

UNIVERZA V LJUBLJANI
BIOTEHNIŠKA FAKULTETA

Živa RAMŠAK

**PRISTOPI K ANALIZI RAZLIČNIH TIPOV
PODATKOV ZA BOLJŠE RAZUMEVANJE
OBRAMBNEGA ODZIVA KROMPIRJA (*Solanum
tuberosum* L.) NA VIRUSNE OKUŽBE**

DOKTORSKA DISERTACIJA

Ljubljana, 2015

UNIVERZA V LJUBLJANI
BIOTEHNIŠKA FAKULTETA

Živa RAMŠAK

**PRISTOPI K ANALIZI RAZLIČNIH TIPOV PODATKOV ZA
BOLJŠE RAZUMEVANJE OBRAMBNEGA ODZIVA
KROMPIRJA (*Solanum tuberosum* L.) NA VIRUSNE OKUŽBE**

DOKTORSKA DISERTACIJA

**APPROACHES TO ANALYSIS OF DIVERSE DATA FOR
INCREASED UNDERSTANDING OF THE POTATO (*Solanum
tuberosum* L.) DEFENSE RESPONSE TO VIRAL INFECTION**

DOCTORAL DISSERTATION

Ljubljana, 2015

Moderately $J = 130$

Cream – Crossroads (Wheels of Fire, 1968)
Words & Music by Robert Johnson

The musical score consists of four staves of music. Staff A (Intro) starts with eighth-note chords and leads into a more complex rhythmic pattern with sixteenth-note figures. Staff B follows with a similar pattern. Staff C continues the rhythmic style. Staff D concludes the section with a series of eighth-note chords.

I went down to the *plant hormone* crossroads, fell down on my knees.

Down to the *plant hormone* crossroads, fell down on my knees.

Asked the Lord above *of Science* for mercy, "Save me if you please."

Explain the inner workings of the crosstalk to me...

I went down to the *plant hormone* crossroads, tried to ~~flag a ride~~ ...understand

Down to the *plant hormone* crossroads, tried to ~~flag a ride~~ *the processes*

~~Nobody seemed to know me~~, everybody passed me by.

None seemed to be too well defined...

Well I'm going down to *workshops Rosedale*, take my rider by my side.

Going down to *workshops Rosedale*, take my rider by my side.

You can still barrelhouse, baby, on the riverside.

You can run, you can run, *but you can't hide from plant*

Run, you can run, ... *hormone interconnectedness*.

And I'm staying at the crossroads, believe I'm sinking down,

... not so much anymore ...

Na podlagi Statuta Univerze v Ljubljani ter po sklepu Senata Biotehniške fakultete in sklepa 42. seje Komisije za doktorski študij UL z dne 25. 9. 2013 (po pooblastilu Senata Univerze z dne 20. 1. 2009) je bilo potrjeno, da kandidatka izpolnjuje pogoje za opravljanje doktorata znanosti na Interdisciplinarnem doktorskem študijskem programu Bioznanosti, znanstveno področje: bioinformatika. Za mentorico je bila imenovana prof. dr. Kristina Gruden.

Doktorsko delo je bilo opravljeno na Oddelku za biotehnologijo in sistemsko biologijo Nacionalnega inštituta za biologijo v Ljubljani, v sodelovanju z Odsekom za tehnologije znanja Instituta »Jožef Stefan«.

Komisija za oceno in zagovor:

Predsednik: prof. dr. Jernej JAKŠE
Univerza v Ljubljani, Biotehniška fakulteta, Oddelek za agronomijo

Članica: prof. dr. Tanja KUNEJ
Univerza v Ljubljani, Biotehniška fakulteta, Oddelek za zootehniko

Član: doc. dr. Tomaž CURK
Univerza v Ljubljani, Fakulteta za računalništvo in informatiko

Datum zagovora: 26. 11. 2015

Podpisana izjavljam, da je disertacija rezultat lastnega raziskovalnega dela. Izjavljam, da je elektronski izvod identičen tiskanemu. Na univerzo neodplačno, neizključno, prostorsko in časovno neomejeno prenašam pravici shranitve avtorskega dela v elektronski obliki in reproduciranja ter pravico omogočanja javnega dostopa do avtorskega dela na svetovnem spletu preko Digitalne knjižnice Biotehniške fakultete.

