COBISS Code 1.02

Agrovoc descriptors: coffea arabica, arabica coffee, coffea canephora, coffea congensis, congusta coffee, robusta coffee, nucleotide sequence, gene expression, hybrids, hybridization, genetic code, nucleotides

Agris category code: F30

Analiza EST klonov križancev *Coffea arabica* X *Coffea canephora* in *Coffea canephora* X *Coffea congensis*

Tina SVETEK[1], Nataša ŠIBANC[2]

Delo je prispelo 20. januarja 2012, sprejeto 14. marca 2012.
Received January 20, 2012; accepted March 14, 2012.

IZVLEČEK

EST ali oznake izraženih zaporedij so DNA zaporedja, dolga od 100 do 800 baznih parov, pridobljena z eno reakcijo določanja nukleotidnega zaporedja cDNA molekulam iz 5' ali 3' smeri. Vsebujejo prepisana, ne nujno pa tudi prevedena, zaporedja genov ter pogosto tudi elemente vektorjev. Predstavljajo presežen nabor izraženih genov v nekem vzorcu in se uporabljajo za študije izražanja genov, iskanje novih genov, raziskave alternativnega izrezovanja intronov idr. Mnogokrat predstavljajo prvo orodje funkcionalne genomike manj raziskanih organizmov. Zaradi presežnosti se jih mnogokrat združuje v gruče. Največjo zbirko EST zaporedij vzdržuje NCBI, imenuje se dbEST in ima več kot 70 milijonov zaporedij. V omenjeni bazi smo poiskali klone EST dveh križancev kave, *Coffea arabica* X *Coffea canephora* ter *Coffea canephora* X *Coffea congensis*, ter s pomočjo BLAST algoritma poiskali katere proteine kodirajo. Najdenim proteinom smo nato določili ontologijo.

Ključne besede: EST, *Coffea arabica* X *Coffea canephora*, *Coffea canephora* X *Coffea congensis*, BLAST, ontologija, bioinformatika

ABSTRACT

EST CLONE ANALYSIS OF TWO COFFEE HYBRIDS (*Coffea arabica* X *Coffea canephora* and *Coffea canephora* X *Coffea congensis*)

Expressed sequence tags (ESTs) are short (from 100 to 800 base pairs) 5' or 3' sequences that are acquired with single pass sequencing of cDNA molecules. They contain transcribed, but not necessarily translated regions of genes and often also vector elements. They represent a redundant set of expressed genes in a given sample and are used in gene expression studies, finding new genes, alternative splicing research etc. ESTs often represent primary tool of functional genomic of orphan crops. They are often clustered due to their redundancy. The largest EST collection named dbEST, it is maintaned by NCBI and contains more than 70 million sequences. In this database, we have searched for EST clones of two coffee hybrids, *Coffea arabica* X *Coffea canephora* and *Coffea canephora* X *Coffea congensis*, and used BLAST algorithm to find out which proteins they are encoding. We have also determined gene ontology of protein hits.

Key words: EST, *Coffea arabica* X *Coffea canephora*, *Coffea canephora* X *Coffea congensis*, BLAST, ontology, bioinformatics

1 UVOD

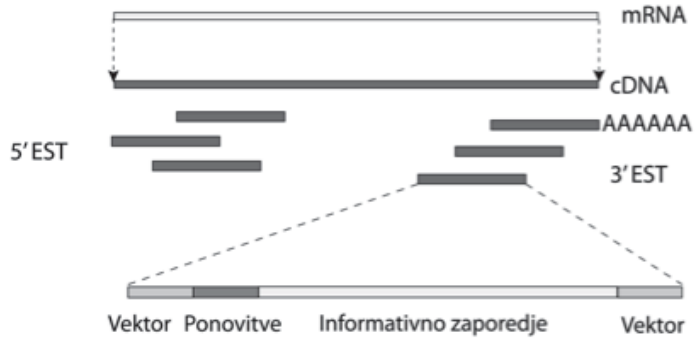
Oznake izraženih zaporedij, EST (angl. expressed sequence tag) so kratki deli DNA zaporedja, ki nastanejo iz določanja nukleotidnega zaporedja enega ali obeh koncev izraženega gena. Zaporedje določimo delom DNA, ki predstavljajo izražene gene določene celice, tkiva ali organa različnih organizmov in uporabimo te oznake za iskanje genov iz kopice kromosomske DNA. EST zaporedja in zaporedja komplementarne DNA (cDNA, angl. complementary DNA) nam omogočajo pregled vzorcev transkriptov, in so pomemben vir transkriptomskih raziskav. EST so presežna zaporedja dolga od 200 do 800 nukleotidnih baz v primeru Sangerjeve tehnologije, dobljena z eno reakcijo sekvenciranja (angl. single pass sequencing) iz cDNA knjižnic. Že za relativno nizko ceno lahko pridobimo večje število EST cDNA klona in tako

