bilo izraženije u slučaju *CeAGp7*, gde je u pet od sedam embriogenih i organogenih uzoraka detektovana značajno smanjena ekspresija: organogeni kalus (oc), adventivni pupoljci poreklom od lista gajeni na hranljivoj podlozi sa 2,4-D/CPPU (ablh), embriogeni kalus (ec), globularni i kotiledonarni somatski embrioni (gse i cse) gajeni na istoj hranljivoj podlozi (videti Tabelu 8). *CeAGp7* je imao značajno povećanu ekspresiju u listovima sakupljenim od biljaka gajenih u prirodi (ln). *CeAGp6* je imao značajno smanjenu ekspresiju u organogenom kalusu (oc) i korenovima iz biljaka koje su cvetale u prirodi (rn), a povišenu u korenu rozete (rr) i nezrelim cvetovima sa biljaka iz prirode (imf). Ekspresija *CeAGp9* značajno je smanjena u oc, a povećana u listovima biljaka iz prirode (ln), dok je *CeAGp10* visoko eksprimiran u rr. Pored ovih *AGp* gena, najveću promenu u ekspresiji u poređenju sa kontrolom je imao *CeFLA3*, koji je slabo eksprimiran u vegetativnim organima iz prirode – listovima (ln), korenu (rn) i stablu (st). Nasuprot *CeFLA3, CeFLA1, CeFAL4* i u manjem obimu *CeFLA5* geni su generalno više eksprimirani u biljkama iz prirode nego u uzorcima biljaka gajenim u uslovima *in vitro* kulture. *CeKLA3* je imao najnižu ekspresiju u adventivnim pupoljcima, bez obzira na poreklo (ablh i abl), kao i u zrelim i nezrelim cvetovima (mf i imf) sa biljaka gajenih u prirodi. Nekoliko gena, kao što su *CeFLA7, CeKLA1, CeKLA6* i *CeKLA7* su imali konstitutivnu ekspresiju (manje relativne promene od log_2 2 puta poređenju sa rl) u svim testiranim uzorcima (Slika 28).

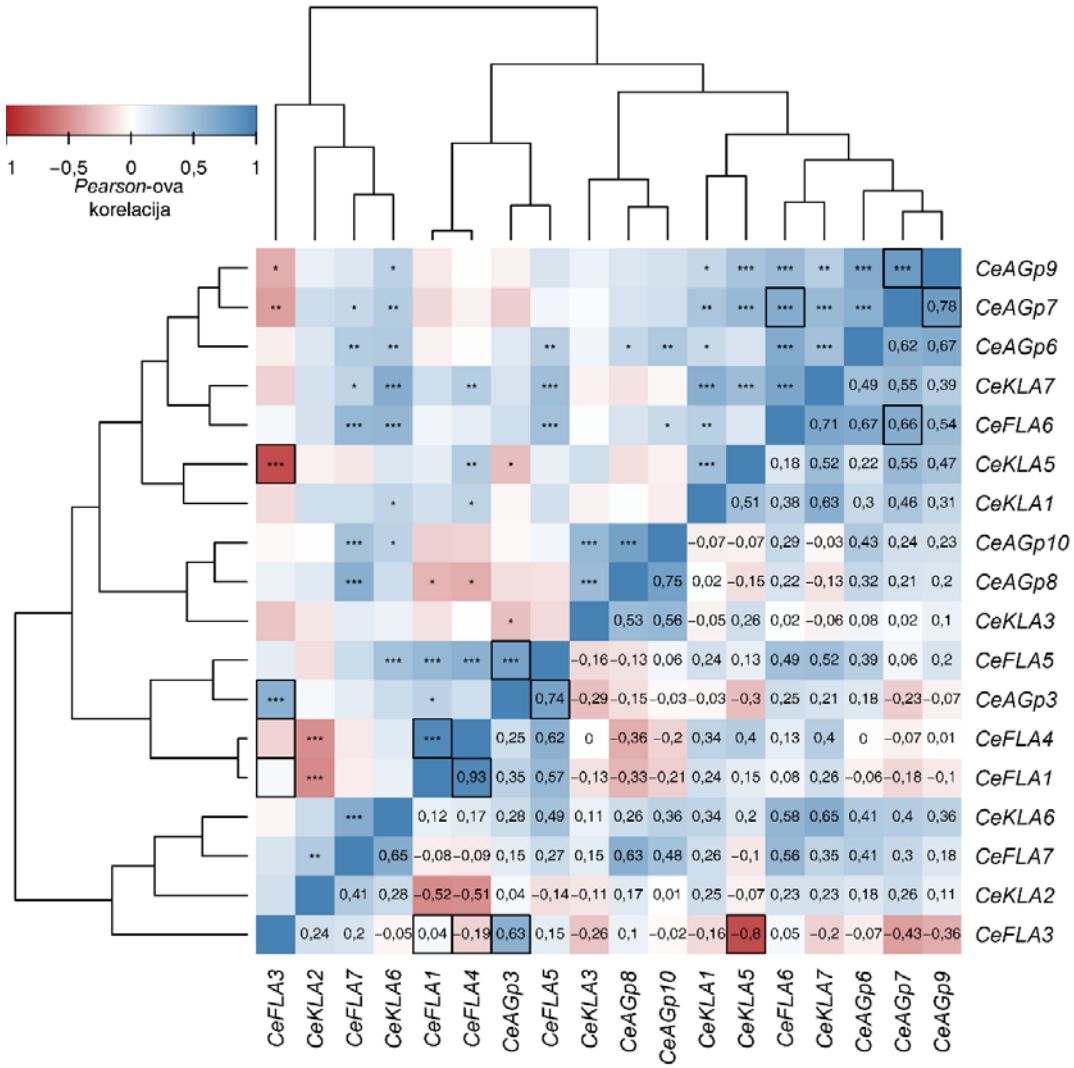


Slika 28. Relativna ekspresija osamnaest proučavanih *AGp*, *FLA* i *KLA* gena u 16 različitih uzoraka tkiva i organa biljaka. Ekspresija je normalizovana na listove rozete gajene u uslovima *in vitro* (rl). Prikazane su srednja vrednost i standardna devijacija za tri biološka ponavljanja. Horizontalne isprekidane linija predstavljaju relativnu promenu ekspresije od *log2* 2 i -2 puta u poređenju sa listovima rozete (rl) gajenim u uslovima *in vitro*. *Welch*-ov t-test je korišćen za poređenje relativne ekspresije gena u svakom od uzoraka sa ekspresijom u listu rozete (rl). Jedna zvezdica (*) ukazuje na značajne razlike u ekspresiji gena u datom uzorku u poređenju sa rl (korigovan p < 0,05); dve zvezdice (**) ukazuju na visoko značajne razlike u ekspresiji gena u datom uzorku u poređenju sa rl (korigovan p < 0,01); Crvena boja zvezdica označava

da je poređenje srednjih vrednosti relativne ekspresije značajno i da je razlika u srednjim vrednostima veća od log_2 2 puta.

Pearson-ovom korelacijom (cor_{Pear}) (Pearson, 1895) su procenjeni linearni odnosi između relativnih ekspresija ispitivanih gena u 16 uzoraka biljnih tkiva (Slika 29). Skoro linearna zavisnost postoji između *CeFLA1* i *CeFLA4* ekspresija (0,93 cor_{Pear}), dok je negativna korelacija izračunata između ekspresija *CeKLA5* i *CeFLA3* (-0,8 cor_{Pear}). Visoka pozitivna korelacija postoji među genima: *CeAGp9*, *CeAGp7*, *CeAGp6*, *CeKLA7* i *CeFLA6* (Slika 29), od kojih svi imaju donekle smanjenu ekspresiju u uzorcima koji se odnose na SE i SO i povećanu ekspresiju u klijancima (Slika 28).

Dalja analiza uzajamnih veza između profila ekspresije parova gena urađena je korišćenjem BCMI (engl. „*bias corrected mutual information*“) koja kvantifikuje količinu informacije koju obrazac ekspresije svakog gena sadrži o ostalim. Najviše BCMI vrednosti su imali parovi sa visokom apsolutnom *Pearson*-ovom korelacijom: *CeAGp7* i *CeAGp9*, *CeAGp7* i *CeFLA6*, *CeAGp3* i *CeFLA5*, *CeFLA3* i *CeKLA5* i *CeFLA3* i *CeAGp3*. Dok su parovi *CeFLA3* sa *CeFLA1* i *CeFLA4* imali visok BCMI, a nisku *Pearson*-ovu korelaciju, što ukazuje na postojanje nelinearne veze koja postoji među obrascima ekspresije ovih gena.



Slika 29. *Pearson*-ova korelaciona toplotna mapa relativne ekspresije gena. Korelacioni koeficijenti su dati u trouglu ispod dijagonale, dok su statističke značajnosti poređenja parova date u trouglu iznad dijagonale: * za $p < 0,05$, ** za $p < 0,01$ i *** za $p < 0,001$. Redovi i

kolone toplotne mape su raspoređeni prema hijerarhijskoj analizi klastera urađenoj na matrici korelacionih distanci ($1-cor_{Pear}$). Aglomeracija klastera je urađena metodom potpunog povezivanja (engl. *„complete linkage“*) i dobijeni dendrogrami su prikazani levo i iznad toplotne mape. Naglašene ćelije (uokvirene pravougaonicima) označavaju najvećih 5% parova na osnovu korigovanog uzajamnog sadržaja informacije (BCMI, engl. „*jackknife bias corrected mutual information*“).

5. Diskusija

5.1. R paket *ragp*

Sa ciljem identifikacije HRGP sekvenci koja bi istovremeno povećala specifičnost identifikovanih sekvenci (da se smanji udeo lažno identifikovanih HRGP sekvenci) uz istovremeno zadržavanje senzitivnosti identifikacije (da se identifikuju skoro sve sekvence koje jesu HRGP) osmišljen je novi pristup identifikaciji ovih sekvenci koji inkorporira mašinsko učenje. Pomenuti pristup bazira se na jednoj od osnovnih odlika HRGP, a to je prisustvo hidroksiprolina. Identifikacija verovatnih hidroksiprolina u sekvencama biljnih proteina se postiže korišćenjem modela MU koji je treniran da predviđa verovatnoću hidroksilacije prolina na osnovu lokalne sekvence. Ovaj pristup je implementiran u *ragp* R paket koji je dostupan pod MIT (engl. „*Massachusetts Institute of Technology*“) licencom na *github* platformi na adresi https://github.com/missuse/ragp.

5.1.1. Predviđanje pozicija hidroksiprolina u proteinskim sekvencama biljaka

Dosadašnji pristupi identifikaciji HRGP sekvenci (Showalter i sar., 2010; Johnson i sar., 2017a; Ma i sar., 2017) su takođe koristili različite modele MU koji su upotrebljavani za predviđanje N-sp i GPI. Glavna inovacija u odnosu na pomenute pristupe, koja je razvijena tokom izrade ove teze, predstavlja model MU koji predviđa verovatnoću hidroksilacije prolina u proteinskim sekvencama biljaka. Tokom treniranja i selekcije modela MU postavljeno je nekoliko ciljeva koje bi on trebalo da zadovolji: da ima sposobnost generalizacije, da evaluacija modela bude objektivna i da je model u stanju da generiše predviđanja brzo, kako bi bile omogućene analize visoke propusnosti celih biljnih proteoma za kratko vreme.

Biljne proteinske sekvence za treniranje modela su preuzete iz UniProtKB/Swiss-Prot (The UniProt Consortium, 2019) baze podataka i rigorozno prekontrolisane kako bi se koristili samo oni prolini/hidroksiprolini u sekvencama za koje se na osnovu dostupne literature može sa sigurnošću utvrditi hidroksilacioni status prolina. Ukupan broj sekvenci sa eksperimentalno anotiranim Hyp u nekim od pozicija je u preuzetim podacima bio 40. U kontekstu modernog MU ovo predstavlja izuzetno skroman skup podataka. Upravo zbog toga je sa posebnom pažnjom osmišljen tok treniranja i evaluacije modela kako bi se anticipirale i izbegle, ili barem ublažile, uobičajene zamke MU (Del Vento i Fanfarillo, 2019). U cilju što objektivnije procene performansi modela, odnosno kako bi se izbegla preoptimistička procena performansi modela u obzir je uzeto sledeće:

- Kada podaci koji se koriste u MU potiču od bioloških sekvenci, često se pristupa uklanjanju homolognih sekvenci da bi se smanjila preoptimistička procena performansi modela tokom validacije ili unakrsne validacije usled prisustva jako sličnih/istih sekvenci u skupovima podataka za treniranje i evaluaciju. Sa druge strane, uklanjanje homolognih sekvenci neizbežno vodi ka gubitku informacije naročito kada se raspolaže sa malim skupom podataka. U ovom slučaju to je naročito izraženo usled preklapanja lokalnih sekvenci iz istog proteina kada su Hyp blizu u sekvenci. Da bi se pomenuti problemi ublažili, korišćena su dva komplementarna pristupa:
    - Smanjenje redundantnosti lokalnih sekvenci sa Hyp u sredini urađeno je na osnovu Levenštajnovih distanci, tako da ne postoje dve lokalne sekvence koje dele više od

90% homologije (u setu sa 21-merama, za kraće k-mere je procenat nešto manji). Dalje uklanjanje homolognih sekvenci (< 90% homologije) značajno je smanjilo skup podataka za treniranje i povećalo nebalansiranost klasa pa nije preduzeto.

- Pošto je većina međusobno homolognih lokalnih sekvenci poticala iz istog proteina, da bi se izbegla subjektivnost prilikom procene performansi modela korišćena je blok unakrsna validacija (naziva se još i grupna unakrsna validacija) u svim fazama treniranja i evaluacije modela. Ovo podrazumeva da se sve lokalne sekvence poreklom od jednog proteina koriste ili za pravljenje modela ili za njegovu evaluaciju. Drugim rečima podela unutar unakrsne validacije se ne generiše na osnovu opservacija već na osnovu grupa opservacija koje potiču od jednog proteina. Ovakav pristup daje objektivniju procenu performansi u poređenju sa korišćenjem uobičajene (nasumične) unakrsne validacije kada opservacije nisu nezavisne (Roberts i sar., 2017). Takođe je vođeno računa da odnos klasa u svakoj od podela unakrsne validacije bude sličan (stratifikacija, Kohavi, 1995).

- Da bi se dobile objektivne procene performansi modela, svi koraci u treniranju modela moraju biti odvojeni od njegove evaluacije (Varma i Simon, 2006; Cawley i Talbot, 2010). Da bi se optimizacija hiperparametara odvojila od evaluacije modela korišćena je unakrsna validacija u dva sloja. Unutrašnji sloj korišćen je za odabir hiperparametara, a spoljašnji sloj je korišćen za procenu performansi modela. Pošto je spoljašnji sloj unakrsne validacije, u slučaju sfs, korišćen i za odabir grupa atributa može se očekivati određena pristrasnost evaluacije u slučaju kada je ova metoda korišćena za odabir atributa. Potpuno nepristrasan pristup bi zahtevao unakrsnu validaciju u tri sloja: unutrašnji sloj za podešavanje hiperparametra, srednji sloj za odabir atributa omotač tipom algoritama i spoljašnji sloj za evaluaciju modela. Zbog dužine trajanja ovakvog procesa, usled treniranja jako velikog broja modela, on nije korišćen u ovom slučaju. Pošto je sfs selekcija bila bazirana na 16 grupa atributa, a ne na pojedinačnim atributima, nije očekivan veliki uticaj na objektivnost ocene performansi.

Kada se razmatra pravljenje modela MU baziranog na proteinskim sekvencama, potrebno je odlučiti na koji način će se numerički kodirati informacija o sekvencama. Proteinske sekvence se mogu numerički kodirati na puno načina koji su opisani u literaturi. Tokom 90-ih i 2000-ih dominantni načini numeričkog kodiranja proteinskih sekvenci ostvarivani su kroz upotrebu različitih domenskih znanja o aminokiselinama i proteinima, koji su zahtevali dobro poznavanje konkretnih problema čije je rešavanje pokušano. Većina ovih metoda se bazira na pojedinim fizičko-hemijskim i biološkim odlikama samih aminokiselina i na kvantifikaciji promene ovih osobina kroz proteinsku sekvencu na određen način (Atchley i sar. 2005; Chou 2000; Chou 2001; Chou 2005; Dubchak i sar., 1995). Deo razvijenih metoda koristio je i učestalost pojedinačnih aminokiselina, dipeptida i tripeptida u proteinima, najčešće predstavljenih u vektorskom prostoru sa manjim brojem dimenzija (Shen i sar., 2007). Sa razvojem neuronskih mreža u poslednjoj deceniji, one postaju najpopularniji algoritam MU za rešavanje kompleksnih problema korišćenjem raznovrsnih skupova podataka od kojih mnoge nije moguće (ili nije jasno kako efektno) predstaviti u formi matrice (dvodimenziona tabela), koja je do skora bila standardan ulaz algoritama MU. Jedan od glavnih razloga za uspešnost neuronskih mreža je mogućnost da pronađu optimalnu numeričku reprezentaciju podataka u vektorskom prostoru sa manje dimenzija za dati problem. Odnosno sama reprezentacija se trenira zajedno sa celim modelom MU. One na ovaj način eliminišu ili drastično smanjuju potrebu za domenskim znanjem vezanim za specifični problem – pitanje >kako da predstavim podatke algoritmu MU?< postaje zastarelo. Sa druge strane, za dobijanje korisnih reprezentacija za rešavanje specifičnog problema potrebna je jako velika količina obeleženih ulaznih podataka (Adadi, 2021). Pošto je opisani skup podataka vezan za hidroksilaciju prolina u biljnim proteinima bio mali odlučeno je da se proteinske sekvence transformišu u numerički

oblik korišćenjem konvencionalnih metoda koje su prethodno opisane, odnosno da se ne koriste neuronske mreže. Na ovakav izbor je uticalo i to što su metode MU za predviđanje Hyp opisane u literaturi takođe koristile konvencionalne pristupe transformacije proteinske sekvence u numerički oblik (Ismail i sar., 2016; Shi i sar., 2015; Qiu i sar., 2016; Xu i sar., 2014).