Živa Ramšak

KLJUČNA DOKUMENTACIJSKA INFORMACIJA

ŠD	Du3
DK	UDK 57:633.491:632(043.3)=163.6
KG	genska ontologija/ko-ekspresijske genske mreže/krompir/krompirjev virus Y/MapMan/matematično modeliranje/pristopi sistemsko biologije/ <i>Solanum tuberosum</i> /virusna okužba
AV	RAMŠAK, Živa, univ. dipl. bioteh.
SA	GRUDEN, Kristina (mentor)
KZ	SI-1000 Ljubljana, Jamnikarjeva 101
ZA	Univerza v Ljubljani, Biotehniška fakulteta, Interdisciplinarni doktorski študij Bioznanosti, področje bioinformatika
LI	2015
IN	PRISTOPI K ANALIZI RAZLIČNIH TIPOV PODATKOV ZA BOLJŠE RAZUMEVANJE OBRAMBNEGA ODZIVA KROMPIRJA (<i>Solanum tuberosum</i> L.) NA VIRUSNE OKUŽBE
TD	Doktorska disertacija
OP	XIV, 166 str., 25 pregl., 52 sl., 2 pril., 314 vir.
IJ	sl
JI	sl/en
AI	S pristopi sistemsko biologije smo analizirali kompleksni odgovor rastline krompirja (<i>Solanum tuberosum</i> L.) na okužbo s krompirjevim virusom Y (PVY). Najprej smo izboljšali funkcionalne opise genov in s tem povezano pripravili pan-genom krompirja po lastni razviti metodologiji združevanja nukleotidnih zaporedij v paralogne skupine. Informacijo vsebovano v pan-genomu krompirja smo združili z informacijo modelne rastline navadnega repnjakovca (<i>Arabidopsis thaliana</i>) in nekaterimi kulturno pomembnimi rastlinskimi vrstami v aplikaciji GoMapMan, katera omogoča lažje usklajevanje informacije med rastlinami in širjenje le-te med raziskovalnimi skupinami. Razširili smo model rastlinskega obrambnega odgovora na patogen, katerega smo analizirali z uporabo pristopov analize omrežij, to je z iskanjem pomembnih žarišč omrežja in funkcionalnih modulov ter analize kratkih poti med točkami. Na podlagi časovne serije transkriptomskih podatkov smo pripravili ko-ekspresijska genska omrežja in razvili sistem za združevanje le-teh z omrežjem znanja. Končno veliko omrežje smo primerjali s povezavami v nadgrajenem modelu rastlinske obrambe, kjer smo se osredotočili na iskanje novih komponent in povezav pomembnih za interakcijo krompirja s krompirjevim virusom Y. Na podlagi dobljenih rezultatov bomo lahko v prihodnje osnovali nove hipoteze, kjer bomo lahko s primerno zasnovanimi poskusi bolje razjasnili koncepte medsebojnega pogovora med rastlinskimi hormoni, ki so udeleženi v delovanju omrežja rastlinske imunske signalizacije.

KEY WORDS DOCUMENTATION

ND Du3
DC UDK 57:633.491:632(043.3)=163.6
CX gene ontology/co-expression gene networks/potato/potato virus Y/MapMan/mathematical modeling/systems biology/*Solanum tuberosum*/viral infection
AU RAMŠAK, Živa, BSc.
AA GRUDEN, Kristina (mentor)
PP SI-1000 Ljubljana, Jamnikarjeva 101
PB University of Ljubljana, Biotechnical Faculty, Interdisciplinary Doctoral Programme in Biosciences, scientific field Bioinformatics
PY 2015
TI APPROACHES TO ANALYSIS OF DIVERSE DATA FOR INCREASED UNDERSTANDING OF THE POTATO (*Solanum tuberosum* L.) DEFENSE RESPONSE TO VIRAL INFECTION
DT Doctoral Dissertation
NO XIV, 166 p., 25 tab., 52 fig., 2 ann., 314 ref.
LA sl
AL sl/en
AB Systems biology approaches were used to analyse the complex response of the potato plant (*Solanum tuberosum* L.) to potato virus Y (PVY) infection. First we improved the potato gene functional annotations, which included the construction of its pan-genome using a self-developed methodology, where nucleotide sequences are combined into paralogous groups. The information contained in the potato pan-genome was combined with information on the model plant *Arabidopsis thaliana* and some culturally important plant species in the application GoMapMan, which enables easier consolidation of information between species and its easy dissemination to other research groups. We expanded the model of plant immune response to a pathogen, which we then analysed using network analysis approaches, i.e. searching for important hub nodes, functional modules of the network and short paths. Co-expression gene networks were prepared on the basis of our own transcriptomics time series data and then further combined with the knowledge network. The resulting network was compared with relations of the expanded model, where we mainly focused on exploring new components and connections important for the potato-PVY interaction. These analyses will lead to formation of future hypotheses, where the results of well defined experiments will offer us new insights into important concepts of plant hormone cross-talk and inner workings of plant immune signalization.