[1] Univerza v Ljubljani, Biotehniška fakulteta, Oddelek za agronomijo, Katedra za genetiko, biotehnologijo, statistiko in žlahtnjenje rastlin, Jamnikarjeva 101, 1000 Ljubljana, Slovenija; univ. dipl. bioteh., mlada raziskovalka, tina.svetek@bf.uni-lj.si

[2] Univerza v Ljubljani, Biotehniška fakulteta, Oddelek za agronomijo, Katedra za aplikativno botaniko, ekologijo, fiziologijo rastlin in informatiko, Jamnikarjeva 101, 1000 Ljubljana, Slovenija; univ. dipl. ing. agr., mlada raziskovalka, natasa.sibanc@bf.uni-lj.si

dobimo informacijo o prepisanih regijah posameznega organizma.
Kopije izraženih genov dobimo iz zaporedij informacijske RNA (mRNA) v celici. Ker RNA ne moremo direktno klonirati, jih z encimom reverzno transkriptazo prepišemo v cDNA. Dobljeno cDNA kloniramo in ustvarimo knjižnico, ki predstavlja set prepisanih genov prvotne celice, tkiva ali organizma. Tem cDNA klonom naključno določimo nukleotidno zaporedje z eno samo reakcijo določanja nukleotidnega zaporedja iz obeh koncev, da dobimo 5' ali 3' zaporedje. Rezultati so lahko presežni EST-ji, saj so lahko cDNA matrice sekvencirane delno, v celotni dolžini ali pa so v knjižnici že same po sebi presežne.

Dias Neto in sod. (2000) so ustvarili novo, cenovno ugodno metodo za pridobivanje velikega števila EST, oznake izraženega okvirja z odprtim bralnim okvirjem, imenovane ORESTES (angl. open reading frame expressed sequence tags). Ta metoda se razlikuje od konvencionalnega pridobivanja EST tako, da pridobimo zaporedje iz sredinske kodirajoče regije, ki je najbolj informativna. ORESTES nukleotidne podatke lahko prav tako najdemo v dbEST bazi.



Slika 1: Osnovne značilnosti klonov EST. Gre za krajše fragmente cDNA, ki lahko poleg kodirajočega zaporedja vsebujejo tudi neprevedena zaporedja (5' ali 3' UTR, angl. untranslated region) ter zaporedja vektorjev.

Figure 1: Basic characteristics of the EST clones. This is a short fragment of cDNA containing coding sequences, and can also contain 5' and 3' untranslated regions and sequences of vectors.

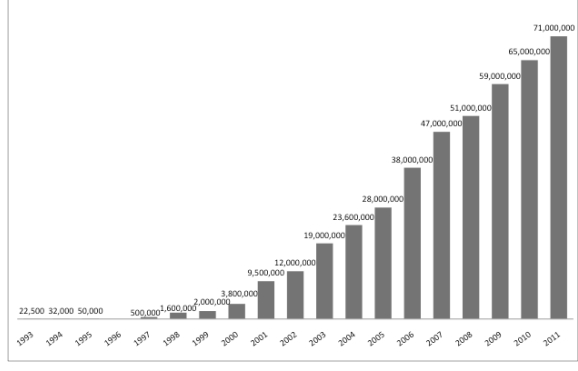
EST zaporedje (Slika 1) je le kratka kopija mRNA, ki je sekvencirana samo enkrat in je zelo podvržena napakam, še posebno na koncih. Kvaliteta zaporedja je ponavadi večja na sredini. Vektorje in ponovljena zaporedja izrežemo v postopku pred-procesiranja EST-jev. Pred-procesiranje zmanjšuje skupne motnje, ki nastanejo pri EST podatkih in tako izboljša učinkovitost nadaljnjih analiz. Splošno, je kvaliteta odčitavanja baz v posameznih EST zaporedjih na začetku slaba (do 20 % v prvih 50 do 100 baznih parih), nato se izboljša in ponovno poslabša proti koncu (Aaronson in sod., 1996). Presežnost in preveč ter premalo zastopani transkritpi so dejanski problemi pri EST podatkih. Razlog je predvsem v različni stopnji izražanja določenih genov v različnih tkivih, deloma pa tudi v neenotnosti protokolov uporabljenih pri pridobivanju EST. Pogosto opažene napake EST so tudi artefakti zaporedij, skupno tudi do 5 % (Aaronson in sod., 1996), ponavljanje baz, še posebno G in T, in slaba kvaliteta zaporedij. Pride lahko tudi do pogoste kontaminacije iz vektorjev, adapterjev in himernih zaporedij, kot tudi iz genomskih DNA fragmentov. Slaba kvaliteta značilnih zaporedij, kratka zaporedja, ponovitve in napake v anotaciji lahko predstavljajo probleme za nadaljnjo analizo. Tudi naravne variacije v procesih, kot je alternativno RNA procesiranje in genomske variacije, nastale zaradi SNP (angl. single nucleotide polymorphism) lahko predstavljajo izzive, saj je težko razlikovati med artefaktnimi in naravno prisotnimi zamenjavami in insercijami ter delecijami v danem podatkovnem setu EST.