Za predviđanje verovatnoće hidroksilacije prolina u zavisnosti od lokalne sekvence izabrano je nekoliko klasičnih načina transformisanja proteinske sekvence u numerički oblik što je rezultovalo sa 16 različitih grupa atributa koji su zajedno imali 1294 jedinstvena atributa (Tabela 4, odeljak 3.1.1.). Iako možda izgleda preterano predstavljati sekvencu dugu 21 aminokiselinu sa oko ~1300 numeričkih vrednosti, treba imati u vidu da one predstavljaju samo mali broj mogućih kombinacija, i da broj različitih 21-mernih sekvenci sa po 10 aminokiselina oko centralnog Pro, koji je neizmenjiv, iznosi 20^{20}. Pošto mnogi od korišćenih atributa nisu značajni za sam model, odnosno ne pružaju nikakve dodatne informacije u odnosu na druge, već prisutne atribute, isprobano je nekoliko metoda za odabir atributa, dve filter metode IGr (Quinlan, 1986) i mRMR (Peng i sar., 2005; Zhao i sar., 2019) i jedna metoda bazirana na omotač algoritmima za odabir atributa sfs (Kohavi i John, 1997). Treba napomenuti da čak i algoritmi MU koji tokom samog treniranja vrše odabir atributa (engl. *„embedded feature selection*“), kao što su XGB i RF mogu imati koristi od prethodnog odabira atributa što je rezultatima i pokazano (Slika 5). Tokom odabira modela evaluirana su četiri algoritma MU: KNN, SVM, RF i XGB, od kojih su SVM i RF već korišćeni za modele koji predviđaju Hyp na osnovu lokalne sekvence (Shi i sar., 2015; Ismail i sar., 2016; Qiu i sar., 2016).

Sve poređene metode za odabir atributa, rezultirale su u malom povećanju performansi modela u poređenju sa modelima bez selekcije atributa, osim za SVM koji je pokazao jako slabe performanse u slučaju bez odabira atributa ili uz mRMR filter algoritam (Slika 5). Model sa najboljim performansama napravljen je korišćenjem kombinacije sfs i XGB algoritama (Slika 5). Ovo nije iznenadjujuće pošto je u nekoliko godina od kada je postao dostupan, XGB (Chen i Guestrin, 2016) algoritam MU se pokazao ili kao najbolji ili među najboljim algoritmima po performansama za rešavanje takozvanih tabelarnih problema (predstavljeni matricom). Ovo je jasno i na osnovu broja uspešnih klonova/adaptacija ovog algoritma kao što su visoko citirani *Microsoft*-ov *LightGBM* (Ke i sar., 2017) i *catboost* (Prokhorenkova i sar., 2018).

Odabrani model je procenjen dodatnom unakrsnom validacijom upotrebom prethodno odabranih grupa atributa (Slika 6) i korišćenjem skupa podataka za evaluaciju koji ni na koji način nije korišćen za treniranje algoritma (Slika 9) i rezultati pokazuju njegove visoke performanse. Performanse modela na skupu podataka za evaluaciju su upoređene sa ekvivalentnim modelima MU opisanim u literaturi (Ismail i sar., 2016; Shi i sar., 2015; Qiu i sar., 2016; Xu i sar., 2014) i samo je jedan od njih - *RF-Hydroxysite* (Ismail i sar., 2016) imao nešto niži ali uporediv nivo performansi (Slika 9). Ovde treba dodati da urađeno poređenje treba uzeti sa rezervom pošto nije poznato da li su, i u kojoj meri, podaci za validaciju korišćeni za poređenje algoritama upotrebljavani pri konstrukciji modela iz literature. Moguće je da je deo performansi *RF-Hydroxysite* posledica preklapanja skupa podataka korišćenih za njegovo treniranje i ovde korišćenog skupa za validaciju. Ovo nije provereno, a čak i da jeste, bilo bi gotovo nemoguće konstruisati skup za validaciju koji ne sadrži nijednu proteinsku sekvencu korišćenu za treniranje modela iz literature. Ne mala prednost modela treniranog u okviru izrade ove teze u odnosu na *RF-Hydroxysite* server je to što omogućava predviđanje pozicija hidroksilacije u čitavim proteomima za relativno kratko vreme. Primera radi, predviđanje na celom proteomu *A. thaliana* (TAIR10) koji se sastoji od nešto više od 48 hiljada proteinskih sekvenci traje oko 15 minuta korišćenjem jedne procesorske niti Intelovog i7 8700 procesora. Paralelizacija predviđanja je trivijalna pošto je potrebno samo podeliti proteinske sekvence u

odgovarajući broj paralelnih niti, a ubrzanje je direktno proporcijalno broju korišćenih procesorskih niti.

Ispitivanje značajnosti atributa za predviđanja modela pokazalo je da postoji relativno mali broj dominantnih atributa koji utiču na performanse modela (Slika 10) što je posledica pre svega velikog preklapanja korišćenih atributa u informaciji koju nose o hidroksilaciji prolina. Ovo ukazuje da je moguće model uprostiti korišćenjem manjeg broja atributa za treniranje, što je kasnije i pokazano (Dragićević i Simonović, 2021). Ovo pak nije dovelo do poboljšanja performansi modela. Dominantni atributi nose već poznate informacije o hidroksilaciji Pro u proteinskim sekvencama biljaka (Canut i sar., 2016; Duruflé i sar, 2017), a to su da lokalna frekvencija prolina u sekvenci povećava verovatnoću hidroksilacije, glavna aminokiselina koja utiče na hidroksilaciju je ona koja prethodi prolinu u sekvenci, i da prolin oko koga se nalaze AG motivi u većem broju, ima veću šansu da bude hidroksilovan.

Visoke performanse modela ne znače da model ne može biti unapređen korišćenjem drugog algoritma MU, skupa atributa, hiperparametara ili slaganjem nekoliko modela MU u još veći ansambl. Mišljenje autora je da se ipak najveći napredak u generalizaciji modela može očekivati ako se više sekvenci, sa većom raznolikošću, iz različitih biljnih vrsta koriste za treniranje. Kako se bude povećavao broj biljnih sekvenci sa eksperimentalno određenim pozicijama Hyp povećavaće se i moć generalizacije modela napravljenih na osnovu njih.

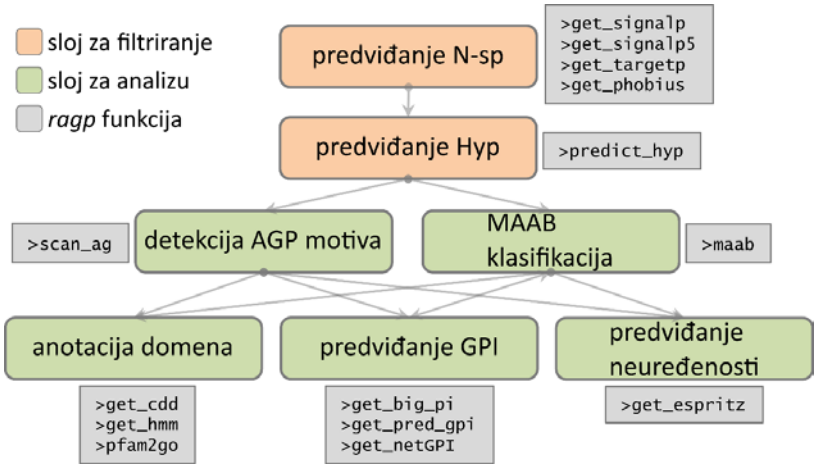
Treba dodati, da je u periodu izrade ove teze zabeležena velika uspešnost lingvističkih modela MU baziranih na neuronskim mrežama sa transformer arhitekturom (Vaswani i sar., 2017), kao i razvoj i široka upotreba takozvanog transfer-učenja (Tan i sar., 2018). Transfer-učenje podrazumeva da se model treniran na jednom skupu podataka za rešavanje nekog problema prilagodi sa drugim skupom podataka za rešavanje nekog alternativnog problema. Ovo je od velikog značaja za modele MU trenirane na proteinskim sekvencama jer omogućava da se sve dostupne proteinske sekvence koriste za pravljenje generalnih lingvističkih modela, samo-supervizovanim učenjem na osnovu bioloških sekvenci, koji se kasnije fino podešavaju za specifične zadatke sa malom količinom obeleženih podataka (Elnaggar i sar., 2020; Rives i sar., 2021; Teufel i sar., 2022). Za sada izgleda da opisana metodologija MU ujedno predstavlja i put za buduća poboljšanja modela za predviđanje hidroksilacije prolina koja ne bi zavisila od tempa eksperimentalne anotacije Hyp u sekvencama. Trenutno jedinu prepreku ovome predstavlja relativno velika računarska zahtevnost prilikom manipulacije ovakvim modelima, i to ne samo tokom treniranja, već i tokom samog predviđanja što ograničava njihovu upotrebnu vrednost.

5.1.2. Identifikacija i analiza HRGP korišćenjem znanja o Hyp

Prema literaturi su sekvence HRGP identifikovane na osnovu aminokiselinskog sastava kombinovanog sa prisustvom karakterističnih aminokiselinskih motiva (Showalter i sar., 2010; Johnson i sar. 2017a; Ma i sar., 2017). Postoji nezanemarljiv broj himernih proteina koji pored HRGP motiva sadrže i specifične domene, a koje je teže identifikovati na osnovu datih osobina. Sekvence koje sadrže domene za koje je poznato da asociraju sa HRGP regionima je moguće identifikovati na osnovu homologije (Showalter i sar., 2010; Simonović i sar., 2016), ali na ovaj način nije moguće identifikovati domene za koje je do sada nepoznato da asociraju sa HRGP regionima. Među HRGP sekvencama, AGP su najkomplikovaniji za identifikaciju pošto nemaju dobro definisane motive - nije poznato koliko AG motiva je neophodno i koliki razmak između njih je dozvoljen da bi motiv bio hidroksilovan, kao preduslov za glikozilaciju. Pronalaženje takvih sekvenci bi bilo jednostavno da su pozicije Hyp poznate unapred, što je

nemoguće zbog cene i truda potrebnog za eksperimentalnu identifikaciju PTM proteina. Potencijalno rešenje je da se postojeća znanja o pozicijama Hyp u proteinskim sekvencama iskoriste za pronalaženje novih HRGP sekvenci u proteomima različitim biljnih vrsta. Pomenut pristup je primenjen u *ragp* paketu (Slika 30) i to kroz dva sloja analize:

- Sloj za filtriranje sekvenci služi za identifikaciju proteina koji imaju N-sp komunikacijom sa *TargetP1.1* (Emanuelsson i sar., 2007), *SignalP4.1* (Petersen i sar., 2011) i *Phobius* (Käll i sar., 2007), a od nedavno je implementiran *SignalP5* (Almagro Armenteros i sar., 2019). Potom se iz skupa sekvenci za koje je predviđeno da sadrže N-sp filtriraju proteini koji sadrže nekoliko predviđenih Hyp korišćenjem implementiranih modela MU.
- Sloj za analizu u kome se filtrirane sekvence analiziraju na prisustvo AG motiva, klasifikuju u MAAB klase (Johnson i sar., 2017a), identifikuju se domeni (Finn i sar., 2011; Lu i sar., 2020), transmembranski regioni (Käll i sar., 2007), potencijalna mesta za vezivanje GPI-sidra (Eisenhaber i sar., 2003; Pierleoni i sar., 2008; Gíslason i sar., 2021) i neuređeni regioni proteina (Walsh i sar., 2012) komunikacijom sa pomenutim Internet serverima. Svi serveri su birani tako da omogućavaju visoku propusnost analiza kako bi *ragp* paket bio upotrebljiv za anotaciju celih proteoma.



Slika 30. Šema analize sekvenci korišćenjem *ragp* paketa. Bojom kajsije predstavljen je sloj za filtriranje sekvenci, zelenom bojom je predstavljen sloj za analizu sekvenci, sivi kvadrati označavaju nazive *ragp* funkcija kojima se postižu opisani zadaci u okviru dva sloja analize

Modeli MU za predviđanje Hyp su ugrađeni u *ragp* funkciju *predict_hyp* i njihova predviđanja se mogu koristiti na dva načina:

- Sekvence koji sadrže manje od određenog broja predviđenih Hyp se ne uzimaju za dalja razmatranja. Preporuka je da minimalan broj predviđenih Hyp u sekvencama bude tri, ali zavisno od cilja može biti veći ili zavisiti od dužine sekvence. Ukoliko je potrebno identifikovati sekvence sa visokom sigurnošću može se podići prag odluke prilikom predviđanja hidroksiporlina.
- Sekvence u kojima su, na osnovu predviđanja Hyp, P zamenjeni sa O mogu se koristiti za skeniranje AG motiva funkcijom *scan_ag* koja, u tom slučaju uzima u obzir samo AG motive u kojima je predviđen Hyp. U ovoj funkciji omogućeno je skeniranje AG motiva postavljanjem minimalnog broja neophodnih motiva, maksimalnog broja

aminokiselina između dva motiva, tipova dipeptida koji se ubrajaju u AG motive kao i mogućnost maskiranja EXT motiva.

Kombinacija navedenih alata predstavlja novu metodologiju za filtriranje i analizu HRGP sekvenci, prvenstveno namenjenu za identifikaciju i analizu AGP sekvenci iz biljnih proteoma. *Ragp* paket je jednostavan za upotrebu; funkcije paketa imaju konzistentan unos i prihvataju širok spektar ulaznih objekata (od bazičnih struktura podataka u R do *.FASTA* fajlova). Podrazumevane vrednosti za dodatne argumente funkcija su pažljivo odabrane i u većini slučajeva se mogu ostaviti kako su i zadate. Izlazne vrednosti funkcija su takođe bazične R strukture kojima se lako može manipulisati korišćenjem različitih R paketa sa specijalizovanim namenama.

5.1.3. Anotacija HRGP sekvenci iz 62 biljna proteoma

Ragp paket je primenjen na proteinske sekvence iz 62 biljna proteoma kako bi se stekao utisak o varijabilnosti identifikovanih HRGP sekvenci u biljnom svetu i sa ciljem testiranja sposobnosti paketa. Filtriranjem sekvenci na osnovu predviđanja Hyp i njihovom MAAB klasifikacijom iz 62 biljna proteoma identifikovano je 99,29% prototipskih HRGP sekvenci (MAAB klase 1-23, Tabela 1, Slika 12) u poređenju sa MAAB klasifikacijom bez predviđanja Hyp, pri čemu su pronađene sve prototipske sekvence HRGP u 51 od 62 analizirana biljna proteoma (Slika 12). Ovo može da znači ili da kriterijumi za MAAB klasifikaciju mogu da uključe i sekvence koje nemaju barem tri Hyp, ili da predviđanja Hyp pomoću modela MU ne funkcionišu jednako dobro na sekvencama poreklom od svih biljnih vrsta. Identifikovan je priličan broj potencijalnih himernih AGP sekvenci: oko 30000 (u proseku oko 500 po proteomu) sa različitim domenima, što ukazuje da prisustvo Hyp u AG motivima nije limitirano na par tipova proteina sa specifičnim domenima. Za domene *Tryp_alpha_amyl* (PF00234) i LTP_2 (PF14368) koji su karakteristični za proteine lipidnog transfera (engl. „*plant lipid-transfer proteins*“) zatim protein sličan plastocijaninu (engl. „*Plastocyanin-like“*, PF02298) i fasciklin (PF02469) od ranije je poznato da su mogu da se kombinuju sa AGP regionima u himernim AGP (Huang i sar., 2021). Za domene X8 (PF07983) i glikozid hidrolaznu familiju 17 (*Glyco_hydro_17* – PF00332) je nedavno, na osnovu bioinformatičke analize sekvenci, predloženo da mogu činiti himerne AGP (Ma i sar., 2017). Čak i kada se kriterijumi predviđanja Hyp povise podizanjem praga odluke što povećava specifičnost, i u obzir uzimaju sekvence koje imaju veći broj predviđenih Hyp i dalje je broj identifikovanih AGP himernih sekvenci znatan (rezultat nije prikazan). Sve ovo uzeto zajedno otvara pretpostavku da O-glikozilacija na Hyp u ekstracelularnim biljnim proteinima predstavlja relativno često pojavu, potencijalno uporedivu sa N-glikozilacijom. Ipak da bi se ovo potvrdilo nije dovoljno čak ni eksperimentalno dokazivanje da je Hyp prisutan u pomenutim proteinima već je potreban i eksperimentalan dokaz da je ovakav Hyp glikozilovan.

Interesantno je da mali broj publikacija dovodi u vezu PK/PTK domen sa AGP regionima (Hwang i sar., 2016; Ma i sar., 2017; Pfeifer i sar., 2020), naročito pošto je on najfrekventnije identifikovan domen u himernim AGP prema analizi urađenoj ovde. Na osnovu prikazane strukture himernih AGP – receptor kinaza *A. thaliana* (Slika 14), može se zaključiti da glikozilovani ekstracelularni regioni ovih proteina verovatno služe za prijem signala u ćelijskom zidu koji se onda šalju dalje u ćeliju posredstvom kinaznog domena. Ovo je u saglasnosti sa predloženom ulogom himernih AGP kao ekstracelularnih senzora (Showalter i Basu, 2016). Na osnovu iznetih rezultata se čini da AGP ne samo da interaguju sa ekstracelularnim regionima receptor kinaza kao što su pretpostavili Showalter i Basu (2016), već mogu da budu i sastavni deo ovih proteina.