KAZALO VSEBINE

KLJUČNA DOKUMENTACIJSKA INFORMACIJA	III
KEY WORDS DOCUMENTATION	IV
KAZALO VSEBINE	V
KAZALO PREGLEDNIC	VIII
KAZALO SLIK	X
KAZALO PRILOG	XII
SLOVARČEK (POJMI, OKRAJŠAVE, SIMBOLI)	XIII
1 UVOD	1
1.1 CILJI IN HIPOTEZE DOKTORSKE NALOGE	3
2 PREGLED OBJAV	5
2.1 OPAZOVANI BIOLOŠKI SISTEM	5
2.1.1 Krompir (<i>Solanum tuberosum L.</i>)	5
2.1.2 Krompirjev virus Y (PVY)	7
2.1.3 Rastlinski obrambni sistem.....	10
2.1.3.1 Splošne signalne poti odziva rastline na okužbo	12
2.1.3.2 Signalne poti odziva rastline ob virusni okužbi	13
2.2 PRISTOPI SISTEMSKE BIOLOGIJE.....	14
2.2.1 Analiza bioloških omrežij	16
2.2.1.1 Topološka analiza omrežij.....	17
2.2.1.2 Pogosto najdeni motivi signalizacijskih omrežij	20
2.2.2 Integracija predhodnega znanja z eksperimentalnimi podatki	22
2.2.3 Dinamično modeliranje sistemov	25
2.2.3.1 Tipi dinamičnih modelov signalizacijskih omrežij	27
2.2.3.2 Izgradnja dinamičnega matematičnega modela.....	29
2.3 JAVNO DOSTOPNO ZNANJE V RASTLINSKI BIOLOGIJI	30
2.3.1 Eksperimentalni podatki.....	30
2.3.2 Ontologije in ortologne skupine	32
3 MATERIALI IN METODE	34
3.1 PRIPRAVA TRANSKRIPTOMSKIH PODATKOV	34
3.1.1.1 Predprocesiranje podatkov	35
3.1.1.2 Priprava linearneg modela ter primerjave med različnimi pogoji	37
3.2 INTEGRACIJA ZAPOREDIJ V OKVIRU ENE RASTLINSKE VRSTE....	37
3.2.1 Priprava pan-genoma krompirja	37
3.2.1.1 Zmanjšanje redundance v množicah podatkov.....	38
3.2.1.2 Primerjava zmanjšanih množic zaporedij.....	39
3.2.1.3 Zmanjšanje števila genskih skupkov preko učenja odločitvenega drevesa....	39