Največja in prosto dostopna baza EST podatkov (71,276,166 EST iz 2325 organizmov, december 2011) je dbEST (http://www.ncbi.nlm.nih.gov/dbEST/). UniGene iz National Center for Biotechnology Information (NCBI) (www.ncbi.nlm.nih.gov/UniGene), Združene države Amerike, shranjuje edinstvene gene in predstavlja nepresežen set gensko orientiranih gruč nastalih iz EST. Drugi specializiran vir EST ustvarjen za specifične organizme pa je na Dana Farber Cancer Institute, predhodno urejen in vzdrževan na The Institute for Genome Research (TIGR; http://compbio.dfci.harvard.edu/tgi/).

Kontaminacija z vektorji je pri EST široko razširjena, in pogosto se del vektorja ali adaptorja, ki smo ga uporabili pri kloniranju sekvencira skupaj z EST zaporedjem. Takšni vektorji morajo biti odstranjeni preden so EST-ji zbrani v gruče. Kontaminacije lahko tako identificiramo in jih izločimo, če primerjamo EST z nepresežnimi bazami vektorjev. Primerni viri za pred-procesiranje EST klonov so podatkovna baza UniVec (http://ncbi.nlm.nih.gov/VecScreen/UniVec.html), programsko orodje za primerjavo EMVec (http://www.ebi.ac.uk/Tools/sss/ncbiblast/vectors.html) in tudi orodje RepeatMasker (http://repeatmasker.org/cgi-bin/WEBRepeatMasker).

Ker so EST podatki presežni in vsako zaporedje vsebuje le majhno informacijo o zaporedju gena, se jih na podlagi identičnosti združuje v gruče. Deloma to delamo zaradi zmanjšanja števila ponovljenih transkriptov, deloma pa tudi zato, da transkripte istega gena združimo v isto gručo, s čimer smo korak bližje celotnemu zaporedju gena. Enostaven način za zbiranje EST je z merjenjem podobnosti sekvenčnih parov med njimi. Te razdalje so potem pretvorjene v binarne enote, glede na to, ali se značilno ujemajo ali ne, in tako je sekvenčni par sprejet v nastajajočo gručo ali izločen iz nje. Dva pristopa zbiranja sta opisala Ptitsyn in Hide (2005) kot zaostreno (angl. stringent) in ohlapno (angl. loose) zbiranje. Zaostren tip zbiranja je konzervativen in temelji na enkratnem zbiranju EST, kar da relativno natančne gruče, vendar so nastala zaporedja krajša z manjšim številom izraženih genov. V nasprotnem primeru, nam ohlapno zbiranje s ponavljanjem poravnav EST zaporedij slabše kvalitete ustvari manj natančna zaporedja, ki pa so daljša in imajo tako večjo pokritost izraženega gena ter podajo boljšo informacijo o alternativnem izrezovanju intronov, vendar obstaja nevarnost da se v gručo vključi tudi paralogna zaporedja. Pristop, ki ga uporablja TIGR je zaostreno zbiranje, UniGene pa je med zaostrenim in ohlapnim zbiranjem.

Najbolj pogosto uporabljeni programi za zbiranje in združevanje EST zaporedij, pridobljenih s Sangerjevo tehnologijo, so Phrap (Ewing in Green, 1998) (http://www.phrap.org), CAP3 (Huang in Madan, 1999) (http://pbil.univ-lyon1.fr/cap3.php) in zelo popularno orodje izdelano na TIGR TGICL (Lee in sod., 2005) (angl. TIGR gene indices clustering tools), ki združuje programa megablast in CAP3. Primerjava teh treh programov (Liang in sod., 2000) je pokazala, da je CAP3 najbolj optimalen za uporabo.