5.2. Identifikacija i analiza *HRGP* gena *C. erythraea*

5.2.1. Veličina i raznolikost HRGP superfamilije proteina *C. erythraea*

Poslednjih godina razvijeno je nekoliko pristupa identifikacije potencijalnih HRGP, naročito AGP sekvenci, u pokušaju da se prevaziđu ograničenja klasičnih pristupa. MAAB sistem klasifikacije razvijen je da bi se prikazalo postojanje kontinuuma HRGP sekvenci sa zajedničkim karakteristikama EXT, AGP i PRP (Johnson i sar., 2017a). Sa druge stane, *ragp* paket je pre svega napravljen za identifikaciju AGP sekvenci sa malom zastupljenošću AG motiva kombinacijom algoritama MU za predviđanje Hyp sa fleksibilnom detekcijom AG motiva koji sadrže predviđen Hyp. Oba pristupa iskorišćena su za identifikaciju HRGP kičice.

Korišćenjem modifikovane MAAB klasifikacije, inkorporirane u *ragp* paket, na predviđenim proteinskim sekvencama *de novo* sastavljenog transkriptoma *C. erythraea* (Ćuković i sar., 2020, https://zenodo.org/record/3591805) 126 transkripata (134 proteinske sekvence) je klasifikovano u MAAB HRGP klase 1-23, dok je preostalih 87 transkripata (kojima odgovara isti broj proteina) klasifikovano kao MAAB klasa 24 (Slika 16A). Tokom analize HRGP iz 62 biljna proteoma (Slika 12) samo četiri vrste biljaka (*Hordeum vulgare, Physcomitrella patens, Triticum aestivum* i *Zea mays*) i jedna vrsta algi (*Chlamydomonas reinhardtii*) je imalo više od 100 proteinskih sekvenci klasifikovanih kao MAAB klase 1-23. Naizgled veliki broj HRGP sekvenci identifikovanih u transkriptomu kičice se može objasniti kombinacijom bioloških i tehničkih razloga. BUSCO procena kompletnosti (Simao i sar., 2015) transkriptoma *C. erythraea* ukazivala je na visok procenat kompletnih sekvenci (95%), ali i visok procenat dupliranih kompletnih sekvenci (54%), verovatno zbog toga što je kičica tetraploid (Ćuković i sar., 2020). Heterozigotni aleli koji se nisu sklopili tokom *de novo* sastavljanja transkriptoma, dovodeći do duplikacija, mogu biti uzrok velikog broja sekvenci klasifikovanih u MAAB grupe 1-23. Razdvajanje alela i paraloga tokom *de novo* sastavljanja transkriptoma je izazovno, posebno za poliploidne organizme, i *Trinity,* kao i drugi programi slične namene, mogu stvoriti himere među njima (Yang i Smith, 2013; Gruenheit i sar., 2012) što dovodi do dodatnog povećanja redundantnosti sekvenci. Iako postoje metode za smanjivanje redundantnosti (Yang i Smith, 2013; Ono i sar., 2015; Kerkvliet i sar., 2019) one nisu ovde primenjene kako bi se izbegao gubitak pravih bioloških entiteta (sekvenci). Zbog svega ovoga direktno poređenje *Phytozome* proteoma, koje uglavnom čine proteomi poreklom od haploidnih genoma, sa proteomom poreklom od *de novo* sastavljenog transkriptoma tetraploidne biljke može dati samo grube procene.

Većina HRGP sekvenci kičice su klasični AGP (MAAB klase 1 i 4), dok su druge klase, osim MAAB klase 24, ili nedovoljno zastupljene ili u potpunosti odsutne (Slika 16A). Primetno je da je broj sekvenci klasifikovan u MAAB klase, gotovo u svakoj grupi, dosta veći u poređenju sa Johnson i sar. (2017b) koji su primenili MAAB klasifikaciju na veliki broj biljnih transkriptoma. Verovatno objašnjenje je to što u ovoj studiji nisu isključene sekvence sa predviđenim domenima ili kraće proteinske sekvence, sa manje od 90 aminokiselina. Smatram da je informativnije ne isključivati sekvence pre klasifikacije, već ih nakon klasifikacije pregledati za prisustvo domena (Slika 16B) i eventualno distribuciju dužine. Drugi potencijalni razlog je prethodno pomenuta redundantnost sekvenci kičice u dostupnom transkriptomu. Činjenica da većina sekvenci kičice koje su klasifikovane kao MAAB 1-23 pripada klasičnim AGP (klase 1 i 4, Slika 16A), ne mora nužno značiti da kičicu karakteriše mala raznovrsnost HRGP sekvenci. Johnson i sar. (2017b) su pokazali da je za MAAB klasifikaciju, ključno koristiti nekoliko različitih dužina k-mera prilikom *de novo* sastavljanja transkriptoma. Kada

se koriste samo 25-mere, što je uobičajeno za većinu programa koji služe za *de novo* sastavljanje transkriptoma na osnovu de Bruijin grafova uključujući i *Trinity* koji je ovde korišćen, primećena je manja zastupljenost visoko repetitivnih HRGP sekvenci kao što su CL-EXT i PRP (Johnson i sar., 2017b). Pošto je postavljeni cilj bio identifikacija AGP sekvenci kičice, i to naročito onih sa malom zastupljenosti AG motiva, dužina k-mera korišćena za *de novo* sastavljanje transkriptoma ne bi trebalo da ima značajan uticaj na pronalaženje sekvenci, što je i potvrđeno velikim brojem pronađenih AGP sekvenci (Slika 16A, 17). Takođe, distribucija sekvenci kičice po MAAB klasama je delimično neutvrđena pošto jednom delu predviđenih proteinskih sekvenci nedostaje C-terminus (transkripti koji nisu sastavljeni u celosti), pa predviđanje GPI, neophodno za određivanje pojedinih MAAB klasa može biti pogrešno.

MAAB klasifikacija se bazira na razvrstavanju sekvenci koje imaju veliki udeo karakterističnih aminokiselina koji čine HRGP motive (preko 45% P, A, S i T u slučaju AGP). Zato ne čudi što su MAAB klasifikacijom prepoznate sekvence koje imaju nekoliko kratkih *Pfam* domena ili češće nijedan (Slika 16A, B). Prisustvo većih domena, kao što je fasciklinski (PF02469), koji je najčešće povezan sa AGP (Johnson i sar., 2017a; Meng i sar., 2020) umanjilo bi karakteristični aminokiselinski sastav ovih sekvenci. Zbog toga sekvence iz samo nekoliko MAAB klasa, pre svega klasični GPI-AGP (klasa 1) imaju *Pfam* domene (Slika 16A). Za neke od ovih domena, kao što su nespecifični proteini za transfer lipida (PF00234 i PF14368) i proteini slični plastocijaninu (PF02298) je odavno poznato da su povezani sa AGP, dok su drugi, kao X8 domen (Showalter i sar., 2010), tek od nedavno povezani sa AGP bioinformatičkom analizom (Ma i sar., 2017; Dragićević i sar., 2020; Pfeifer i sar., 2020).

Namena *ragp* paketa, pored pronalaženja predstavnika prototipskih HRGP preko MAAB klasifikacije, je identifikacija sekvenci koje sadrže kratke regione nalik AGP. Brojne sekvence su identifikovane korišćenjem relaksirane i stroge pretrage motiva (822 i 330 proteinskih sekvenci, Slika 15) od kojih su mnoge asocirane sa različitim *Pfam* domenima (Slika 17). Pored već korišćene relaksirane detekcije AG motiva kakva je korišćena za analizu sekvenci poreklom iz 62 biljna proteoma (odeljci 4.1.7. i 5.1.3.) dodata je i takozvana stroga detekcija koja podrazumeva prisustvo većeg broja AG motiva sa Hyp i manji razmak između njih u sekvencama, kako bi se procenile potencijalno lažno identifikovane sekvence relaksiranom pretragom. Oko 40% sekvenci identifikovanih relaksiranom pretragom, pronađeno je i prilikom stroge pretrage. Ovo ne znači da je preostalih 60% sekvenci lažno detektovano, ali se na osnovu distribucije sekvenci sa *Pfam* domenima, identifikovanim sa ove dve pretrage, može zaključiti koje sekvence potencijalno nisu AGP. Nijedna sekvenca koja je sadržala PF04674 domen (engl. „*phosphate-induced protein 1 conserved region*“) nije prošla strogu detekciju pa je moguće da predstavlja lažno identifikovane AGP prilikom relaksirane pretrage. Strogom pretragom identifikovano je malo sekvenci sa forminskim domenom (engl. „*Formin Homology 2 Domain*“, PF02181.24), što potencijalno ukazuje da ove sekvence verovatno nisu himerni AGP, ili da sadrže isključivo kratke nizove AG motiva sačinjene od tri dipeptida. Ostali frekventno pronađeni domeni u sekvencama kičice, identifikovani kao AGP (Slika 17), su već povezani sa AGP na osnovu bioinformatičkih (Showalter i sar., 2010; Ma i sar., 2017; Dragićević i sar., 2020) ili eksperimentalnih dokaza (Johnson i sar., 2003; Mashiguchi i sar., 2004).

5.3. Analiza ekspresije odabranih *AGP* gena kičice

U prvom delu istraživanja obuhvaćenih ovom doktorskom disertacijom napravljen je R paket za identifikaciju i analizu AGP sekvenci sa ciljem identifikacije sekvenci koje kodiraju arabinogalaktanske proteine kičice. Od identifikovanih AGP sekvenci, izabrano je 18 (po šest iz svake grupe: FLA, KLA i AGp) za detaljniju filogenetsku analizu, a ispitana je i njihova ekspresija u različitim organima kao odgovor na mehaničke povrede i dugotrajno gajenje na različitim koncentracijama βGlcY kao i tokom različitih morfogenetskih puteva *in vitro* i u različitim delovima biljaka iz prirode.

5.3.1. Filogenetske veze i strukturne odlike izabranih sekvenci FLA, KLA i AGp

FLA predstavljaju detaljno proučavanu familiju proteina sa pretpostavljenom ulogom u adheziji ćelija (Meng i sar., 2020; He i sar., 2019; Johnson i sar., 2003; Seifert, 2018; Shafee i sar., 2020). Izabrani su za dalja proučavanja u ovom radu jer je prethodno pokazano da njihova ekspresija raste tokom morfogeneze kičice gajene u uslovima *in vitro* (Simonović i sar., 2015). Većinu odabranih FLA kičice karakteriše prisustvo jednog ili dva fasciklinska domena (PF02469) koji su okruženi AG motivima i za sve, sem za CeFLA3, je predviđeno prisustvo GPI-sidra (Slika 18). CeFLA3 je parcijalna sekvenca kojoj nedostaje C-terminus tako da je moguće da i je i ovaj protein pričvršćen za membranu GPI-sidrom. C-terminalni TM region predviđen kod većine FLA sekvenci (Slika 18) je odlika pre-proteina, pošto GPI signalna sekvenca sadrži niz hidrofobnih aminokiselina sa osobinama transmembranskog α-heliksa koji pričvršćuje pre-proteine za luminalnu stranu membrane ER dok se ne odigra transamidacija sa GPI (Kinoshita i Fujita, 2016; Galian i sar., 2012).

KLA geni su proučavani jer su AG motivi od skora povezani sa nizom biljnih receptor kinaza (Hervé i sar., 2016; Pfeifer i sar., 2020; Ma i sar., 2017; Dragićević i sar., 2020). Kao i kod srodnih sekvenci iz drugih biljnih vrsta, KLA sekvence kičice sastoje se od ekstracelularnog regiona sa AG motivima, TM regiona (koji nedostaje samo kod CeKLA5) i intracelularnog protein kinaznog domena (Slika 19). Prethodno je pokazano da su protein kinazni (PF00069.25) i protein tirozin kinazni (PF007714.17) domen najčešći *Pfam* domeni na osnovu *Pfam* 32 baze nađeni u proteinskim sekvencama sa AG motivima kada se koristi relaksirana pretraga među 62 analizirana biljna proteoma (Slika 13). Ažuriranjem *Pfam* baze na verziju 33, promenjena je anotacija PF007714.17 iz „protein tirozin kinaze“ u „protein tirozin i serin/treonin kinaze“ (verzija PF007714.18) što je u skladu sa činjenicom da je većina identifikovanih receptor kinaza kičice sa AG motivima sa predviđenim Hyp homologna sa serin/treonin kinazama iz drugih biljaka. Prema Unirpot bazi podataka su kao serin/treonin kinaze anotirani homolog CeKLA3 iz *C. canephora* CDP09635 (Uniprot: A0A068UM75), homolog CeKLA5 iz *C. canephora* CDP10466 (Unirpot: A0A068UPV3), homolog CeKLA2 iz *C. canephora* CDP03159 (Unirpot: A0A068U4T9) i homolog CeKLA7 iz *G. max* KRH17308 (Unirpot: C6ZRR4).

Sekvence kičice CeKLA1, CeKLA6 i CeKLA7, zajedno sa članovima klastera C, D i E (Slika 19), su receptor kinaze sa leucinom bogatim ponovcima u ekstracelularnom regionu (LRR-RLK), himerni AGP uključeni u protein-protein interakcije i signalnu transdukciju (Ma i sar., 2017). LRR je često asociran sa AGP i HRGP regionima (Slike 13A, 16B, 17) i ima ulogu u dimerizaciji molekula koji ga sadrže (Ma i sar., 2017). CeKLA3 i homologne sekvence

iz A1 podklastera (Slika 19) pripadaju familiji cisteinom-bogatih receptora nalik kinazama (CRK) (Chen, 2001; Zuo i sar., 2020). Njih karakteriše prisustvo N-terminalnog domena (PF01657) sa konzerviranim Cys, koji grade disulfidne mostove, a koji je zadužen za odgovor biljaka na stres izazvan solima i gljivicama (PF01657).

AG peptidi, kao podgrupa kratkih klasičnih AGP, teško se pronalaze na osnovu homologije pošto najveći deo sekvence čine signalni peptidi N-sp i GPI-sp koji se odsecaju tokom sazrevanja pre-proteina i koji najčešće nisu karakteristični za ovu klasu proteina. Preostali deo je varijabilne sekvence, sa rasutim AG motivima. Ipak, jedan AG peptid, CeAGp3, identifikovan je u prethodnoj verziji transkriptoma *C. erythraea* (Malkov i Simonović, 2011) pretragom po homologiji sa poznatim AGp sekvencama (Simonović i sar., 2015) zahvaljujući prisustvu domena nepoznate funkcije DUF1070 (engl. „*domain of unknown function*“). Kasnije je pokazano da DUF1070 predstavlja GPI signalnu sekvencu, kojoj prethode AG motivi, karakterističnu za AG peptide (Simonović i sar., 2016), tako da je anotacija domena u *Pfam* bazi promenjena u arabinogalaktanski peptid (PF06376). Koliko je poznato, PF06376 je jedini domen pronađen isključivo u AGP i HRGP sekvencama. Zajedno sa CeAGp3, CeAGp7 i CeAGp9, koji sadrže PF06376 domen, jedan GPI-usidren peptid bez PF06376 domena (CeAGp6) kao i dve parcijalne sekvence (CeAGp8 i CeAGp10) izabrane su za dalju analizu (Slika 20).

Većina proučavanih FLA i KLA sekvenci kičice grupisana je zajedno sa homolognim sekvencama iz taksonomski najbliže vrste *C. canephora*, što ukazuje na međusobnu filogenetsku povezanost sekvenci u ovim proteinskim familijama. Takvo saznanje je bilo očekivano s obzirom da obe grupe sekvenci sadrže relativno duge i konzervirane domene sa strukturno i funkcionalno ograničenim evolutivnim stopama. Sa druge strane, filogenetsko stablo AGp ne pokazuje takav trend i sadrži samo jedan pouzdan klaster (Slika 20) zahvaljujući prisustvu PF06376 domena koji omogućava ekstrakciju određenog filogenetskog signala.

5.3.2. Odabir referentnih gena za ispitivanje ekspresije *AGP* gena

Za robusno određivanje ekspresije gena potrebno je koristiti adekvatne referentne gene za normalizaciju (Vandesompele i sar., 2002; Andersen i sar., 2004). Zbog toga je ispitana stabilnost ekspresije jedanaest gena tokom odgovora kičice na mehaničku povredu (Slika 21). Za ispitivanje stabilnosti ekspresije *AGP* gena prilikom gajenja na različitim koncentracijama βGlcY, odabrana su četiri gena (*TBP1*, *RPL2*, *AK* i *RAN*) koja su u radu Ćuković i sar. (2020), pokazala stabilan nivo ekspresije. Pokazano je da *TBP1*, *RPL2*, *AK* i *RAN* imaju stabilnu ekspresiju prilikom odgovora biljke, gajene u uslovima *in vitro*, na mehaničku povredu (Slika 21), kao i u većini grupisanja uzoraka poreklom od biljaka iz prirode, biljaka gajenih u uslovima *in vitro* i uzoraka iz različitih morfogenetskih faza i procesa kičice (Ćuković i sar. 2020). Ovi geni su i prethodno ispitivani kao kandidati za referentne gene u različitim eksperimentalnim uslovima kod različitih vrsta biljaka. *RPL2* je odabran kao jedan od gena sa najstabilnijom ekspresijom u biljkama *Solanum lycopersicum* koje su rasle u uslovima nedostatka azota u zemljištu, niskih temperatura i nedovoljno svetlosti (Løvdal i Lillo, 2009), u različitim razvojnim fazama i uslovima biotičkog stresa jabuke (Kumar i Singh, 2015) i nakon infekcije *Cucumis melo* sa *Fusarium oxysporum* (Sestili i sar., 2014). Pokazano je da se *TBP1* stabilno eksprimira u eksperimentima sa papajom gajenom u različitim uslovima (Zhu i sar., 2012) kao i kod *Solanum melongena* L. tokom faze sazrevanja ploda (Kanakachari i sar., 2016). *AK* nije često korišćen kao referentni gen, mada je pokazao stabilnost ekspresije kod četinara u različitim razvojnim fazama (de Vega-Bartol i sar., 2011). *RAN* kodira GTP-vezujuće

proteine i bio je među najbolje rangiranim referentnim genima u različitim uzorcima tkiva banane (Chen i sar., 2011), a pokazao je veliku stabilnost i u eksprerimentima sa papajom (Zhu i sar., 2012).