3.2.1.4	Priprava končnih paralognih skupin v programu Pajek	40
3.2.2	Priprava pan-transkriptoma tobaka in povezovanje z mikromrežnimi zasnovami	42
3.3	GOMAPMAN: POVEZOVANJE ZNANJA MED RASTLINSKIMI VRSTAMI	43
3.3.1	Glavni gradniki podatkovne baze GoMapMan	45
3.3.1.1	Ontologija MapMan	45
3.3.1.2	Rastlinske vrste.....	45
3.3.1.3	Ortologne skupine	46
3.3.1.4	Povezovanje z znanjem v drugih javnih bazah podatkov	48
3.3.2	Funkcionalnosti aplikacije GoMapMan	48
3.3.2.1	Proces avtomatske konsolidacije genskih funkcionalnih pripisov	48
3.3.2.2	Ročno popravljanje genskih funkcionalnih opisov	51
3.3.2.3	Povezovanje z zunanjimi aplikacijami	51
3.4	INTEGRACIJA HETEROGENIH TIPOV PODATKOV	52
3.4.1	Nadgradnja strukturnega modela rastlinske obrambe.....	53
3.4.1.1	Nadgradnja modela imunske signalizacije rastline navadni repnjakovec	53
3.4.1.2	Integracija javno dostopnega znanja.....	55
3.4.1.3	Priprava omrežja imunske signalizacije krompirja	56
3.4.2	Priprava ko-ekspresijskih omrežij.....	57
3.4.2.1	Priprava ko-ekspresijskih omrežij v programu BioLayout.....	57
3.4.2.2	Priprava tarčnih ko-ekspresijskih omrežij	58
3.4.3	Povezovanje omrežja imunske signalizacije podatki eksperimentov	59
3.4.4	Analiza pripravljenih omrežij	60
3.4.4.1	Nenadzorovana analiza omrežij.....	60
3.4.4.2	Izboljšava modela rastlinske obrambe.....	60
3.4.4.3	Analiza biološke hipoteze.....	61
3.5	RAZVOJ DINAMIČNEGA MODELA RASTLINSKE OBRAMBE	62
3.5.1	Predpriprava.....	62
3.5.2	Koraki v valu Bayesove linearne metodologije parametrizacije	64
3.5.2.1	Vzorčenje z latinsko hiperocko in numerična integracija	64
3.5.2.2	Priprava linearnih modelov in emulatorja	65
3.5.2.3	Izračun neverjetnosti in izbor ne-neverjetnih kombinacij parametrov	66
4	REZULTATI.....	67
4.1	STATISTIČNA ANALIZA TRANSKRIPTOMSKIH PODATKOV	67
4.2	INTEGRACIJA ZAPOREDIJ V OKVIRU ENE RASTLINSKE VRSTE....	73
4.2.1	Pan-genom krompirja	73
4.2.2	Pan-transkriptom tobaka.....	77

4.3	GOMAPMAN: POVEZOVANJE ZNANJA MED RASTLINSKIMI VRSTAMI	78
4.3.1	Implementacija podatkovne baze GoMapMan	79
4.3.1.1	Popravljanje ontologije MapMan	80
4.3.1.2	Vnos rastlinskih vrst	82
4.3.1.3	Vnos ortognih skupin	83
4.3.1.4	Povezovanje z znanjem v drugih javnih bazah podatkov	86
4.3.2	Funkcionalnosti aplikacije GoMapMan.....	87
4.3.2.1	Avtomatski popravki preko konsolidacije genskih funkcionalnih anotacij....	88
4.3.2.2	Ročno popravljanje funkcionalnih opisov genov v ontologiji.....	90
4.4	INTEGRACIJA HETEROGENIH TIPOV PODATKOV	93
4.4.1	Nadgradnja strukturnega modela rastlinske obrambe.....	94
4.4.1.1	Priprava modela imunske signalizacije navadnega repnjakovca.....	94
4.4.1.2	Integracija javno dostopnega znanja.....	96
4.4.1.3	Priprava omrežij rastlinske imunske signalizacije.....	99
4.4.2	Priprava ko-ekspresijskih omrežij.....	102
4.4.3	Združitev transkriptomskih podatkov z omrežjem PIS	105
4.4.4	Analiza pripravljenih omrežij	108
4.4.4.1	Analiza vplivov virusnih proteinov na rastlinsko imunsko signalizacijo....	108
4.4.4.2	Izboljšava modela rastlinske obrambe.....	109
4.4.4.3	Analiza biološke hipoteze.....	110
4.5	RAZVOJ MATEMATIČNEGA MODELA RASTLINSKE OBRAMBE ..	114
5	RAZPRAVA.....	120
5.1	INTEGRACIJA PODATKOV ZAPOREDIJ GENOMOV OZ. TRANSKRIPTOMOV	121
5.2	APLIKACIJA GOMAPMAN IN NJENE PREDNOSTI.....	124
5.3	INTEGRACIJA JAVNO DOSTOPNEGA IN EKSPERIMENTALNEGA ZNANJA.....	126
5.4	TVORBA HIPOTEZ IN DINAMIČNO MODELIRANJE.....	130
6	SKLEPI	132
7	POVZETEK (SUMMARY)	135
7.1	POVZETEK.....	135
7.2	SUMMARY.....	137
8	VIRI	139
	ZAHVALA	
	PRILOGE	