Slika 2: Število zaporedij v bazi dbEST po letih. Od ustanovitve leta 1992 je število klonov EST strmo naraščalo. Večina zaporedij je bilo humanega izvora, saj so bili EST pomembno orodje pri odkrivanju novih genov v človeškem genomu.

Figure 2: Number of sequences in the dbEST database through the years. Since its beginnings in 1992, the number of ESTs has been growing rapidly. The majority of sequences comes from human, since the ESTs have been an important tool in finding new genes in the human genome.

Ko pridobimo skupno zaporedje iz sestavljenih EST, jim lahko pripišemo funkcije, do katerih pridemo s pomočjo iskanja podobnosti z že znanimi zaporedji v podatkovnih bazah. Najbolj univerzalno in znano orodje za iskanje podobnosti med zaporedji v bazah je BLAST (Altschul in sod., 1997) (angl. The basic local alignment search Tool) na strežniku NCBI (http://blast.ncbi.nlm.nih.gov/Blast.cgi) ali v obliki namizne različice programa. BLAST primerja nukleotidna, proteinska ali prevedena nukleotidna zaporedja z zaporedji iz baze podatkov in izračuna statistično značilnost med ujemanji. BLAST lahko uporabljamo za iskanje funkcionalnih in evolucijskih relacij med zaporedji in tudi za pomoč pri identifikaciji članov iz določene družine genov. Obstajajo različni algoritmi primerjav, BLASTN za iskanje nukleotidov v bazah nukleotidnih zaporedij, BLASTP za iskanje proteinov v proteinskih bazah, BLASTX prav tako za iskanje v proteinski bazi vendar za primerjavo uporabi prevedeno nukleotidno zaporedje, TBLASTN išče proteinska ujemanja s prevedeno nukleotidno bazo in TBLASTX išče podobnosti prevedenih nukleotidnih zaporedij s prevedeno nukleotidno bazo na aminokislinskem nivoju.

Zgodovina EST-jev sovpada z začetki avtomatiziranega določanja nukleotidnega zaporedja (približno leta 1990). EST-ji so igrali pomembno vlogo pri odkrivanju genov v projektu človeškega genoma, saj je bilo v začetku identificiranih in fizično kartiranih zelo malo človeških genov (Adams in sod., 1991). Uporabljali so jih za iskanje novih genov, za kartiranje genov na kromosome in za raziskovanje profila izražanja genov. EST zaporedja so uporabna tudi za študije strukture genov, alternativnega izrezovanja intronov ter diferencialno izraženih genov (na primer primerjava med zdravim in bolnim tkivom). Podatki o EST-jih so hitro postali množično uporabljeni in popularni, kar se odraža tudi v

hitrem naraščanju le-teh v bazi dbEST (Slika 2). V več kot 19 letih, odkar obstaja omenjena podatkovna zbirka, je število EST zaporedij iz dobrih 22.000 zraslo na več kot 71 milijonov. Prvi objavljeni EST so prihajali iz sedmih organizmov, danes je v dbEST zastopanih preko 2.000 različnih organizmov. V vrhu po številu EST zaporedij so: človek, miš, koruza, prašič, navadni repnjakovec, govedo, zebrica, soja, Xenopus, riž, pšenica in podgana. Danes EST predstavljajo večino (približno 60 %) zaporedij v podatkovni zbirki GenBank. Tudi v zadnjih letih, ko se močno uveljavljajo nove in hitrejše metode določanja nukleotidnega zaporedja celih genomov, število EST vztrajno narašča. Vzrok za to je verjetno dejstvo, da se nove metode določevanja nukleotidnega zaporedja celotnih genomov še uveljavljajo, za manjše laboratorije pa je za enkrat še vedno enostavnejše in cenejše pridobivanje EST zaporedij. Sčasoma se bodo EST verjetno umaknili in jih bodo nadomestile novejše, hitrejše in zanesljivejše metode, kot je na primer RNA-Seq (Ozsolak in sod, 2009).