5.3.3 Ekspresija odabranih *AGP* gena nakon mehaničke povrede lista i korena kičice gajene u uslovima *in vitro*

Biljke su, kao sesilni organizmi, konstantno izloženi nekom vidu stresa iz spoljašnje sredine od kojih su najčešće mehaničke povrede usled delovanja različitih faktora spoljašnje sredine (León i sar., 2001; Savatin i sar., 2014) ili usled napada patogena.

U literaturi je implicirana uloga AGP u odgovoru na povrede kod različitih biljnih vrsta, pre svega na osnovu promene ekspresije gena koji ih kodiraju. Kod paradajza se nakon mehaničke povrede povećava ekspresija *LeAGP1* koji kodira AGP bogate lizinom i *SlAGP4* koji kodira klasičan AGP. Nasuprot ovome, homolog *LeAGP1* iz *Nicotiana alata*, *NaAGP4* imao je snižen nivo ekspresije ubrzo nakon povrede tkiva ili infekcije patogenom (Gilson i sar., 2001). Ovo ukazuje da čak i homologni AGP mogu imati različite uloge prilikom odgovora različitih biljnih vrsta na povredu.

Kod kičice je relativno mali broj ispitivanih *AGP* gena pokazao značajnu promenu ekspresije nakon povrede lista ili korena (Slike 24 i 25). Statistički značaj tih promena je u većini slučajeva prouzrokovan malom varijansom ekspresije u biološkim ponavljanjima. Među ispitivanim genima se izdvajaju *CeAGp6,* koji je pokazao najveću promenu ekspresije nakon povreda listova od svih ispitivanih gena, i *CeAGp7* sa nešto manjom promenom ekspresije (Slika 24). Oba gena su imali povećanu ekspresiju, koja je kulminirala 6 h nakon povrede lista. Interesantno je da su ovi geni pokazali relativno veliku promenu u ekspresiji i nakon povreda korena sa tom razlikom što je zapaženo smanjenje njihove ekspresije (Slika 25). Treba dodati da je inicijalan (kontrola, 0 h) nivo ekspresije oba gena viši u korenu, a pomenuta razlika je izraženija za *CeAGp6*. *CeAGp6* kodira propeptid od 64 aminokiseline, koji se, nakon uklanjanja N-sp i GPI signalne sekvence redukuje na dužinu od 16 aminokiselina, sa četiri predviđena Hyp u sklopu AG motiva (27-AOfeafAOAOAOTaes-42). CeAGp7 je propeptid dug 74 aminokiseline sa PF06376 domenom i GPI-sidrom (Slika 20); zreo protein nakon obrade je dug samo 12 aminokiselina sa tri predviđena Hyp u sklopu AG motiva (33-qAOAOAOAatsd-44). Oba proteina pokazuju sličnost sa proteinima *A. thaliana*: CeAGp6 pokazuje sličnost sa At1g55330 (AGP21, Slika 20), a CeAGp7 sa At5g24105 (AGP41, Slika 20). Biološke funkcije arabinogalaktanskih peptida nisu dobro definisane, a jedna od retkih publikacija o funkciji AGp upravo ističe ulogu AtAGP21 u formiranju korenskih dlaka *A. thaliana* (Borassi i sar., 2020). Pokazano je da nedostatak AtAGP21 (T-DNA insercioni mutant) dovodi do abnormalnog (ektopičnog) razvoja korenskih dlaka koji je sličan kao kod brasinosteroidnih (BR) mutanata. Autori su zaključili da AtAGP21 verovatno modifikuje odgovor ko-receptornog para kinaza BRI1-BAK1 (engl. „*Brassinosteroid insensitive 1 - BRI1 Associated receptor Kinase 1*") na BR ili se na drugi način ukršta sa signalnim putem BR preko do sada neidentifikovanih proteina na površini ćelije. Na osnovu ovoga jasno je da su usidreni AG peptidi u stanju da učestvuju u prenosu signala iz ćelijskog zida u ćeliju i da interaguju sa fitohormonskim signalnim putevima. Pošto odgovor na povredu uključuje kompleksnu regulatornu mrežu u kojoj se ukrštaju signalni putevi etilena, jasmonske, salicilne i abscisinske kiseline (Savatin i sar., 2014), kao i brasinosteroida (Saini i sar., 2015), neophodna su dodatna istraživanja da bi se pokazao fiziološki značaj indukcije *CeAGp6* i *CeAGp7* povredama u listu, odnosno njihova represija nakon povreda korena. Ova istraživanje bi trebala da obuhvate

fenotipizaciju mutanata sa sniženom ili isključenom ekspresijom pomenutih gena, kao i ispitivanje aktivnosti njihovih promotora. Pomenuti posao će drastično biti olakšan kada genom kičice bude sekvenciran.

5.3.4 Ekspresija odabranih *AGP* gena nakon dugotrajnog izlaganja eksplantata lista i korena kičice gajenih u uslovima *in vitro* različitim koncentracijama βGlcY

βGlcY reagens specifično precipitira AGP pa se zbog toga koristi za proučavanje njihove funkcije. Tretman ćelijske kulture *A. thaliana* sa βGlcY reagensom izazvao je odgovor sličan odgovoru izazvanom povredama (Guan i Nothnagel, 2004). Pokazano je da vijabilnost ćelijskih kultura *A. thaliana* drastično opada 36 h nakon dodavanja βGlcY, a da βGlcY dovodi do pojačavanja strukture ćelijskog zida i sinteze polisaharida kaloze (Guan i Nothnagel, 2004) što je karakteristično za odgovor biljke na povredu (Chen i Kim, 2009). Gubitak vijabilnosti u ćelijama *A. thaliana* indukovan βGlcY dešava se procesom programirane ćelijske smrti (Gao i Showalter, 1999). Pretpostavka je da agregacija AGP na ćelijskoj membrani, indukovana βGlcY, dovodi do fizičkog stresa koji direktno oštećuje membranu, na sličan način kao insekti ili mehaničke povrede.

Agregacija AGP, nakon dodavanja βGlcY u hranljivu podlogu, smanjuje frekvenciju formiranja sekundarnih somatskih embriona kod *Bactris gasipaes*, ukazujući da su AGP značajni za razvoj somatskih embriona (Steinmacher i sar., 2012). Interakcija βGlcY i AGP utiče na razviće embriona *Brassica napus* (Tang i sar., 2006), kao i na razviće somatskih embriona *Euphorbia pulcherrima* (Saare-Surminski i sar., 2000). Kod kičice βGlcY dovodi inhibicije somatske embrogeneze i organogeneze (Simonović i sar., 2015), kao i do stimulacije regeneracije pupoljaka na korenovima u kulturi *in vitro* (Trifunović i sar., 2014). Upravo je zbog ovih efekata βGlcY na morfogenezu kičice ispitivan uticaj dugotrajnog izlaganja eksplantata različitim koncentracijama ovog reagensa na ekspresiju 18 odabranih *AGP* gena. Dodatak βGlcY je indukovao relativno male promene u ekspresiji ispitivanih gena, uglavnom bez statističke značajnosti (Slike 26 i 27). Statistički značajne promene su zabeležene jedino kod *CeAGp3* u uzorcima lista i korena, kao i *CeFLA5* u uzorcima korena, kod kojih dolazi do povećanja ekspresije sa povećanjem koncentracije βGlcY u hranljivoj podlozi. Ovo je očekivano na osnovu literature (Trifunović i sar., 2014; Simonović i sar., 2015): biljka odgovara na nespecifičnu precipitaciju AGP izazvanu βGlcY reagensom pojačanom ekspresijom gena koji ih kodiraju. Na osnovu ovog rezultata nije moguće spekulisati o ulogama ispitivanih *AGP* gena. Drastično veće promene u ekspresiji ispitivanih gena bi bile očekivane u kratkom periodu nakon izlaganja βGlcY, ali ni to ne bi ukazalo na njihove potencijalne uloge. Generalno se na osnovu eksperimenata u kojima se ispituju fiziološki efekti βGlcY teško mogu izvući zaključci o potencijalnoj ulozi pojedinačnih gena, pošto reagens deluje na celu heterogenu AGP familiju koja se, kao što je u ranijim poglavljima opisano, sastoji od jako velikog broja članova uključenih u veliki broj procesa kod biljaka.

5.3.5. Ekspresija odabranih *AGP* gena u različitim tkivima biljaka gajenih *in vitro* i delovima biljaka iz prirode

Uloga AGP tokom SE i SO na eksplantatima listova i korenova kičice pokazana je korišćenjem niza tehnika zasnovanih na upotrebi βGlcY (Simonović i sar., 2015; Simonović i

sar., 2021; Trifunović i sar. 2014; Trifunović i sar. 2015;) i monoklonalnih antitela (Filipović i sar., 2021), kao i analizom ekspresije gena (Simonović i sar., 2015). Jedino istraživanje gde je ispitivana ekspresija *AGP* gena kičice tokom morfogeneze (Simonović i sar., 2015), bazirano je na transkriptomu upitne pokrivenosti (Malkov i Simonović, 2011), u kome su AGP sekvence identifikovane isključivo na osnovu homologije, usled čega su pronađene samo četiri sekvence (Simonović i sar., 2015), od kojih se dve, *CeAGp3* i *CeFLA1,* nalaze među 18 odabranih gena. Razvoj *ragp* paketa kombinovan sa sastavljanjem transkriptoma kičice sa dobrom pokrivenošću (Ćuković i sar., 2020) uz uspostavljanje eksperimentalnog sistema za ispitivanje morfogenetskih procesa kod kičice (Ćuković i sar., 2020), omogućilo je profilisanje ekspresije *AGP* gena tokom ovih procesa. Uzorci biljaka gajenih u prirodi i u uslovima *in vitro,* tkiva iz različitih faza SE i SO, eksplantata gajenih na hranljivim podlogama sa ili bez regulatora rastenja, kao i uzorci biljaka gajenih u različitim uslovima osvetljenja (Tabela 8) omogućavaju barem grubu procenu efekata *in vitro* kulture, regulatora rastenja i razvojnih faza na ekspresiju odabranih AGP gena (Slika 28).

Pregled profila genske ekspresije *AGP* ukazuje da su, u poređenju sa listovima biljaka u fazi rozete gajenim u uslovima *in vitro* (rl), koji su korišćeni kao referentni uzorci, ostali uzorci tkiva biljaka gajenih u uslovima *in vitro,* posebno korenovi rozete biljaka (rr) i celi klijanci (sd), imali povećan nivo ekspresije AG peptida, naročito *CeAGp10*, ali uporediv nivo ekspresije FLA i KLA transkripata (Slika 28). Nekoliko *AGP*, uključujući *CeAGp8*, *CeAGp10*, *CeFLA3* i *CeKLA3* imali su povišenu ekspresiju u delovima biljaka gajenih u uslovima *in vitro* (uzorci rl, rr, sd i rc) u odnosu na biljke iz prirode (uzorci ln, rn, st, mf i imf, Tabela 7 i Slika 28). Nasuprot njima, FLA transkripti *CeFLA1* i *CeFLA4* su najzastupljeniji među uzorcima iz prirode, posebno u stablu (st). Pošto je stablo biljaka kičice koje rastu u prirodi dosta žilavije u odnosu na stablo biljaka gajenih u uslovima *in vitro,* može se pretpostaviti da pomenuti FLA, kao ćelijski adhezioni molekuli, doprinose čvrstoći stabla. Dva FLA *A. thaliana*, *AtFLA11* i *AtFLA12*, slične građe kao *CeFLA1* i *CeFLA4* sa jednim FAS domenom okruženim AG motivima (Slika 18), su značajni za mehanička svojstva stabla (MacMillan i sar., 2010). Pretpostavlja se da FLA sa jednim FAS domenom doprinose čvrstini i elastičnosti stabla tako što utiču na depoziciju celuloze, pa samim tim i na integritet matrice ćelijskog zida (MacMillan i sar., 2010). *CeFLA1* i *CeFLA4* imaju slične profile ekspresije u različitim uzorcima tkiva (Slika 28) i u odgovoru na povrede lista i korena (Slike 24 i 25), pri čemu su strukturno i filogenetski bliski (Slika 18). Nije iznenađujuće da su ova dva gena imala najveću korelaciju ekspresije u odnosu na sve ostale parove gena (0,93, Slika 29). Pored toga *CeFLA1 i CeFLA4*, zajedno sa još nekoliko gena, imaju višu ekspresiju u embriogenom kalusu u odnosu na ostala embirogena tkiva (ec, Slika 28), što ukazuje na njihovu potencijalnu ulogu u inicijaciji SE. *CeFLA1* (prethodno označen kao *CeAGp1*, GenBank:KC733882, (Simonović i sar., 2015)) je imao povećanu ekspresiju u eksplantatima listova kičice nakon 10 dana gajenja u kulturi na hranljivoj podlozi sa 2,4-D i CPPU, tokom indirektne SE. Na osnovu profila njegove ekspresije zaključeno je da verovatno ne učestvuje u procesu organogeneze (Simonović i sar., 2015). Sličan zaključak se nameće i na osnovu profila ekspresije prikazanih u ovom radu pošto *CeFLA1* nema povećanu ekspresiju u uzorcima tkiva uzetim tokom procesa organogeneze (oc, ablh, abl i abr, Slika 28)

Značaj AGP tokom SE kod kičice je do sada potvrđen u brojnim radovima (Simonović i sar., 2015; Filipović i sar., 2021; Simonović i sar., 2021). Činjenica da postoji dinamična promena AGP epitopa tokom razvoja somatskih embriona (Filipović i sar., 2021) ukazuje da verovatno postoji više AGP (ili njihovih glikoformi), uključenih u sam proces, nego što je detektovano do sada.

Što se tiče procesa SO u uslovima *in vitro*, nijedan od izabranih 18 gena čija je ekspresija ispitivana, nije pokazao povećan nivo ekspresije u organogenom kalusu (oc) ili u

adventivnim pupoljcima nezavisno od njihovog porekla (ablh, abl i abr), u poređenju sa rl (Slika 28). Ipak, *CeAGp7*, *CeAGp9* i, u manjoj meri, *CeAGp6* i *CeKLA3* imaju nešto sniženu ekspresiju u nekoliko uzoraka koji se odnose na proces SO. Prethodno je pokazano da je ekspresija *CeAGp3* indukovana tokom direktnog razvića korena u mraku i direktnog razvića izdanaka na svetlosti iz eksplantata listova kičice gajenih na hranljivoj podlozi bez regulatora rastenja, kao i u eksplantatima gajenim na hranljivoj podlozi sa 2,4-D i CPPU na svetlosti, gde se indirektna organogeneza i SE odigravaju simultano (Simonović i sar., 2015), dok je prema rezultatima predstavljenim u ovoj tezi ekspresija *CeAGp3* stabilna u svim uzorcima koji predstavljaju i SE i SO i nepromenjena u odnosu na rl (Slika 28). Pošto je eksperimentalni postupak korišćen za profilisanje ekspresije *AGP* u ovom radu (organogeni kalus i adventivni pupoljci su odvojeni od eksplantata radi RNK ekstrakcije) različit od onog opisanog u radu Simonović i sar. (2015) gde su korišćeni celi eksplantati listova sa razvijenim adventivnim pupoljcima i/ili somatskim embrionima, direktno poređenje rezultata nije moguće. Učešće AGP u procesu organogeneze je pokazano kod pšenice (Konieczny i sar., 2007), šećerne repe (Wisniewska i sar., 2007) i grejpfruta (Orbović i sar., 2013). Pošto većina ispitivanih *AGP* gena ima relativno stabilnu, ili čak sniženu ekspresiju u uzorcima uzetim tokom procesa organogeneze (Slika 28) verovatno je da izabrani skup od 18 gena jednostavno nije uključio gene sa značajnom ulogom tokom ovog procesa.

Povrede predstavljaju sastavni deo manipulacije biljnog tkiva u uslovima *in vitro* i mogu indukovati morfogenetske procese (Méndez-Hernández i sar., 2019; Santarem i sar., 1997; Bhatia i sar., 2005; Lup i sar., 2016) pa se može postaviti pitanje da li se poređenjem profila ekspresije gena nakon povrede (Slike 24 i 25) sa profilima u različitim uzorcima koji predstavljaju razvojne procese (Slika 28) mogu izabrati kandidati koji povezuju ova dva procesa, te bi zbog toga bili interesantni za dalje istraživanje. *CeAGp6* i *CeAGp7*, za koje je pokazano da su uključeni u odgovor na povrede lista, sa vrhuncem ekspresije 6 h nakon povreda (Slika 24), imali su smanjenu ekspresiju tokom organogeneze i SE (Slike 28) i teško se mogu smatrati vezom između ova dva procesa. Nasuprot njima, *CeFLA1*, koji je neznatno indukovan nakon povređivanja lista, sa značajnim povećanjem ekspresije posle 48 h od povrede tkiva (Slika 24), mogao bi biti uključen u procese koji se u biljci dešavaju nakon povređivanja, kao što je zaceljivanje. Pošto je *CeFLA1* indukovan u embriogenom kalusu (Slika 28), on potencijalno može biti deo mreže signala koja povezuje povrede i SE.