KAZALO PREGLEDNIC

Pregl. 1: Tipi bioloških omrežij in pomen točk ali povezav v njih.	17
Pregl. 2: Rastlinske vrste vključene v aplikaciji GoMapMan.	45
Pregl. 3: Transkriptomi in mikromrežne zasnove v aplikaciji GoMapMan.	46
Pregl. 4: Seznam dokaznih oznak kot jih uporablja Gene Ontology (GO).	51
Pregl. 5: Pregled podatkov iz literature v velikem omrežju znanja navadnega repnjakovca.	55
Pregl. 6: Seznam izbranih genov za iskanje najkrajših poti in sprehodov.	61
Pregl. 7: Začetne vrednosti za 12 komponent etilenskega podmodela.	64
Pregl. 8: Število diferencialno izraženih genov pri obeh tipih normalizacije in njunem preseku.	72
Pregl. 9: Zmanjšanje redundancy zaporedij po uporabi algoritma CD-HIT na posebej genskih modelih in posebej množicah zaporedij unigene.	73
Pregl. 10: Rezultati validacije treh algoritmov učenja odločitvenih dreves.	74
Pregl. 11: Primer paralognne skupine za etilensko odzivni faktor 5 (ERF5).	75
Pregl. 12: Izboljšava strukture ontologije MapMan za biotski stres.	81
Pregl. 13: Število proteinsko kodirajočih genov za vsako rastlino in njihova vključenost v katerokoli od ortolognih skupin.	85
Pregl. 14: Izboljšava funkcionalnih opisov genov različnih rastlinskih vrst s pomočjo konsolidacije znanja v aplikaciji GoMapMan.	90
Pregl. 15: Pregled velikost manjših podomrežij, pridobljenih iz posameznih podatkovnih virov.	97
Pregl. 16: Pokritje genov navadnega repnjakovca in/ali krompirja glede na ortologni tip.	99
Pregl. 17: Število prevedenih genov v velikem omrežju znanja iz navadnega repnjakovca na krompir je odvisno od izbranega ortolognega tipa.	100
Pregl. 18: Velikosti omrežij znanja po abstrakciji na ortologne skupine in nato razširitvi na posamezne gene.	100
Pregl. 19: Pregled povezav rangiranih omrežij PIS glede na izvorni seznam iz modela rastlinske imunske signalizacije.	101
Pregl. 20: Število ortolognih skupin v rangiranih podomrežjih PIS s pripisanimi mikromrežnimi sondami.	102
Pregl. 21: Velikosti omrežij po združevanju velikega omrežja znanja z rezultati ko-ekspresijske analize.	105
Pregl. 22: Pokritje povezav preko različnih virov podatkov glede na pripravljen »zlati standard«.	107
Pregl. 23: Rezultati TRANSFAC analize zaporedja promotorske regije gena <i>npr1</i>	113

Pregl. 24: PlantCare analiza zaporedja promotorske regije gena <i>npr1</i>	113
Pregl. 25: Zmanjšanje števila kombinacij parametrov v odvisnosti od izbora linearnih modelov.....	117