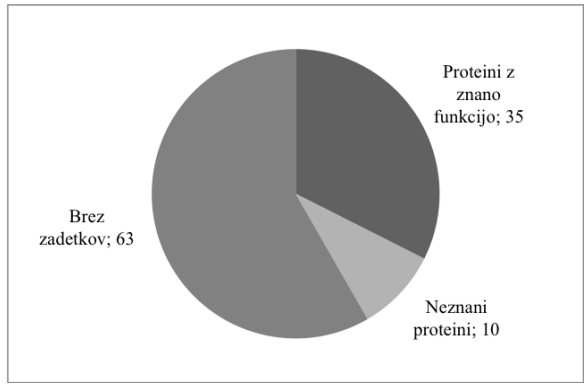
Namen prispevka je predstaviti EST, njihovo pridobivanje ter napake in rešitve napak, ki se pri tem lahko pojavijo. Zastavili smo si tudi praktični primer uporabe EST z analizo klonov dveh medvrstnih križancev rodu kave (*Coffea* sp.). Na tem primeru smo analizirali pridobljena EST zaporedja in primerjali kodirane proteine obeh križancev med seboj z uporabo genske ontologije.

2 MATERIAL IN METODE

V podatkovni bazi dbEST smo poiskali 108 EST zaporedij križanca *Coffea arabica* X *Coffea canephora* ter 145 zaporedij križanca *Coffea canephora* X *Coffea congensis*. Obe skupini zaporedij smo v FASTA obliki zbrali v dveh ločenih tekstovnih datotekah. Nato smo uporabili NetBlast program, ki omogoča iskanje po NCBI bazah brez internetnega vmesnika preko ukazne vrstice na lastnem računalniku. Program smo uporabili zato, ker omogoča iskanje z več zaporedji hkrati in omogoča delo z večimi zaporedji naenkrat, brez posredovanja uporabnika. Za iskanje zaporedij, podobnih našim EST zaporedjem, smo uporabili BLASTX algoritem, ki primerja prevedeno nukleotidno zaporedje z aminokislinskimi zaporedji. Izpis rezultatov smo omejili z vrednostjo e manjšo od 0,01. Vrednost e (angl. expected value) je pri BLAST analizi parameter, ki označuje pričakovano število naključnih zadetkov v dani bazi. Manjši kot je e, manj naključnih ujemanj lahko pričakujemo in bolj signifikanta je poravnava. Tekstovno datoteko z rezultati smo spremenili v tabelarično obliko z doma napisano PERL skripto, za kar smo uporabili BioPerl paket, modul BIO:searchIO (skripta je na voljo pri avtoricah). Ker smo delali v okolju Windows, smo potrebovali tudi nameščen Perl jezikovni program (ActivePerl). V tabeli z rezultati smo imeli izpisane akcesije EST zaporedij, njihove dolžine, akcesije njihovih zadetkov, opise zadetkov, aminokislinska zaporedja, dolžino poravnave, e vrednost in rezultat poravnave. Iz najboljšega anotiranega zadetka smo sklepali kateri protein kodira posamezno EST zaporedje. Za vsak kodiran protein smo nato s pomočjo baz UniProt (http://www.uniprot.org/) in Gene Ontology (http://www.geneontology.org/) določili vse tri ontologije, torej kje v celici se protein nahaja, kakšna je njegova molekularna funkcija in v kakšnem biološkem procesu sodeluje.

3 REZULTATI IN RAZPRAVA

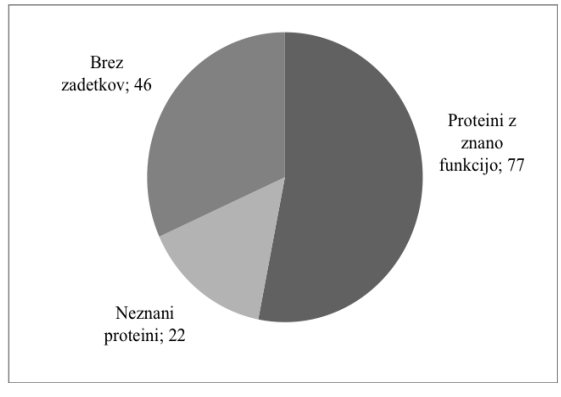
Izmed 108 klonov EST iz križanca *C. arabica* X *C. canephora*, ki smo jih našli v bazi dbEST, jih je imelo zadetke po izvedbi BLAST algoritma le 45 (Slika 3). Petintrideset od teh zadetkov so proteini, ki imajo znano funkcijo, 10 zadetkov pa so proteini z neznano funkcijo ali pa hipotetični in predvideni proteini. Do napovedi za hipotetičen protein ponavadi pride pri analizi genoma, kjer se najde dovolj velik odprt bralni okvir, za katerega se predvideva, da verjetno kodira nek proteinski produkt, vendar pa ni eksperimentalnih dokazov za obstoj proteina *in vivo*. Včasih imajo zaporedja predvidenih proteinov značilne regije, kot npr. določene funkcionalne domene, na podlagi katerih se lahko sklepa, kakšno funkcijo bi protein imel, če bi se dejansko izražal.