6. Zaključak

U skladu sa postavljenim ciljevima i iznetim rezultatima, može se zaključiti:

- Procenom performansi modela za previđanje verovatnoće hidroksilacije prolina na osnovu lokalne sekvence biljnih proteina, korišćenjem četiri algoritma MU: KNN, SVM, RF i XGB u kombinaciji sa tri tipa odabira atributa: mRMR, IGr i sfs, pokazano je da kombinacija sfs i XGB algoritama najbolje generalizuje.
- Trenirani modeli MU za previđanje verovatnoće hidroksilacije prolina predstavljaju centralni deo nove metodologije za identifikaciju i analizu HRGP i AGP sekvenci inkorporirane u R paket *ragp* koji je dostupan za korišćenje svima sa osnovnim znanjem R programskog jezika. Dalje unapređenje preciznosti modela MU za predviđanje pozicija Hyp u sekvencama biljnih proteina će neminovno unaprediti specifičnost i senzitivnost identifikovanih sekvenci.
- Razvijena metodologija je primenjena za identifikaciju HRGP i AGP sekvenci u velikom broju biljnih proteoma uključujući i proteom kičice. Identifikovan je veliki broj i heterogenost himernih AGP sekvenci, što potencijalno znači da O-glikozilacija na Hyp kod biljaka nije ograničena na manji skup prototipskih sekvenci. Identifikovani *HRGP* geni predstavljaju preduslov i resurs za buduća istraživanja uloga članova ove proteoglikanske familije
- Ispitivanjem uticaja rastućih koncentracija βGlcY u hranljivim podlogama na ekspresiju odabranih *AGP* gena, može se zaključiti da je dodatak βGlcY indukovao relativno male promene u ekspresiji ispitivanih gena, uglavnom bez statističke značajnosti, pa na osnovu dobijenih rezultata nije moguće doneti zaključke o potencijalnoj ulozi ispitivanih gena.
- Ispitivanjem ekspresije odabranih *AGP* gena nakon mehaničke povrede tkiva pokazano je da je *CeAGp6,* koji kodira 18 aminokiselina dug AG peptid, jako indukovan nakon povrede lista, a reprimiran nakon povrede korena. Bilo bi zanimljivo funkcionalno okarakterisati ovaj gen kako bi se utvrdila njegova uloga tokom odgovora biljnog tkiva na mehaničku povredu.
- FLA transkripti, *CeFLA1* i *CeFLA4* su najzastupljeniji među uzorcima iz prirode, posebno u stablu pa je moguće da doprinose čvrstini i elastičnosti stabla kao što je prethodno pokazano za njihove homologe iz *A.thaliana*.
- Nijedan od izabranih 18 gena čija je ekspresija ispitivana, nije pokazao značajno povećanje ekspresije u tkivima iz različitih faza SO u uslovima *in vitro*. Prema tome je verovatno da izabrani skup od 18 *AGP* gena nije uključio gene sa značajnom ulogom tokom ovog procesa.
- *CeFLA1,* koji je pokazao povećanje ekspresije tokom 48 h nakon povrede lista, a indukovan je i u embriogenom kalusu, mogao bi biti deo mreže koja povezuje povrede i somatsku embriogenezu.

7. Literatura

Aberham A, Pieri V, Croom EM, Ellmerer E, Stuppner H. 2011. Analysis of iridoids, secoiridoids and xanthones in *Centaurium erythraea*, *Frasera caroliniensis* and *Gentiana lutea* using LC-MS and RP-HPLC. J Pharm Biomed Anal 54(3):517-525.

Adadi A. 2021. A survey on data efficient algorithms in big data era. J Big Data 8(1):24.

Almagro Armenteros JJ, Tsirigos KD, Sønderby CK, Petersen TN, Winther O, Brunak S, von Heijne G, Nielsen H. 2019. SignalP 5.0 improves signal peptide predictions using deep neural networks. Nat Biotechnol 37(4):420-423.

Alonso JM, Stepanova AN, Leisse TJ, Kim CJ, Chen H, Shinn P, Stevenson DK, Zimmerman J, Barajas P, Cheuk R, Gadrinab C, Heller C, Jeske A, Koesema E, Meyers CC, Parker H, Prednis L, Ansari Y, Choy N, Deen H, Geralt M, Hazari N, Hom E, Karnes M, Mulholland C, Ndubaku R, Schmidt I, Guzman P, Aguilar-Henonin L, Schmid M, Weigel D, Carter DE, Marchand T, Risseeuw E, Brogden D, Zeko A, Crosby WL, Berry CC, Ecker JR. 2003. Genome-Wide Insertional Mutagenesis of *Arabidopsis thaliana*. Science 301(5633):653-657.

Amanchy R, Kandasamy K, Mathivanan S, Periaswamy B, Reddy R, Yoon WH, Joore J, Beer MA, Cope L, Pandey A. 2011. Identification of Novel Phosphorylation Motifs Through an Integrative Computational and Experimental Analysis of the Human Phosphoproteome. J Proteomics Bioinform 4(2):22-35.

Andersen CL, Jensen JL, Ørntoft TF. 2004. Normalization of Real-Time Quantitative Reverse Transcription-PCR Data: A Model-Based Variance Estimation Approach to Identify Genes Suited for Normalization, Applied to Bladder and Colon Cancer Data Sets. Cancer Res 64(15):5245-50.

Atchley WR, Zhao J, Fernandes AD, Drüke T. 2005. Solving the protein sequence metric problem. Proc Natl Acad Sci USA 102(18):6395-6400.

Baek M, DiMaio F, Anishchenko I, Dauparas J, Ovchinnikov S, Lee GR, Wang J, Cong Q, Kinch LN, Schaeffer RD, Millán C, Park H, Adams C, Glassman CR, DeGiovanni A, Pereira JH, Rodrigues AV, van Dijk AA, Ebrecht AC, Opperman DJ, Sagmeister T, Buhlheller C, Pavkov-Keller T, Rathinaswamy MK, Dalwadi U, Yip CK, Burke JE, Garcia KC, Grishin NV, Adams PD, Read RJ, Baker D. 2021. Accurate prediction of protein structures and interactions using a three-track neural network. Science 373(6557):871-876.

Bai L, Zhang G, Zhou Y, Zhang Z, Wang W, Du Y, Wu Z, Song CP. 2009. Plasma membrane-associated proline-rich extensin-like receptor kinase 4, a novel regulator of Ca signalling, is required for abscisic acid responses in *Arabidopsis thaliana*. Plant J 60(2):314-327.

Baldi P, Brunak S. 2001. Bioinformatics: The Machine Learning Approach, 2nd edition. MIT Press, Cambridge, Massachusetts, pp.452

Battaglia M, Solórzano RM, Hernández M, Cuéllar-Ortiz S, García-Gómez B, Márquez J, Covarrubias AA. 2007. Proline-rich cell wall proteins accumulate in growing regions and phloem tissue in response to water deficit in common bean seedlings. Planta 225(5):1121-1133.

Baumberger N, Ringli C, Keller B. 2001. The chimeric leucine-rich repeat/extensin cell wall protein LRX1 is required for root hair morphogenesis in *Arabidopsis thaliana*. Genes Dev 15(9):1128-1139.

Benjamini Y, Hochberg Y. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J R Stat Soc Series B Stat Methodol 57(1):289-300.

Bhatia P, Ashwath N, Midmore DJ. 2005. Effects of genotype, explant orientation, and wounding on shoot regeneration in tomato. In Vitro Cell Dev Biol Plant 41(4):457-464.

Biecek P. 2018. Dalex: Explainers for complex predictive models in R. J Mach Learn Res 19:1−5.

Bischl B, Richter J, Bossek J, Horn D, Thomas J, Lang M. 2017. mlrMBO: A Modular Framework for Model-Based Optimization of Expensive Black-Box Functions.

Bischl B, Lang M, Kotthoff L, Schiffner J, Richter J, Studerus E, Casalicchio G, Jones ZM. 2016. mlr: Machine learning in R. J Mach Learn Res 17:1–5.

Bogdanović MD, Ćuković KB, Subotić AR, Dragićević MB, Simonović AD, Filipović BK, Todorović SI. 2021. Secondary Somatic Embryogenesis in *Centaurium erythraea* Rafn. Plants 10(2):199.

Borassi C, Gloazzo Dorosz J, Ricardi MM, Carignani Sardoy M, Pol Fachin L, Marzol E, Mangano S, Rodríguez Garcia DR, Martínez Pacheco J, Rondón Guerrero YDC, Velasquez SM, Villavicencio B, Ciancia M, Seifert G, Verli H, Estevez JM. 2020. A cell surface arabinogalactan-peptide influences root hair cell fate. New Phytol 227(3):732-743.

Borassi C, Sede AR, Mecchia MA, Salgado Salter JD, Marzol E, Muschietti JP, Estevez JM. 2016. An update on cell surface proteins containing extensin-motifs. J Exp Bot 67(2):477-487.

Breiman L. 2001. Random Forests. Mach Learn 45(1):5-32.

Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, Agarwal S, Herbert-Voss A, Krueger G, Henighan TJ, Child R, Ramesh A, Ziegler DM, Wu J, Winter C, Hesse C, Chen M, Sigler E, Litwin M, Gray S, Chess B, Clark J, Berner C, McCandlish S, Radford A, Sutskever I, Amodei D. 2020. Language Models are Few-Shot Learners. CoRR

Cannon MC, Terneus K, Hall Q, Tan L, Wang Y, Wegenhart BL, Chen L, Lamport DT, Chen Y, Kieliszewski MJ. 2008. Self-assembly of the plant cell wall requires an extensin scaffold. Proc Natl Acad Sci USA 105(6):2226-2231.

Canut H, Albenne C, Jamet E. 2016. Post-translational modifications of plant cell wall proteins and peptides: A survey from a proteomics point of view. Biochim Biophys Acta Proteins Proteom 1864(8):983-990.

Castilleux R, Plancot B, Gügi B, Attard A, Loutelier-Bourhis C, Lefranc B, Nguema-Ona E, Arkoun M, Yvin J-C, Driouich A, Vicré M. 2019. Extensin arabinosylation is involved in root response to elicitors and limits oomycete colonization. Ann Bot 125(5):751-763.

Castilleux R, Plancot B, Vicré M, Nguema-Ona E, Driouich A. 2021. Extensin, an underestimated key component of cell wall defence? Ann Bot 127(6):709-713.

Cawley GC, Talbot NLC. 2010. On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation. J Mach Learn Res 11:2079-2107.

Chalkia D, Nikolaidis N, Makalowski W, Klein J, Nei M. 2008. Origins and evolution of the formin multigene family that is involved in the formation of actin filaments. Mol Biol Evol 25(12):2717-2733.

Chen L, Zhong HY, Kuang JF, Li JG, Lu WJ, Chen JY. 2011. Validation of reference genes for RT-qPCR studies of gene expression in banana fruit under different experimental conditions. Planta 234(2):377-390.

Chen M, Liu Q, Chen S, Liu Y, Zhang C, Liu R. 2019. XGBoost-Based Algorithm Interpretation and Application on Post-Fault Transient Stability Status Prediction of Power System. IEEE Access 7:13149-13158.

Chen T, Guestrin C. 2016. XGBoost: A Scalable Tree Boosting System. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Association for Computing Machinery, New York, NY, USA. p785–794.

Chen X-Y, Kim J-Y. 2009. Callose synthesis in higher plants. Plant Signal Behav 4(6):489-492.

Chen Z. 2001. A superfamily of proteins with novel cysteine-rich repeats. Plant Physiol 126(2):473-476.

Cheung AY, Niroomand S, Zou Y, Wu H-M. 2010. A transmembrane formin nucleates subapical actin assembly and controls tip-focused growth in pollen tubes. Proc Natl Acad Sci 107(37):16390-16395.

Chollet F, Allaire JJ. 2018. Deep learning with R. Shelter Island, NY USA: Manning Publications Company

Chou KC. 2000. Prediction of protein subcellular locations by incorporating quasi-sequence-order effect. Biochem Biophys Res Commun 278(2):477-483.

Chou KC. 2001. Prediction of protein cellular attributes using pseudo-amino acid composition. Proteins 43(3):246-255.

Chou KC. 2005. Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. Bioinformatics 21(1):10-19.

Claesen M, De Moor B. 2015. Hyperparameter Search in Machine Learning. Proceedings of the 11th Metaheuristics International Conference.

Cohen J. 1960. A Coefficient of Agreement for Nominal Scales. Educ Psychol Meas 20:37-46.

Coimbra S, Costa M, Jones B, Mendes MA, Pereira LG. 2009. Pollen grain development is compromised in *Arabidopsis agp6 agp11* null mutants. J Exp Bot 60(11):3133-3142.

Cortes C, Vapnik V. 1995. Support-vector networks. Mach Learn 20(3):273-297.

Costa M, Pereira AM, Pinto SC, Silva J, Pereira LG, Coimbra S. 2019. In silico and expression analyses of fasciclin-like arabinogalactan proteins reveal functional conservation during embryo and seed development. Plant Reprod 32(4):353-370.

Ćuković K, Dragićević M, Bogdanović M, Paunović D, Giurato G, Filipović B, Subotić A, Todorović S, Simonović A. 2020. Plant regeneration in leaf culture of *Centaurium*

erythraea Rafn. Part 3: *de novo* transcriptome assembly and validation of housekeeping genes for studies of *in vitro* morphogenesis. Plant Cell, Tissue Organ Cult 141(2):417-433.

De Vega-Bartol JJ, Santos R, Simões M, Miguel C. 2011. Evaluation of reference genes for quantitative PCR analysis during somatic embryogenesis in conifers. BMC Proc 32(5):715-29

Deepak S, Shailasree S, Kini RK, Muck A, Mithöfer A, Shetty SH. 2010. Hydroxyproline-rich Glycoproteins and Plant Defence. J Phytopathol 158(9):585-593.

Del Vento D, Fanfarillo A. 2019. Traps, Pitfalls and Misconceptions of Machine Learning applied to Scientific Disciplines. In: Proceedings of the Practice and Experience in Advanced Research Computing on Rise of the Machines (learning). Chicago, IL, USA: Association for Computing Machinery. p1-8.

Desnoyer N, Palanivelu R. 2020. Bridging the GAPs in plant reproduction: a comparison of plant and animal GPI-anchored proteins. Plant Reprod 33(3-4), 129-142.

Dilokpimol A, Geshi N. 2014. *Arabidopsis thaliana* glucuronosyltransferase in family GT14. Plant Signal Behav 9(6):28891.

Doll S, Burlingame AL. 2015. Mass Spectrometry-Based Detection and Assignment of Protein Posttranslational Modifications. ACS Chem Biol 10(1):63-71.

Dragićević MB, Paunović DM, Bogdanović MD, Todorović SI, Simonović AD. 2020. ragp: Pipeline for mining of plant hydroxyproline-rich glycoproteins with implementation in R. Glycobiology 30(1):19-35.

Dragićević MB, Simonović AD. 2021. Arabinogalactan protein mining and diversity - the case of *Centaurium erythraea*. Biol Serb 43(1):4–11.

Drost H-G, Paszkowski J. 2017. Biomartr: genomic data retrieval with R. Bioinformatics 33(8):1216-1217.

Dubchak I, Muchnik I, Holbrook SR, Kim SH. 1995. Prediction of protein folding class using global description of amino acid sequence. Proc Natl Acad Sci USA 92(19):8700-8704.

Duclercq J, Sangwan-Norreel B, Catterou M, Sangwan RS. 2011. *De novo* shoot organogenesis: from art to science. Trends Plant Sci 16(11):597-606.

Duruflé H, Hervé V, Balliau T, Zivy M, Dunand C, Jamet E. 2017. Proline Hydroxylation in Cell Wall Proteins: Is It Yet Possible to Define Rules? Front Plant Sci 8:1802.

Duvenaud D, Maclaurin D, Aguilera-Iparraguirre J, Gómez-Bombarelli R, Hirzel T, Aspuru-Guzik A, Adams R. 2015. Convolutional Networks on Graphs for Learning Molecular Fingerprints. In the Proceedings of Adv Neural Inf Process Syst 28 (NIPS 2015), Montreal, Canada. 2215-2223.

Dvoráková L, Srba M, Opatrny Z, Fischer L. 2012. Hybrid proline-rich proteins: novel players in plant cell elongation? Ann Bot 109(2):453-462.

Eddy SR. 2011. Accelerated Profile HMM Searches. PLoS Comput Biol 7(10):e1002195.

Eisenhaber B, Wildpaner M, Schultz CJ, Borner GH, Dupree P, Eisenhaber F. 2003. Glycosylphosphatidylinositol lipid anchoring of plant proteins. Sensitive prediction from

sequence- and genome-wide studies for Arabidopsis and rice. Plant Physiol 133(4):1691-1701.

Ellis M, Egelund J, Schultz CJ, Bacic A. 2010. Arabinogalactan-Proteins: Key Regulators at the Cell Surface? Plant Physiol 153(2):403-419.

Elnaggar A, Heinzinger M, Dallago C, Rihawi G, Wang Y, Jones L, Gibbs T, Feher T, Angerer C, Bhowmik D, Rost B. 2020. ProtTrans: Towards Cracking the Language of Life's Code Through Self-Supervised Deep Learning and High Performance Computing. IEEE Trans Pattern Anal Mach Intell PP

Emanuelsson O, Brunak S, von Heijne G, Nielsen H. 2007. Locating proteins in the cell using TargetP, SignalP and related tools. Nat Protoc 2(4):953-971.

Faye L, Boulaflous A, Benchabane M, Gomord V, Michaud D. 2005. Protein modifications in the plant secretory pathway: current status and practical implications in molecular pharming. Vaccine 23(15):1770-1778.

Filipović BK, Simonović AD, Trifunović MM, Dmitrović SS, Savić JM, Jevremović SB, Subotić AR. 2015. Plant regeneration in leaf culture of *Centaurium erythraea* Rafn. Part 1: The role of antioxidant enzymes. Plant Cell, Tissue Organ Cult 121(3):703-719.

Filipović BK, Trifunović-Momčilov MM, Simonović AD, Jevremović SB, Milošević SM, Subotić AR. 2021. Immunolocalization of some arabinogalactan protein epitopes during indirect somatic embryogenesis and shoot organogenesis in leaf culture of centaury (*Centaurium erythraea* Rafn). In Vitro Cell Dev Biol Plant 57:470-480.

Finn RD, Clements J, Eddy SR. 2011. HMMER web server: interactive sequence similarity searching. Nucleic Acids Res 39:W29-W37.

Fix E, Hodges JL. 1951. Discriminatory analysis-nonparametric discrimination: consistency properties. (Report). USAF School of Aviation Medicine, Randolph Field, Texas

Fragkostefanakis S, Dandachi F, Kalaitzis P. 2012. Expression of arabinogalactan proteins during tomato fruit ripening and in response to mechanical wounding, hypoxia and anoxia. Plant Physiol Biochem 52:112-118.

Friedman JH. 2001. Greedy function approximation: A gradient boosting machine. Ann Stat 29(5):1189-1232.

Galian C, Björkholm P, Bulleid N, von Heijne G. 2012. Efficient Glycosylphosphatidylinositol (GPI) Modification of Membrane Proteins Requires a C-terminal Anchoring Signal of Marginal Hydrophobicity*. J Biol Chem 287(20):16399-16409.

Gao M, Showalter AM. 1999. Yariv reagent treatment induces programmed cell death in Arabidopsis cell cultures and implicates arabinogalactan protein involvement. Plant J 19(3):321-331.

Gašić K, Hernandez A, Korban SS. 2004. RNA extraction from different apple tissues rich in polyphenols and polysaccharides for cDNA library construction. Plant Mol Biol Rep 22(4):437-438.

Gaspar Y, Johnson K, McKenna J, Bacic A, Schultz C. 2001. The complex structures of arabinogalactan-proteins and the journey towards understanding function. Plant Mol Biol 47:161-176.

Gaspar YM, Nam J, Schultz CJ, Lee LY, Gilson PR, Gelvin SB, Bacic A. 2004. Characterization of the Arabidopsis lysine-rich arabinogalactan-protein *AtAGP17* mutant (*rat1*) that results in a decreased efficiency of agrobacterium transformation. Plant Physiol 135(4):2162-71.

Geshi N, Johansen JN, Dilokpimol A, Rolland A, Belcram K, Verger S, Kotake T, Tsumuraya Y, Kaneko S, Tryfona T, Dupree P, Scheller HV, Höfte H, Mouille G. 2013. A galactosyltransferase acting on arabinogalactan protein glycans is essential for embryo development in Arabidopsis. Plant J 76(1):128-137.

Gilson P, Gaspar YM, Oxley D, Youl JJ, Bacic A. 2001. NaAGP4 is an arabinogalactan protein whose expression is suppressed by wounding and fungal infection in *Nicotiana alata*. Protoplasma 215(1-4):128-139.

Gíslason MH, Nielsen H, Almagro Armenteros JJ, Johansen AR. 2021. Prediction of GPI-anchored proteins with pointer neural networks. Curr Res Biotechnol 3:6-13.

Goodwin S, McPherson JD, McCombie WR. 2016. Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet 17(6):333-351.

Gorres KL, Raines RT. 2010. Prolyl 4-hydroxylase. Crit Rev Biochem Mol Biol 45(2):106-124.

Grantham R. 1974. Amino acid difference formula to help explain protein evolution. Science 185(4154):862-864.

Grigorescu S, Trasnea B, Cocias T, Macesanu G. 2020. A survey of deep learning techniques for autonomous driving. J Field Robot 37(3):362-386.

Gruenheit N, Deusch O, Esser C, Becker M, Voelckel C, Lockhart P. 2012. Cutoffs and k-mers: implications from a transcriptome study in allopolyploid plants. BMC Genom 13(1):92.

Guan Y, Nothnagel EA. 2004. Binding of Arabinogalactan Proteins by Yariv Phenylglycoside Triggers Wound-Like Responses in Arabidopsis Cell Cultures. Plant Physiol 135(3):1346-1366.

Hamza N, Berke B, Cheze C, Agli A-N, Robinson P, Gin H, Moore N. 2010. Prevention of type 2 diabetes induced by high fat diet in the C57BL/6J mouse by two medicinal plants used in traditional treatment of diabetes in the east of Algeria. J Ethnopharmacol 128(2):513-518.

He J, Zhao H, Cheng Z, Ke Y, Liu J, Ma H. 2019. Evolution Analysis of the Fasciclin-Like Arabinogalactan Proteins in Plants Shows Variable Fasciclin-AGP Domain Constitutions. Int J Mol Sci 20(8):1945.

Held MA, Tan L, Kamyab A, Hare M, Shpak E, Kieliszewski MJ. 2004. Di-isodityrosine is the intermolecular cross-link of isodityrosine-rich extensin analogs cross-linked *in vitro*. J Biol Chem 279(53):55474-55482.

Hervé C, Siméon A, Jam M, Cassin A, Johnson KL, Salmeán AA, Willats WG, Doblin MS, Bacic A, Kloareg B. 2016. Arabinogalactan proteins have deep roots in eukaryotes: identification of genes and epitopes in brown algae and their role in *Fucus serratus* embryo development. New Phytol 209(4):1428-1441.

Hijazi M, Durand J, Pichereaux C, Pont F, Jamet E, Albenne C. 2012. Characterization of the arabinogalactan protein 31 (AGP31) of *Arabidopsis thaliana*: new advances on the Hyp-O-glycosylation of the Pro-rich domain. J Biol Chem 287(12):9623-9632.

Hijazi M, Velasquez SM, Jamet E, Estevez JM, Albenne C. 2014. An update on post-translational modifications of hydroxyproline-rich glycoproteins: toward a model highlighting their contribution to plant cell wall architecture. Front Plant Sci 5:395.

Hu Y, Qin Y, Zhao J. 2006. Localization of an arabinogalactan protein epitope and the effects of Yariv phenylglycoside during zygotic embryo development of *Arabidopsis thaliana*. Protoplasma 229(1):21-31.

Huang H, Miao Y, Zhang Y, Huang L, Cao J, Lin S. 2021. Comprehensive Analysis of Arabinogalactan Protein-Encoding Genes Reveals the Involvement of Three *BrFLA* Genes in Pollen Germination in *Brassica rapa*. Int J Mol Sci 22(23):13142.

Huang J, Kim CM, Xuan YH, Liu J, Kim TH, Kim BK, Han CD. 2013. Formin homology 1 (OsFH1) regulates root-hair elongation in rice (*Oryza sativa*). Planta 237(5):1227-1239.

Hwang Y, Lee H, Lee Y-S, Cho H-T. 2016. Cell wall-associated ROOT HAIR SPECIFIC 10, a proline-rich receptor-like kinase, is a negative modulator of Arabidopsis root hair growth. J Exp Bot 67(6):2007-2022.

Ismail HD, Newman RH, Kc DB. 2016. RF-Hydroxysite: a random forest based predictor for hydroxylation sites. Mol Biosyst 12(8):2427-2435.

James G, Witten D, Hastie T, Tibshirani R. 2017. An Introduction to Statistical Learning with Applications in R. 7nd edition, Springer, New York

Johnson KL, Cassin AM, Lonsdale A, Bacic A, Doblin MS, Schultz CJ. 2017a. A motif and amino acid bias bioinformatics pipeline to identify hydroxyproline-rich glycoproteins. Plant Physiol 174(2):886-903.

Johnson KL, Cassin AM, Lonsdale A, Wong GK-S, Soltis DE, Miles NW, Melkonian M, Melkonian B, Deyholos MK, Leebens-Mack J, Rothfels CJ, Stevenson DW, Graham SW, Wang X, Wu S, Pires JC, Edger PP, Carpenter EJ, Bacic A, Doblin MS, Schultz CJ. 2017b. Insights into the Evolution of Hydroxyproline-Rich Glycoproteins from 1000 Plant Transcriptomes. Plant Physiol 174(2):886-903.

Johnson KL, Jones BJ, Bacic A, Schultz CJ. 2003. The fasciclin-like arabinogalactan proteins of Arabidopsis. A multigene family of putative cell adhesion molecules. Plant Physiol 133(4):1911-1925.

Jones DT, Taylor WR, Thornton JM. 1992. The rapid generation of mutation data matrices from protein sequences. Comput Appl Biosci 8(3):275-282.

Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, Bridgland A, Meyer C, Kohl SAA, Ballard AJ, Cowie A, Romera-Paredes B, Nikolov S, Jain R, Adler J, Back T, Petersen S, Reiman D, Clancy E, Zielinski M, Steinegger M, Pacholska M, Berghammer T, Bodenstein S, Silver D, Vinyals O, Senior AW, Kavukcuoglu K, Kohli P, Hassabis D. 2021. Highly accurate protein structure prediction with AlphaFold. Nature 596(7873):583-589.

Jung I, Matsuyama A, Yoshida M, Kim D. 2010. PostMod: sequence based prediction of kinase-specific phosphorylation sites with indirect relationship. BMC Bioinform 11:S10.

Käll L, Krogh A, Sonnhammer EL. 2007. Advantages of combined transmembrane topology and signal peptide prediction--the Phobius web server. Nucleic Acids Res 35:429-432.

Kanakachari M, Solanke AU, Prabhakaran N, Ahmad I, Dhandapani G, Jayabalan N, Kumar PA. 2016. Evaluation of Suitable Reference Genes for Normalization of qPCR Gene Expression Studies in Brinjal (*Solanum melongena* L.) During Fruit Developmental Stages. Appl Biochem Biotechnol 178(3):433-450.

Karami O, Aghavaisi B, Mahmoudi Pour A. 2009. Molecular aspects of somatic-to-embryogenic transition in plants. J Chem Biol 2(4):177-190.

Kawamoto T, Noshiro M, Shen M, Nakamasu K, Hashimoto K, Kawashima-Ohya Y, Gotoh O, Kato Y. 1998. Structural and phylogenetic analyses of RGD-CAP/beta ig-h3, a fasciclin-like adhesion protein expressed in chick chondrocytes. Biochim Biophys Acta 1395(3):288-292.

Kawashima S, Kanehisa M. 2000. AAindex: amino acid index database. Nucleic Acids Res 28(1):374.

Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q, Liu T-Y. 2017. LightGBM: a highly efficient gradient boosting decision tree. In: Proceedings of the 31st International Conference on Neural Information Processing Systems Curran Associates Inc., Red Hook, NY, USA, 3149–3157

Kerkvliet J, de Fouchier A, van Wijk M, Groot AT. 2019. The Bellerophon pipeline, improving *de novo* transcriptomes and removing chimeras. Ecol Evol 9(18):10513-10521.

Kieliszewski MJ, Lamport DT. 1994. Extensin: repetitive motifs, functional sites, post-translational codes, and phylogeny. Plant J 5(2):157-172.

Kieliszewski MJ, Lamport DTA, Tan L, Cannon MC. 2010. Hydroxyproline-Rich Glycoproteins: Form and Function. Annu Plant Rev 41:321-342.

Kim JE, Kim SJ, Lee BH, Park RW, Kim KS, Kim IS. 2000. Identification of motifs for cell adhesion within the repeated domains of transforming growth factor-beta-induced gene, betaig-h3. J Biol Chem 275(40):30907-30915.

Kinoshita T, Fujita M. 2016. Thematic Review Series: Glycosylphosphatidylinositol (GPI) Anchors: Biochemistry and Cell Biology Biosynthesis of GPI-anchored proteins: special emphasis on GPI lipid remodeling. J Lipid Res 57(1):6-24.

Kitazawa K, Tryfona T, Yoshimi Y, Hayashi Y, Kawauchi S, Antonov L, Tanaka H, Takahashi T, Kaneko S, Dupree P, Tsumuraya Y, Kotake T. 2013. β-Galactosyl Yariv Reagent Binds to the β-1,3-Galactan of Arabinogalactan Proteins. Plant Physiol 161(3):1117-1126.

Kjellbom P, Snogerup L, Stöhr C, Reuzeau C, McCabe PF, Pennell RI. 1997. Oxidative cross-linking of plasma membrane arabinogalactan proteins. Plant J 12(5):1189-1196.

Knoch E, Dilokpimol A, Tryfona T, Poulsen CP, Xiong G, Harholt J, Petersen BL, Ulvskov P, Hadi MZ, Kotake T, Tsumuraya Y, Pauly M, Dupree P, Geshi N. 2013. A β-glucuronosyltransferase from *Arabidopsis thaliana* involved in biosynthesis of type II

arabinogalactan has a role in cell elongation during seedling growth. Plant J 76(6):1016-1029.

Kobayashi Y, Motose H, Iwamoto K, Fukuda H. 2011. Expression and Genome-Wide Analysis of the Xylogen-Type Gene Family. Plant Cell Physiol 52(6):1095-1106.

Kobe B, Deisenhofer J. 1994. The leucine-rich repeat: a versatile binding motif. Trends Biochem Sci 19(10):415-421.

Kohavi R. 1995. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. In Proceedings of the 14th International Joint Conference on Artificial Intelligence Montreal, Canada 2:1137–1143

Kohavi R, John GH. 1997. Wrappers for feature subset selection. Artif Intell 97(1):273-324.

Konieczny R, Swierczyńska J, Czaplicki AZ, Bohdanowicz J. 2007. Distribution of pectin and arabinogalactan protein epitopes during organogenesis from androgenic callus of wheat. Plant Cell Rep 26(3):355-363.

Kotch FW, Guzei IA, Raines RT. 2008. Stabilization of the Collagen Triple Helix by O-Methylation of Hydroxyproline Residues. J Am Chem Soc 130(10):2952-2953.

Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. 2015. Machine learning applications in cancer prognosis and prediction. Comput Struct Biotechnol J 13:8-17.

Kumar G, Singh A. 2015. Reference gene validation for qRT-PCR based gene expression studies in different developmental stages and under biotic stress in apple. Sci Hortic 197:597-606.

Lamport DT, Kieliszewski MJ, Showalter AM. 2006. Salt stress upregulates periplasmic arabinogalactan proteins: using salt stress to analyse AGP function. New Phytol 169(3):479-92.

Lamport D, Northcote D. 1960. Hydroxyproline in primary cell walls of higher plants. Nature 188:665-666.

Larrañaga P, Calvo B, Santana R, Bielza C, Galdiano J, Inza I, Lazano JA, Armañanzas R, Santafe G, Pérez A, Robles V. 2006. Machine learning in bioinformatics. Brief Bioinformatics 7:86-112.

Lee CB, Swatek KN, McClure B. 2008. Pollen proteins bind to the C-terminal domain of *Nicotiana alata* pistil arabinogalactan proteins. J Biol Chem 283(40):26965-26973.

Lee KJD, Sakata Y, Mau S-L, Pettolino F, Bacic A, Quatrano RS, Knight CD, Knox JP. 2005. Arabinogalactan proteins are required for apical cell extension in the moss *Physcomitrella patens*. Plant Cell 17(11):3051-3065.

León J, Rojo E, Sánchez-Serrano JJ. 2001. Wound signalling in plants. J Exp Bot 52(354):1-9.

Li J, Yu M, Geng LL, Zhao J. 2010. The fasciclin-like arabinogalactan protein gene, *FLA3*, is involved in microspore development of Arabidopsis. Plant J 64(3):482-497.

Libbrecht MW, Noble WS. 2015. Machine learning applications in genetics and genomics. Nat Rev Genet 16(6):321-332.

Liu C, Mehdy MC. 2007. A Nonclassical Arabinogalactan Protein Gene Highly Expressed in Vascular Tissues, *AGP31*, Is Transcriptionally Repressed by Methyl Jasmonic Acid in Arabidopsis. Plant Physiol 145(3):863-874.

Liu X, Wolfe R, Welch LR, Domozych DS, Popper ZA, Showalter AM. 2016. Bioinformatic Identification and Analysis of Extensins in the Plant Kingdom. PLoS One 11(2): e0150177.

Livak KJ, Schmittgen TD. 2001. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. Methods 25(4):402-408.

Løvdal T, Lillo C. 2009. Reference gene selection for quantitative real-time PCR normalization in tomato subjected to nitrogen, cold, and light stress. Anal Biochem 387:238-242.

Lu S, Wang J, Chitsaz F, Derbyshire MK, Geer RC, Gonzales NR, Gwadz M, Hurwitz DI, Marchler GH, Song JS, Thanki N, Yamashita RA, Yang M, Zhang D, Zheng C, Lanczycki CJ, Marchler-Bauer A. 2019. CDD/SPARCLE: the conserved domain database in 2020. Nucleic Acids Res 48(D1):D265-D268.

Lup SD, Tian X, Xu J, Pérez-Pérez JM. 2016. Wound signaling of regenerative cell reprogramming. Plant Sci 250:178-187.

Ma H, Zhao H, Liu Z, Zhao J. 2011. The phytocyanin gene family in rice (*Oryza sativa* L.): genome-wide identification, classification and transcriptional analysis. PLoS One 6(10):e25184.

Ma H, Zhao J. 2010. Genome-wide identification, classification, and expression analysis of the arabinogalactan protein gene family in rice (*Oryza sativa* L.). J Exp Bot 61(10):2647-2668.

Ma Y, Yan C, Li H, Wu W, Liu Y, Wang Y, Chen Q, Ma H. 2017. Bioinformatics Prediction and Evolution Analysis of Arabinogalactan Proteins in the Plant Kingdom. Front Plant Sci 8(66).