Slika 3: Rezultati uporabe algoritma BLAST na 108 EST iz *C. arabica* X *C. canephora*.

Figure 3: Results of BLAST algorithm used on 108 ESTs from *C. arabica* X *C. Canephor.*

Izmed 145 klonov EST križanca *C. canephora* X *C. congensis* iz baze dbEST, je imelo zadetke po uporabi BLAST algoritma 99 zaporedij (Slika 4). Kar 77 je bilo proteinov z znano funkcijo, neznani ali napovedani proteini pa so predstavljali 22 zadetkov (Slika 4).

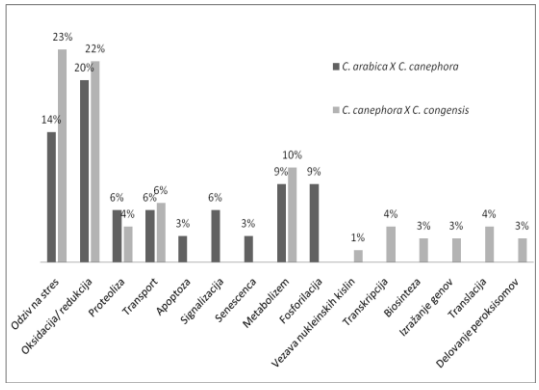


Slika 4: Rezultati uporabe algoritma BLAST na 145 EST iz *C. canephora* X *C. congensis.*

Figure 4: Results of BLAST algorithm used on 108 ESTs from *C. canephora* X *C. congensis.*

Na splošno je imela večina EST iz *C. arabica* X *C. canephora* zadetke z visokimi vrednostmi e, dolžine poravnav so bile krajše kot pri EST iz *C. canephora* X *C. congensis*. Vzrok za zaporedja brez podobnih zadetkov iz baz je lahko v tem, da smo iskali po proteinski bazi, EST pa lahko vsebujejo tudi neprevedene (UTR) regije. Možen vzrok, vendar manj verjeten, je tudi da gre za nova, specifična zaporedja, ki jih še ni v bazi. Med pridobivanjem EST pa je lahko prišlo tudi do genomskih kontaminacij, ki se prav tako kažejo v zaporedjih brez podobnih zadetkov.

Proteine, ki jih kodirajo kloni EST obeh križancev, smo razporedili glede na lokacijo v celici, njihovo molekularno funkcijo in biološki proces, v katerem sodelujejo.

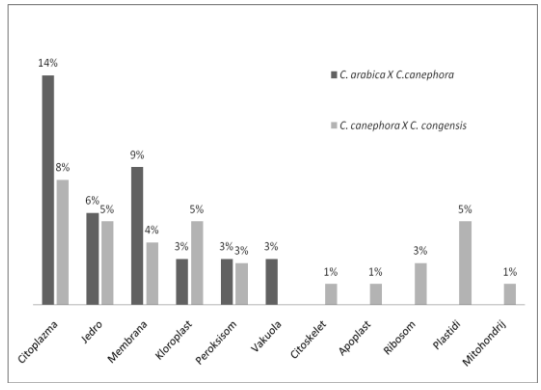


Slika 5: Biološki procesi v katerih sodelujejo proteini, ki jih kodirajo EST zaporedja obeh križancev. Predstavljeni so odstotki vseh proteinov z znano funkcijo.

Figure 5: Biological processes that involve proteins, coded by EST sequences of both hybrids. Shown in the figure are percentages of all proteins with known function.

Od 35 določenih proteinov iz križanca *C. arabica* X *C. canephora* jih je podatke o biološki vlogi vsebovalo 24. Iz drugega križanca smo našli tovrstne podatke za 60 od 77 proteinov. Slika 5 prikazuje odstotek določenih proteinov iz vsakega križanca, ki nastopajo v določenem biološkem procesu. Nekateri proteini nastopajo v več kot enem procesu. Največ proteinov ima vlogo v odzivu na stres – tu gre predvsem za veliko število proteinov toplotnega šoka, ki so se pojavili kot zadetki po uporabi algoritma BLAST. Naslednjo veliko skupino prestavljajo proteini, ki sodelujejo v redukciji ali oksidaciji. Vzrok za velik delež teh proteinov je verjetno v tem, da so redoks procesi osnovni procesi v celicah, mnogi izmed proteinov pa niso imeli natančneje določene vloge v tem procesu (na primer oksidacija točno določenih spojin). Tretja večja skupina bioloških procesov je metabolizem – v tej kategoriji so prisotni proteini, ki sodelujejo v metabolizmu ogljikovih hidratov in drugih bioloških molekul ter proteini, ki nimajo natančneje določene vloge v metabolizmu. Težko je narediti primerjavo proteinov obeh križancev, ker je že samo število zadetkov in določenih proteinov zelo različno (križanec *C. canephora* X *C. congensis* ima določenih dvakrat več proteinov kot primerjani križanec). Razumljivo je, da se v obeh rastlinah pojavljajo temeljni proteini, nujni za preživetje celice, kot so metabolni encimi, redoks encimi in transportni proteini. Zanimivo je, da so v križancu *C. arabica* X *C. canephora* prisotni tudi proteini, ki sodelujejo v senescenci in apoptozi.

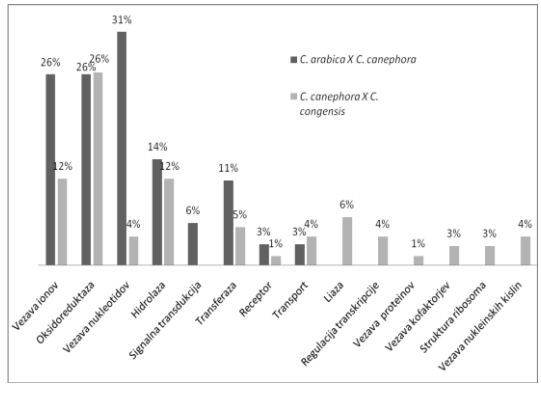
Podatke o celični lokaciji proteinov smo našli za 12 proteinov iz *C. arabica* X *C. canephora* in za 28 proteinov iz C. canephora X C. congensis (Slika 6). Po pričakovanjih je večina proteinov locirana v citoplazmi, jedru ali membrani. *C. arabica* X *C. canephora* ima manjši delež jedrnih proteinov, kar sovpada z zgornjo ugotovitvijo, da ima tudi manj proteinov, ki sodelujejo pri izražanju genov.



Slika 6: Lokacija proteinov, ki jih kodirajo EST obeh križancev. Predstavljeni so odstotki vseh proteinov z znano funkcijo.

Figure 6: Location of proteins, coded by EST sequences of both hybrids. Shown in the figure are percentages of all proteins with known function.

Lokacije proteinov v celici so bile bolj natančno določene v križancu *C. canephora* X *C. congensis*, saj smo našli primere proteinov citoskeleta, ribosomalnih proteinov ter več plastidnih proteinov.



Slika 7: Molekularne funkcije v katerih sodelujejo proteini, ki jih kodirajo EST zaporedja obeh križancev. Predstavljeni so odstotki vseh proteinov z znano funkcijo.

Figure 7: Molecular functions involving proteins, coded by EST sequences of both hybrids. Shown in the figure are percentages of all proteins with known function.

Podatke o molekularni funkciji proteinov smo našli za 26 od 35 proteinov iz *C. arabica* X *C. canephora* in za 46 od 77 proteinov iz *C. canephora* X *C. congensis* (Slika 7). Tudi pri tej ontologiji se je pojavljal problem različne mere natančnosti, s katero so bili anotirani proteini. Po pričakovanjih glede na rezultate, prikazane na sliki 5, ima veliko število proteinov oksidoreduktazno aktivnost. Za križanca *C. arabica* X *C. canephora* so značilni še proteini, ki vežejo ione, tako kovinske kot tudi druge, ter proteini, ki vežejo nukleotide, ni pa prisotnih proteinov, ki so značilni za procese nadzorovanja transkripcije. V obeh rastlinah so prisotne hidrolaze (proteaze, peptidaze, fosfataze, esteraze idr.), saj sodelujejo v pomembnih metabolnih procesih. Tudi v tem primeru je težko primerjati najdene proteine med obema križancema, ker je najdeno število proteinov z določeno funkcijo zelo različno. Kot v obeh prej omenjenih primerih, je za križanca *C. canephora* X *C. congensis* značilna večja pestrost molekularnih funkcij proteinov. Za obe rastlini pa velja, da ima mnogo proteinov več različnih molekularnih vlog (na primer: kinaza ima istočasno tranferazno aktivnost, veže pa tudi določeno molekulo).