MacMillan CP, Mansfield SD, Stachurski ZH, Evans R, Southerton SG. 2010. Fasciclin-like arabinogalactan proteins: specialization for stem biomechanics and cell wall architecture in Arabidopsis and Eucalyptus. Plant J 62(4):689-703.

Majewska-Sawka A, Nothnagel EA. 2000. The Multiple Roles of Arabinogalactan Proteins in Plant Development. Plant Physiol 122(1):3-10.

Malkov SN, Simonović AD. 2011. Shotgun assembly of *Centaurium erythraea* transcriptome. 19th Symposium of the Serbian plant physiology society, Book of abstracts:16.

Mashiguchi K, Urakami E, Hasegawa M, Sanmiya K, Matsumoto I, Yamaguchi I, Asami T, Suzuki Y. 2008. Defense-Related Signaling by Interaction of Arabinogalactan Proteins and β-Glucosyl Yariv Reagent Inhibits Gibberellin Signaling in Barley Aleurone Cells. Plant Cell Physiol 49(2):178-190.

Mashiguchi K, Yamaguchi I, Suzuki Y. 2004. Isolation and Identification of Glycosylphosphatidylinositol-Anchored Arabinogalactan Proteins and Novel β-Glucosyl Yariv-Reactive Proteins from Seeds of Rice (*Oryza sativa*). Plant Cell Physiol 45(12):1817-1829.

Matthew L. 2004. RNAi for Plant Functional Genomics. Comp Funct Genomics 5:941402.

Matthews BW. 1975. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim Biophys Acta 405(2):442-451.

Maurer J, Pereira-Netto A, Pettolino F, Gaspar Y, Bacic A. 2010. Effects of Yariv dyes, arabinogalactan-protein binding reagents, on the growth and viability of Brazilian pine suspension culture cells. Trees - Struct Funct 24:391-398.

Méndez-Hernández HA, Ledezma-Rodríguez M, Avilez-Montalvo RN, Juárez-Gómez YL, Skeete A, Avilez-Montalvo J, De-la-Peña C, Loyola-Vargas VM. 2019. Signaling Overview of Plant Somatic Embryogenesis. Front Plant Sci 10(77).

Meng J, Hu B, Yi G, Li X, Chen H, Wang Y, Yuan W, Xing Y, Sheng Q, Su Z, Xu C. 2020. Genome-wide analyses of banana fasciclin-like AGP genes and their differential expression under low-temperature stress in chilling sensitive and tolerant cultivars. Plant Cell Reports 39(6):693-708.

Meyer D. 2018. Support Vector Machines. https://cran.rproject.org/web/packages/e1071/vignettes/svmdoc.pdf

Moller I, Marcus SE, Haeger A, Verhertbruggen Y, Verhoef R, Schols H, Ulvskov P, Mikkelsen JD, Knox JP, Willats W. 2008. High-throughput screening of monoclonal antibodies against plant cell wall glycans by hierarchical clustering of their carbohydrate microarray binding profiles. Glycoconj J 25(1):37-48.

Motose H, Sugiyama M, Fukuda H. 2004. A proteoglycan mediates inductive interaction during plant vascular development. Nature 429(6994):873-878.

Mroueh M, Saab Y, Rizkallah R. 2004. Hepatoprotective activity of *Centaurium erythraea* on acetaminophen-induced hepatotoxicity in rats. Phytother Res 18(5):431-433.

Murashige T, Skoog F. 1962. A Revised Medium for Rapid Growth and Bio Assays with Tobacco Tissue Cultures. Physiol Plant 15(3):473-497.

Nakhamchik A, Zhao Z, Provart NJ, Shiu S-H, Keatley SK, Cameron RK, Goring DR. 2004. A Comprehensive Expression Analysis of the Arabidopsis Proline-rich Extensin-like Receptor Kinase Gene Family using Bioinformatic and Experimental Approaches. Plant Cell Physiol 45(12):1875-1881.

Nikolić M, Zečević A. 2019. Mašinsko učenje, Matematički fakultet Univerziteta u Beogradu

Nguema-Ona E, Vicré-Gibouin M, Cannesan MA, Driouich A. 2013. Arabinogalactan proteins in root-microbe interactions. Trends Plant Sci 18(8):440-449.

Ogawa-Ohnishi M, Matsubayashi Y. 2015. Identification of three potent hydroxyproline O-galactosyltransferases in Arabidopsis. The Plant Journal 81(5):736-746.

Ogawa-Ohnishi M, Matsushita W, Matsubayashi Y. 2013. Identification of three hydroxyproline O-arabinosyltransferases in *Arabidopsis thaliana*. Nat Chem Biol 9(11):726-730.

Oka T, Saito F, Shimma Y-i, Yoko-o T, Nomura Y, Matsuoka K, Jigami Y. 2009. Characterization of Endoplasmic Reticulum-Localized UDP-d-Galactose: Hydroxyproline O-Galactosyltransferase Using Synthetic Peptide Substrates in Arabidopsis. Plant Physiol 152(1):332-340.

Ono H, Ishii K, Kozaki T, Ogiwara I, Kanekatsu M, Yamada T. 2015. Removal of redundant contigs from de novo RNA-Seq assemblies via homology search improves accurate detection of differentially expressed genes. BMC genom 16:1031-1031.

Orbović V, Göllner EM, Soria P. 2013. The effect of arabinogalactan proteins on regeneration potential of juvenile citrus explants used for genetic transformation by *Agrobacterium tumefaciens*. Acta Physiol Plant 35(5):1409-1419.

Pardy C. 2020. mpmi: Mixed-Pair Mutual Information Estimators. R package version 0.43.1. https://CRAN.R-project.org/package=mpmi

Parker CE, Mocanu V, Mocanu M, Dicheva N, Warren MR. 2010. Mass Spectrometry for Post-Translational Modifications. In: Alzate O, editor. Neuroproteomics. Boca Raton (FL): CRC Press/Taylor & Francis Chapter 6

Paulsen BS, Craik DJ, Dunstan DE, Stone BA, Bacic A. 2014. The Yariv reagent: behaviour in different solvents and interaction with a gum arabic arabinogalactan-protein. Carbohydr Polym 106:460-468.

Pearson, K (1895). Notes on regression and inheritance in the case of two parents. Proc R Soc Lond 58: 240–242.

Peng H, Long F, Ding C. 2005. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans Pattern Anal Mach Intell 27(8):1226-1238.

Pereira AM, Pereira LG, Coimbra S. 2015. Arabinogalactan proteins: rising attention from plant biologists. Plant Reprod 28(1):1-15.

Pérez-Pérez Y, Carneros E, Berenguer E, Solís M-T, Bárány I, Pintos B, Gómez-Garay A, Risueño MC, Testillano PS. 2019. Pectin De-methylesterification and AGP Increase Promote Cell Wall Remodeling and Are Required During Somatic Embryogenesis of *Quercus suber*. Front Plant Sci 9:1915.

Perkins JR, Dawes JM, McMahon SB, Bennett DL, Orengo C, Kohl M. 2012. ReadqPCR and NormqPCR: R packages for the reading, quality checking and normalisation of RT-qPCR quantification cycle (Cq) data. BMC Genom 13:296.

Petersen TN, Brunak S, von Heijne G, Nielsen H. 2011. SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat Methods 8(10):785-786.

Pfeifer L, Shafee T, Johnson KL, Bacic A, Classen B. 2020. Arabinogalactan-proteins of *Zostera marina* L. contain unique glycan structures and provide insight into adaption processes to saline environments. Sci Rep 10(1):8232.

Pierleoni A, Martelli PL, Casadio R. 2008. PredGPI: a GPI-anchor predictor. BMC Bioinform 9(1):392.

Poon S, Heath RL, Clarke AE. 2012. A chimeric arabinogalactan protein promotes somatic embryogenesis in cotton cell culture. Plant Physiol 160(2):684-695.

Prokhorenkova L, Gusev G, Vorobev A, Dorogush AV, Gulin A. 2018. CatBoost: unbiased boosting with categorical features. In Proceedings of the 32nd International Conference on Neural Information Processing Systems (NIPS'18). Curran Associates Inc., Red Hook, NY, USA, 6639–6649

Qiu WR, Sun BQ, Xiao X, Xu ZC, Chou KC. 2016. iHyd-PseCp: Identify hydroxyproline and hydroxylysine in proteins by incorporating sequence-coupled effects into general PseAAC. Oncotarget 7(28):44310-44321.

Qu Y, Egelund J, Gilson PR, Houghton F, Gleeson PA, Schultz CJ, Bacic A. 2008. Identification of a novel group of putative *Arabidopsis thaliana* β-(1,3)-galactosyltransferases. Plant Mol Biol 68(1):43-59.

Quinlan JR. 1986. Induction of decision trees. Mach Learn 1(1):81-106.

R Core Team. 2020. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Rajkomar A, Dean J, Kohane I. 2019. Machine Learning in Medicine. N Engl J Med 380(14):1347-1358.

Rives A, Meier J, Sercu T, Goyal S, Lin Z, Liu J, Guo D, Ott M, Zitnick CL, Ma J, Fergus R. 2021. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proc Natl Acad Sci 118(15):e2016239118.

Roberts DR, Bahn V, Ciuti S, Boyce MS, Elith J, Guillera-Arroita G, Hauenstein S, Lahoz-Monfort JJ, Schröder B, Thuiller W, Warton DI, Wintle BA, Hartig F, Dormann CF. 2017. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography 40(8):913-929.

Roh Y, Heo G, Whang S. 2019. A Survey on Data Collection for Machine Learning: A Big Data - AI Integration Perspective. IEEE Trans Knowl Data Eng PP(99):1-1.

Rosso MG, Li Y, Strizhov N, Reiss B, Dekker K, Weisshaar B. 2003. An *Arabidopsis thaliana* T-DNA mutagenized population (GABI-Kat) for flanking sequence tag-based reverse genetics. Plant Mol Biol 53(1-2):247-259.

Saare-Surminski K, Preil W, Knox JP, Lieberei R. 2000. Arabinogalactan proteins in embryogenic and non-embryogenic callus cultures of *Euphorbia pulcherrima*. Physiol Plant 108(2):180-187.

Saini S, Sharma I, Pati PK. 2015. Versatile roles of brassinosteroid in plants in the context of its homoeostasis, signaling and crosstalks. Front Plant Sci 6(950).

Santarem ER, Pelissier B, Finer JJ. 1997. Effect of explant orientation, pH, solidifying agent and wounding on initiation of soybean somatic embryos. In Vitro Cell Dev Biol Plan. 33(1):13-19.

Savatin DV, Gramegna G, Modesti V, Cervone F. 2014. Wounding in the plant tissue: the defense of a dangerous passage. Front Plant Sci 5(470).

Schliep KP. 2010. phangorn: phylogenetic analysis in R. Bioinformatics 27(4):592-593.

Schliep KP, Hechenbichler K. 2016. kknn: Weighted k-Nearest Neighbors. R package version 1.3.1. https://CRAN.R-project.org/package=kknn

Schneider G, Wrede P. 1994. The rational design of amino acid sequences by artificial neural networks and simulated molecular evolution: de novo design of an idealized leader peptidase cleavage site. Biophys J 66(2 Pt 1):335-344.

Schultz C, Gilson P, Oxley D, Youl J, Bacic A. 1998. GPI-anchors on arabinogalactan-proteins: implications for signalling in plants. Trends Plant Sci 3(11):426-431.

Schultz CJ, Johnson KL, Currie G, Bacic A. 2000. The Classical Arabinogalactan Protein Gene Family of Arabidopsis. Plant Cell 12(9):1751-1767.

Schultz CJ, Rumsewicz MP, Johnson KL, Jones BJ, Gaspar YM, Bacic A. 2002. Using Genomic Resources to Guide Research Directions. The Arabinogalactan Protein Gene Family as a Test Case. Plant Physiol 129(4):1448-1463.

Schwartz D, Chou MF, Church GM. 2009. Predicting protein post-translational modifications using meta-analysis of proteome scale data sets. Mol Cell Proteomics 8(2):365-379.

Seifert GJ. 2018. Fascinating Fasciclins: A Surprisingly Widespread Family of Proteins that Mediate Interactions between the Cell Exterior and the Cell Surface. Int J Mol Sci 19(6):1628.

Seifert GJ, Roberts K. 2007. The Biology of Arabinogalactan Proteins. Annu Rev Plant Biol 58(1):137-161.

Serpe MD, Nothnagel EA. 1994. Effects of Yariv phenylglycosides on Rosa cell suspensions: Evidence for the involvement of arabinogalactan-proteins in cell proliferation. Planta 193(4):542-550.

Sessions A, Burke E, Presting G, Aux G, McElver J, Patton D, Dietrich B, Ho P, Bacwaden J, Ko C, Clarke JD, Cotton D, Bullis D, Snell J, Miguel T, Hutchison D, Kimmerly B, Mitzel T, Katagiri F, Glazebrook J, Law M, Goff SA. 2002. A High-Throughput Arabidopsis Reverse Genetics System. Plant Cell 14(12):2985-2994.

Sestili S, Sebastiani MS, Belisario A, Ficcadenti N. 2014. Reference gene selection for gene expression analysis in melon infected by *Fusarium oxysporum* f.sp. melonis. J Plant Biochem Biotechnol 23(3):238-248.

Shafee T, Bacic A, Johnson K. 2020. Evolution of Sequence-Diverse Disordered Regions in a Protein Family: Order within the Chaos. Mol Biol Evol 37(8):2155-2172.

Shen J, Zhang J, Luo X, Zhu W, Yu K, Chen K, Li Y, Jiang H. 2007. Predicting protein-protein interactions based only on sequences information. Proc Natl Acad Sci USA 104(11):4337-4341.

Shi S-P, Chen X, Xu H-D, Qiu J-D. 2015. PredHydroxy: computational prediction of protein hydroxylation site locations based on the primary structure. Mol Biosyst 11(3):819-825.

Showalter AM. 1993. Structure and function of plant cell wall proteins. Plant Cell 5(1):9-23.

Showalter AM. 2001. Arabinogalactan-proteins: structure, expression and function. Cell Mol Life Sci 58(10):1399-1417.

Showalter AM, Basu D. 2016. Glycosylation of arabinogalactan-proteins essential for development in Arabidopsis. Commun Integr Biol 9(3):e1177687.

Showalter AM, Keppler B, Lichtenberg J, Gu D, Welch LR. 2010. A Bioinformatics Approach to the Identification, Classification, and Analysis of Hydroxyproline-Rich Glycoproteins. Plant Physiol 153(2):485-513.

Shpak E, Leykam JF, Kieliszewski MJ. 1999. Synthetic genes for glycoprotein design and the elucidation of hydroxyproline-O-glycosylation codes. Proc Natl Acad Sci 96(26):14736-14741.

Shu H, Xu L, Li Z, Li J, Jin Z, Chang S. 2014. Tobacco arabinogalactan protein NtEPc can promote banana (*Musa* AAA) somatic embryogenesis. Appl Biochem Biotechnol 174(8):2818-2826.

Silva NF, Goring DR. 2002. The proline-rich, extensin-like receptor kinase-1 (PERK1) gene is rapidly induced by wounding. Plant Mol Biol 50(4-5):667-685.

Silver D, Schrittwieser J, Simonyan K, Antonoglou I, Huang A, Guez A, Hubert T, Baker L, Lai M, Bolton A, Chen Y, Lillicrap T, Hui F, Sifre L, van den Driessche G, Graepel T, Hassabis D. 2017. Mastering the game of Go without human knowledge. Nature 550(7676):354-359.

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31(19):3210-3212.

Simonović AD, Dragićević MB, Bogdanović MD, Trifunović-Momčilov MM, Subotić AR, Todorović SI. 2016. DUF1070 as a signature domain of a subclass of arabinogalactan peptides. Arch Biol Sci 68:737-746.

Simonović AD, Filipović BK, Trifunović MM, Malkov SN, Milinković VP, Jevremović SB, Subotić AR. 2015. Plant regeneration in leaf culture of *Centaurium erythraea* Rafn. Part 2: the role of arabinogalactan proteins. Plant Cell, Tissue Organ Cult 121(3):721-739.

Simonović AD, Trifunović-Momčilov MM, Filipović BK, Marković MP, Bogdanović MD, Subotić AR. 2021. Somatic Embryogenesis in *Centaurium erythraea* Rafn-Current Status and Perspectives: A Review. Plants 10(1):70.

Sommer-Knudsen J, Bacic A, Clarke AE. 1998. Hydroxyproline-rich plant glycoproteins. Phytochemistry 47(4):483-497.

Steinmacher DA, Saare-Surminski K, Lieberei R. 2012. Arabinogalactan proteins and the extracellular matrix surface network during peach palm somatic embryogenesis. Physiol Plant 146(3):336-349.

Stratford S, Barnes W, Hohorst DL, Sagert JG, Cotter R, Golubiewski A, Showalter AM, McCormick S, Bedinger P. 2001. A leucine-rich repeat region is conserved in pollen extensin-like (Pex) proteins in monocots and dicots. Plant Mol Biol 46(1):43-56.

Student Gosset, WS 1908. The probable error of a mean. Biometrika, 6(1):1–25.

Subotić A, Budimir S, Grubišić D, Momčilović I. 2003. Direct regeneration of shoots from hairy root cultures of *Centaurium erythraea* inoculated with *Agrobacterium rhizogenes*. Biol Plant 47:617-619.

Subotić A, Janković T, Jevremović S, Grubišić D. 2006. Plant tissue culture and secondary metabolites productions of *Centaurium erythraea* Rafn., a medicinal plant. In: Teixeira da Silva JA, (ed.), Floriculture, Ornamental and Plant Biotechnology: Advances and Topical Issues. Global Science Books UK, Vol. II, pp. 564-570.