4 SKLEPI

EST kloni so primerno orodje za analizo izražanja genov v nekem vzorcu, iskanje novih genov ter raziskovanje alternativnega izrezovanja intronov. Njihovo pridobivanje je relativno cenovno ugodno in enostavno, vendar pa uporabo EST v zadnjem času nadomeščajo nove tehnike masovnega paralelnega sekvenciranja RNA (angl. RNA-seq). V raziskavi smo analizirali 108 EST zaporedij križanca *C. arabica* X *C. canephora*, ter 145 EST zaporedij križanca *C. canephora* X *C. congensis* iz baze dbEST. Po uporabi BLAST algoritma smo pri prvem križancu našli 35 zadetkov, ki predstavljajo proteine z znano funkcijo, pri drugem križancu pa je bilo takih zadetkov 77. Večina najdenih proteinov je locirana v citoplazmi, jedru in membrani. Najbolj pogoste molekularne funkcije, ki jih opravljajo identificirani proteini, so vezava nukleotidov, vezava ionov ter oksidoreduktazne funkcije. Na podlagi analize je razvidno, da sta si križanca različna tako v številu EST klonov najdenih v bazi, kot tudi v karakteristiki proteinskih zaporedij, ki jih EST kloni kodirajo.

5 ZAHVALA

Večji del izdelka je bil pripravljen kot seminarska naloga pod vodstvom prof. dr. Gregorja Anderluha, prof. dr. Blaža Zupana in prof. dr. Uroša Petroviča pri predmetu Bioinformatika na doktorskem študiju Biomedicina, smer Genetika (Tina Svetek) in doktorskem študiju Bioznanost, smer Biologija (Nataša Šibanc).

6 VIRI

Adams M. D., Kelley J. M., Gocayne J. G., Dubnick M., Polymeropoulos M. H., Xiam H., Merril C.R., Wu A., Olde B., Moreno R. F., Kerlavage A.R., McMombie R., Venter J.C. 1991. Complementary DNA sequencing: Expressed Sequence Tags and Human Genome Project. Science, 252: 1651-1656.

Aaronson J. S., Eckman B., Blevins R.A., Borkowski J.A., Myerson J., Imran S., Elliston K.O. 1996. Toward the development of a gene index to the human genome: an assessment of the nature of high-throughput EST sequence data. Genome Research, 6 (9): 829-45.

Altschul S. F., Madden T. L., Schaffer A. A., Zhang J., Zhang Z., Miller W. Lipman D. J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research, 25: 3389–402.

Dias Neto E., Correa R. G., Verjovski-Almeida S., Briones M. R., Nagai M. A., da Silva W. Jr., Zago M. A., Bordin S., Costa F. F., Goldman G. H., Carvalho A. F., Matsukuma A., Baia G. S., Simpson D. H., Brunstein A., de Oliveira P. S., Bucher P., Jongeneel C. V., O'Hare M. J., Soares F., Brentani R. R., Reis L. F., de Souza S. J., Simpson A. J. 2000. Shotgun sequencing of the human transcriptome with ORF expressed sequence tags. Proceedings of the National Academy of Sciences USA, 97 (7): 3491-3496.

Ewing B., Green P. 1998. Base - calling of automated sequencer traces using phred. II. Error probabilities. Genome Research, 8: 186–194.

Huang X., Madan A. 1999. CAP3: a DNA sequence assembly program. Genome Research, 9: 868–877.

Lee Y., Tsai J., Sunkara S., Karamycheva S., Pertea G., Sultana R., Antonescu V.,

Chan A., Cheung F., Quackenbush J. 2005. The TIGR Gene Indices: clustering and assembling EST and known genes and integration with eukaryotic genomes. Nucleic Acids Research, 33: D71–4.

Liang F., Holt I., Pertea G., Karamycheva S., Salzberg S. L., Quackenbush J. 2000. An optimized protocol for analysis of EST sequence. Nucleic Acids Research, 28 (18): 3657-3665.

Ozsolak F., Platt A. R., Jones D. R., Reifenberger J. G., Sass L. E., McInerney P., Thompson J. F., Bowers J., Jarosz M., Milos J. M. 2009. Direct RNA sequencing. Nature Letters, 461: 814 – 819.

Ptitsyn A. in Hide W. 2005. CLU: a new algorithm for EST clustering. BMC Bioinformatics, 6: S3.