Suzuki T, Narciso JO, Zeng W, van de Meene A, Yasutomi M, Takemura S, Lampugnani ER, Doblin MS, Bacic A, Ishiguro S. 2017. KNS4/UPEX1: A Type II Arabinogalactan β-(1,3)-Galactosyltransferase Required for Pollen Exine Development. Plant Physiol 173(1):183-205.

Šiler B, Avramov S, Banjanac T, Cvetković J, Nestorović Živković J, Patenković A, Mišić D. 2012. Secoiridoid glycosides as a marker system in chemical variability estimation and chemotype assignment of *Centaurium erythraea* Rafn from the Balkan Peninsula. Ind Crops Prod 40:336-344.

Šiler B, Živković S, Banjanac T, Cvetković J, Nestorović Živković J, Ćirić A, Soković M, Mišić D. 2014. Centauries as underestimated food additives: antioxidant and antimicrobial potential. Food Chem 147:367-376.

Tan C, Sun F, Kong T, Zhang W, Yang C, Liu C. 2018. A Survey on Deep Transfer Learning. In: Kůrková V, Manolopoulos Y, Hammer B, Iliadis L, Maglogiannis I. (eds) Artificial Neural Networks and Machine Learning – ICANN 2018. Lecture Notes in Computer Science, vol 11141. Springer, Cham.

Tan L, Mort A. 2020. Extensins at the front line of plant defence. A commentary on: 'Extensin arabinosylation is involved in root response to elicitors and limits oomycete colonization'. Ann Bot 125(5), vii-viii.

Tan L, Showalter A, Egelund J, Hernandez-Sanchez A, Doblin M, Bacic A. 2012. Arabinogalactan-proteins and the research challenges for these enigmatic plant cell surface proteoglycans. Front Plant Sci 3(140).

Tang X-C, He Y-Q, Wang Y, Sun M-X. 2006. The role of arabinogalactan proteins binding to Yariv reagents in the initiation, cell developmental fate, and maintenance of microspore embryogenesis in *Brassica napus* L. cv. Topas. J Exp Bot 57(11):2639-2650.

Teufel F, Almagro Armenteros JJ, Johansen AR, Gíslason MH, Pihl SI, Tsirigos KD, Winther O, Brunak S, von Hejine G, Nielsen H. 2022. SignalP 6.0 predicts all five types of signal peptides using protein language models. Nat Biotechnol

The UniProt Consortium. 2017. UniProt: The universal protein knowledgebase. Nucleic Acids Res. 45:D158–D169.

The UniProt Consortium. 2019. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res 47(D1):D506-D515.

Trifunović M, Cingel A, Simonović A, Jevremović S, Petrić M, Dragićević I, Motyka V, Dobrev P, Zahajská L, Subotić A. 2013. Overexpression of Arabidopsis cytokinin oxidase/dehydrogenase genes *AtCKX1* and *AtCKX2* in transgenic *Centaurium erythraea* Rafn. Plant Cell, Tissue Organ Cult 115:139–150.

Trifunović M, Subotić A, Petrić M, Jevremović S. 2015. The Role of Arabinogalactan Proteins in Morphogenesis of *Centaurium erythraea* Rafn *In Vitro*. In: Rybczyński JJ, Davey MR, Mikuła A, editors. The Gentianaceae - Volume 2: Biotechnology and Applications. Berlin, Heidelberg: Springer Berlin Heidelberg. p 113-138.

Trifunović M, Tadić V, Petrić M, Jontulović D, Jevremović S, Subotić A. 2014. Quantification of arabinogalactan proteins during *in vitro* morphogenesis induced by β-d-glucosyl Yariv reagent in *Centaurium erythraea* root culture. Acta Physiol Plant 36(5):1187-1195.

Tryfona T, Theys TE, Wagner T, Stott K, Keegstra K, Dupree P. 2014. Characterisation of FUT4 and FUT6 α-(1→2)-Fucosyltransferases Reveals that Absence of Root Arabinogalactan Fucosylation Increases Arabidopsis Root Growth Salt Sensitivity. PLoS One 9(3):e93291.

Van der Loo MPJ. 2014. The stringdist Package for Approximate String Matching. R Journal 6(1):111-122.

Van Hengel AJ, Van Kammen A, De Vries SC. 2002. A relationship between seed development, Arabinogalactan-proteins (AGPs) and the AGP mediated promotion of somatic embryogenesis. Physiol Plant 114(4):637-644.

Van Rossum F. 2009. Succession stage variation in population size in an early-successional herb in a peri-urban forest. Acta Oecol 35(2):261-268.

Vandesompele J, De Preter K, Pattyn F, Poppe B, Roy N, De Paepe A, Speleman F. 2002. Accurate Normalization of Real-Time Quantitative RT-PCR Data by Geometric Averaging of Multiple Internal Control Genes. Genome Biol 3:00341-003411.

Varma S, Simon R. 2006. Bias in error estimation when using cross-validation for model selection. BMC Bioinform 7:91.

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A, Kaiser L, Polosukhin I. 2017. Attention Is All You Need. Adv Neural Inf Process Syst 5998-6008.

Velasquez SM, Ricardi MM, Dorosz JG, Fernandez PV, Nadra AD, Pol-Fachin L, Egelund J, Gille S, Harholt J, Ciancia M, Verli H, Pauly M, Bacic A, Olsen CE, Ulvskov P, Petersen BL, Somerville C, Iusem ND, Estevez JM. 2011. O-glycosylated cell wall proteins are essential in root hair growth. Science 332(6036):1401-1403.

Velasquez SM, Ricardi MM, Poulsen CP, Oikawa A, Dilokpimol A, Halim A, Mangano S, Denita Juarez SP, Marzol E, Salgado Salter JD, Dorosz JG, Borassi C, Möller SR, Buono R, Ohsawa Y, Matsuoka K, Otegui MS, Scheller HV, Geshi N, Petersen BL, Iusem ND, Estevez JM. 2015. Complex regulation of prolyl-4-hydroxylases impacts root hair expansion. Mol Plant 8(5):734-746.

Villa-Rivera MG, Cano-Camacho H, López-Romero E, Zavala-Páramo MG. 2021. The Role of Arabinogalactan Type II Degradation in Plant-Microbe Interactions. Front Microbiol 12(2458).

Walsh I, Martin AJM, Di Domenico T, Tosatto SCE. 2012. ESpritz: accurate and fast prediction of protein disorder. Bioinformatics 28(4):503-509.

Walter P, Johnson AE. 1994. Signal sequence recognition and protein targeting to the endoplasmic reticulum membrane. Annu Rev Cell Biol 10:87-119.

Warnes GR, Bolker B, Bonebakker L, Gentleman R, Huber W, Liaw A, Lumley T, Maechler M, Magnusson A, Moeller S, Schwartz M, Venables B. 2020. gplots: Various R Programming Tools for Plotting Data. R package version 3.1.1. https://CRAN.R-project.org/package=gplots

Welch BL. 1947. The generalisation of student's problems when several different population variances are involved. Biometrika 34(1-2):28-35.

Whelan S, Goldman N. 2001. A General Empirical Model of Protein Evolution Derived from Multiple Protein Families Using a Maximum-Likelihood Approach. Mol Biol Evol 18(5):691-699.

Wickham H. 2018a. httr: Tools for Working with URLs and HTTP. R Package Version 1.4.0. https://CRAN.R-project.org/package=httr

Wickham H, Hester J, Ooms J. 2018b. xml2: Parse XML. R Package Version 1.2.0. https://CRAN.R-project.org/package=xml2

Wilkinson S. 2018. kmer: an R package for fast alignment-free clustering of biological sequences. R package version 1.1.1. https://cran.r-project.org/package=kmer

Wiśniewska E, Majewska-Sawka A. 2007. Arabinogalactan-proteins stimulate the organogenesis of guard cell protoplasts-derived callus in sugar beet. Plant Cell Rep 26(9):1457-1467.

Woody ST, Austin-Phillips S, Amasino RM, Krysan PJ. 2007. The WiscDsLox T-DNA collection: an arabidopsis community resource generated by using an improved high-throughput T-DNA sequencing pipeline. J Plant Res 120(1):157-165.

Wright ES. 2015. DECIPHER: harnessing local sequence context to improve protein multiple sequence alignment. BMC Bioinform 16(1):322.

Wright MN, Ziegler A. 2017. ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. J Stat Softw 77(1):1-17.

Xiao N, Cao D-S, Zhu M-F, Xu Q-S. 2015. protr/ProtrWeb: R package and web server for generating various numerical representation schemes of protein sequences. Bioinformatics 31(11):1857-1859.

Xie D, Ma L, Samaj J, Xu C. 2011. Immunohistochemical analysis of cell wall hydroxyproline-rich glycoproteins in the roots of resistant and susceptible wax gourd cultivars in response to *Fusarium oxysporum* f. sp. *Benincasae* infection and fusaric acid treatment. Plant Cell Rep 30(8):1555-1569.

Xu Y, Wen X, Shao XJ, Deng NY, Chou KC. 2014. iHyd-PseAAC: predicting hydroxyproline and hydroxylysine in proteins by incorporating dipeptide position-specific propensity into pseudo amino acid composition. Int J Mol Sci 15(5):7594-7610.

Yahraus T, Chandra S, Legendre L, Low PS. 1995. Evidence for a Mechanically Induced Oxidative Burst. Plant physiol 109(4):1259-1266.

Yang H, Wang M, Liu X, Zhao X-M, Li A. 2021. PhosIDN: an integrated deep neural network for improving protein phosphorylation site prediction by combining sequence and protein–protein interaction information. Bioinformatics 37(24):4668-4676.

Yang Y, Smith SA. 2013. Optimizing *de novo* assembly of short-read RNA-seq data for phylogenomics. BMC Genomics 14(1):328.

Yariv J, Lis H, Katchalski E. 1967. Precipitation of arabic acid and some seed polysaccharides by glycosylphenylazo dyes. Biochem J 105(1):1c-2c.

Yariv J, Rapport MM, Graf L. 1962. The interaction of glycosides and saccharides with antibody to the corresponding phenylazo glycosides. Biochem J 85(2):383-388.

Zhang Y, Showalter AM. 2020. CRISPR/Cas9 Genome Editing Technology: A Valuable Tool for Understanding Plant Cell Wall Biosynthesis and Function. Front Plant Sci 11.

Zhang Y, Yang J, Showalter AM. 2011. AtAGP18, a lysine-rich arabinogalactan protein in *Arabidopsis thaliana*, functions in plant growth and development as a putative co-receptor for signal transduction. Plant Signal Behav 6(6):855-857.

Zhao Z, Anand R, Wang M. 2019. Maximum Relevance and Minimum Redundancy Feature Selection Methods for a Marketing Machine Learning Platform. IEEE International Conference on Data Science and Advanced Analytics (DSAA) pp 442-452.

Zhou K. 2019. Glycosylphosphatidylinositol-Anchored Proteins in Arabidopsis and One of Their Common Roles in Signaling Transduction. Front Plant Sci 10:1022.

Zhou L, Thornburg R. 1999. Wound-Inducible Genes in Plants. In: Reynolds PHS, ed. Inducible Gene expression. Wallingford, UK: CAB International, 127–156.

Zhu X, Li X, Chen W, Chen J, Lu W, Chen L, Fu D. 2012. Evaluation of new reference genes in papaya for accurate transcript normalization under different experimental conditions. PLoS One 7(8):e44405.

Zuo C, Liu H, Lv Q, Chen Z, Tian Y, Mao J, Chu M, Ma Z, An Z, Chen B. 2020. Genome-Wide Analysis of the Apple (*Malus domestica*) Cysteine-Rich Receptor-Like Kinase (CRK) Family: Annotation, Genomic Organization, and Expression Profiles in Response to Fungal Infection. Plant Mol Biol Rep 38(1):14-24.

Biografija

Danijela M. Paunović (devojačko prezime Jontulović) rođena je 12.06.1985. u Čačku gde je završila osnovnu i srednju školu (Gimnazija, prirodno-matematički smer). Tokom srednje škole bila je polaznik Regionalnog centra za talente Čačak i učesnik na republičkim takmičenjima iz biologije u organizaciji centra, gde je 2004. osvojila prvo mesto. Biološki fakultet Univerziteta u Beogradu, smer Molekularna biologija i fiziologija, upisala je školske 2004/2005. godine. Diplomski rad pod nazivom „Detekcija mutacija B-RAF onkogena u uzorcima tumora pluća SSCP metodom i direktnim sekvenciranjem“ uradila je u laboratoriji Instituta za medicinska istraživanja, VMA. Doktorske studije na Biološkom fakultetu Univerziteta u Beogradu, studijski program biologija, modul Fiziologija i molekularna biologija biljaka, upisala je školske 2013/2014. godine. U školskoj 2013/2014. i 2014/2015. učestovala je u izvođenju nastave na predmetima Fiziologija biljaka i Rastenje i razviće biljaka na katedri za Fiziologiju biljaka, Biološkog fakulteta Univerziteta u Beogradu. Kao istraživač pripravnik 2015. godine zaposlena je na Odeljenju za fiziologiju biljaka Instituta za biološka istraživanja „Siniša Stanković”, Instituta od nacionalnog značaja za Republiku Srbiju, Univerziteta u Beogradu. Zvanje istraživač saradnik stekla je 2018. godine. Učestvovala je u realizaciji nacionalnog projekta TR31019: „Razvoj i primena biotehnoloških postupaka u dobijanju zdravog sadnog materijala ukrasnih biljaka” finansiranog od strane Ministarstva prosvete, nauke i tehnološkog razvoja Republike Srbije, pod rukovodstvom dr Angeline Subotić. Danijela Paunović je član Društva za fiziologiju biljaka Srbije (DFBS) i Srpskog biološkog društva (SBD).

Изјава о ауторству

Потписана: **Данијела М. Пауновић**

Број индекса: Б3037/2013

Изјављујем

да је докторска дисертација под насловом:

Идентификација *AGP* гена кичице (*Centaurium erythraea*, Gentianaceae) и праћење њихове експресије у одговору на механичке повреде биљног ткива гајеног *in vitro*

- резултат сопственог истраживачког рада,
- да предложена дисертација у целини ни у деловима није била предложена за добијање било које дипломе према студијским програмима других високошколских установа;
- да су резултати коректно наведени и
- да нисам кршио/ла ауторска права и користио интелектуалну својину других лица.

Потпис докторанда

У Београду, _____ _____

Изјава о истоветности штампане и електронске верзије докторског рада

Име и презиме аутора: **Данијела М. Пауновић**

Број индекса: **Б3037/2013**

Студијски програм: **Биологија (Молекуларна биологија и физиологија биљака)**

Наслов рада: **Идентификација *AGP* гена кичице (*Centaurium erythraea,* Gentianaceae) и праћење њихове експресије у одговору на механичке повреде биљног ткива гајеног *in vitro***

Ментори: **др Милан Драгићевић, виши научни сарадник**

др Ивана Драгићевић, ванредни професор

Изјављујем да је штампана верзија мог докторског рада истоветна електронској верзији коју сам предао/ла за објављивање на порталу **Дигиталног репозиторијума Универзитета у Београду.**

Дозвољавам да се објаве моји лични подаци везани за добијање академског звања доктора наука, као што су име и презиме, година и место рођења и датум одбране рада.

Ови лични подаци могу се објавити на мрежним страницама дигиталне библиотеке, у електронском каталогу и у публикацијама Универзитета у Београду.

Потпис докторанда

У Београду, _____ _____

Изјава о коришћењу

Овлашћујем Универзитетску библиотеку „Светозар Марковић“ да у Дигитални репозиторијум Универзитета у Београду унесе моју докторску дисертацију под насловом:

Идентификација *AGP* гена кичице (*Centaurium erythraea*, Gentianaceae) и праћење њихове експресије у одговору на механичке повреде биљног ткива гајеног *in vitro*

која је моје ауторско дело.

Дисертацију са свим прилозима предао/ла сам у електронском формату погодном за трајно архивирање.

Моју докторску дисертацију похрањену у Дигитални репозиторијум Универзитета у Београду могу да користе сви који поштују одредбе садржане у одабраном типу лиценце Креативне заједнице (Creative Commons) за коју сам се одлучио/ла.

① Ауторство (CC BY)

2. Ауторство – некомерцијално (CC BY-NC)

3. Ауторство – некомерцијално – без прерада (CC BY-NC-ND)

4. Ауторство – некомерцијално – делити под истим условима (CC BY-NC-SA)

5. Ауторство – без прерада (CC BY-ND)

6. Ауторство – делити под истим условима (CC BY-SA)

(Молимо да заокружите само једну од шест понуђених лиценци. Кратак опис лиценци је саставни део ове изјаве).

Потпис докторанда

У Београду, _____ _____

1. **Ауторство.** Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, и прераде, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце, чак и у комерцијалне сврхе. Ово је најслободнија од свих лиценци.
2. **Ауторство – некомерцијално**. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, и прераде, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце. Ова лиценца не дозвољава комерцијалну употребу дела.
3. **Ауторство – некомерцијално – без прерада.** Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, без промена, преобликовања или употребе дела у свом делу, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце. Ова лиценца не дозвољава комерцијалну употребу дела. У односу на све остале лиценце, овом лиценцом се ограничава највећи обим права коришћења дела.
4. **Ауторство – некомерцијално – делити под истим условима.** Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, и прераде, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце и ако се прерада дистрибуира под истом или сличном лиценцом. Ова лиценца не дозвољава комерцијалну употребу дела и прерада.
5. **Ауторство – без прерада.** Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, без промена, преобликовања или употребе дела у свом делу, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце. Ова лиценца дозвољава комерцијалну употребу дела.
6. **Ауторство – делити под истим условима.** Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, и прераде, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце и ако се прерада дистрибуира под истом или сличном лиценцом. Ова лиценца дозвољава комерцијалну употребу дела и прерада. Слична је софтверским лиценцама, односно лиценцама отвореног кода.