KAZALO SLIK

Sl. 1: Shema Potyvirusnega genoma in prevedenih proteinov (Ivanov in sod., 2014: 3)	8
Sl. 2: Simptomi okužbe, ki jih na rastlini krompirja lahko povzroča virus PVY (Karasev in Gray, 2013: 2).....	9
Sl. 3: Modela tekmovanja med patogeni/virusi in rastlinami (Nakahara in Masuta, 2014: 3)	11
Sl. 4: Pristop sistemске biologije (Gutiérrez in sod., 2005: 2).....	15
Sl. 5: Topološka organizacija omrežij (Windram in sod., 2014: 4)	18
Sl. 6: Pogosti motivi v omrežjih (Muraro in sod., 2012: 9)	20
Sl. 7: Pristopi inference funkcionalnih interakcij in rekonstrukcije omrežij (Moreno-Risueno in sod., 2010): 2).....	23
Sl. 8: Eksperimentalni načrt proučevanja dinamike tolerančne interakcije med krompirjem in virusom.	34
Sl. 9: Koraki združevanja večih različnih množic zaporedij v skupne paralogne genske skupine.....	38
Sl. 10: Izbrane točke uporabljenе v enačbi izračuna uteži omrežja genov krompirja.....	41
Sl. 11: Relacijski diagram (ER) za bazo <i>gomapman</i>	44
Sl. 12: Predstavitev tvorbe šibkih komponent iz treh nizov ortolognih skupkov.	47
Sl. 13: Shema niza konsolidacijskih odločitvenih pravil za izboljšavo genskih funkcionalnih pripisov implementirana v aplikaciji GoMapMan.....	49
Sl. 14: Ontologija reakcij modela rastlinske imunske signalizacije (PIS).	54
Sl. 15: Taksonomija komponent modela rastlinske imunske signalizacije (PIS).	54
Sl. 16: Priprava podomrežij rastlinske imunske signalizacije preko določanja bližnjih sosedov.	57
Sl. 17: Etilenski podmodel uporabljen v pristopu Bayesove linearne metodologije.	62
Sl. 18: Primer priprave navadnih diferencialnih enačb za preprost sistem, ki vključuje aktivacijo in represijo genov (Karlebach in Shamir, 2008: 2).....	63
Sl. 19: Sistem definiranih navadnih diferencialnih enačb za etilenski podmodel.....	63
Sl. 20: Shema dela predstavlajoč naloge opravljene v doktorskem delu.	67
Sl. 21: Koraki kvalitete kontrole pred normalizacijo.	68
Sl. 22: Primerjava normalizacij s prikazom okvirjev z ročaji ter gostotnimi diagrami. ..	70
Sl. 23: Rezultat razvrščanja podatkov mikromrež v obliki dendrograma.	70
Sl. 24: Toplotni graf podatkov mikromrež.....	71
Sl. 25: Predstavitev paralogne skupine genov krompirja.	76
Sl. 26: Predstavitev paralogne skupine transkriptov tobaka.	77
Sl. 27: Shematska predstavitev aplikacije GoMapMan in njenih funkcionalnosti.	78
Sl. 28: Pogled ontološkega drevesa aplikacije GoMapMan.....	79

Sl. 29: Pogled informacije o posameznem genu v aplikaciji GoMapMan.....	83
Sl. 30: Pogled informacije o ortologni skupini v aplikaciji GoMapMan.....	84
Sl. 31: Porazdelitev velikosti šibkih komponent pridobljenih iz treh ortolognih nizov...	86
Sl. 32: Prikaz povezav na zunanje baze podatkov v aplikaciji GoMapMan.....	87
Sl. 33: Ponazoritev postopka konsolidacije primera ortologne skupine.	88
Sl. 34: Primer genov v ortologni skupini in njihovih funkcionalnih anotacij.	89
Sl. 35: Informacija o zgodovini gena v ontologiji MapMan, kot jo vidi prijavljen uporabnik.	91
Sl. 36: Vmesnik za vnašanje ročnih popravkov aplikacije GoMapMan.	92
Sl. 37: Koraki nadgradnje strukturnega modela rastlinske imunske signalizacije in njegove analize.	93
Sl. 38: Nadgrajen model rastlinske imunske signalizacije.	95
Sl. 39: Uporabljeni podatki za izgradnjo velikega omrežja znanja navadnega repnjakovca.	96
Sl. 40: Vennov diagram pokritja genov vsebovanih v velikem omrežju navadnega repnjakovca.....	98
Sl. 41: Vizualizacija zgrajenega velikega omrežja znanja navadnega repnjakovca.	98
Sl. 42: Tri omrežja programa BioLayout izhajajoč iz seznama sond omrežja StPIS-g1.	103
Sl. 43: Končno tarčno ko-ekspresijsko omrežje izluščeno preko uporabe etilenskih genov kot vabe na eksperimentalnih podatkih mikromrež.	104
Sl. 44: Interakcije med virusnimi in gostiteljivimi proteini.	108
Sl. 45: Izbrana skupina MCODE za nadaljne <i>in silico</i> analize.	110
Sl. 46: Vse najkrajše poti med vozlišči <i>etr1</i> (pot etilena) in <i>npr1</i> (pot salicilne kisline).	111
Sl. 47: Zaporedje protomorske regije (2,000 nk) pred začetkom gena <i>npr1</i> iz genoma krompirja.....	112
Sl. 48: Porazdelitev parametrov v analizi sprememb izražanja genov kontrolnih vzorcev v zgrajeni latinski hiperkocki za pet parametrov sinteze.	114
Sl. 49: Rezultati numerične integracije za 12 komponent etilenskega podmodela za spremembe izražanja genov v kontrolnih listih.	115
Sl. 50: Graf vrednosti rezultatov linearnih modelov glede na vrednosti parametrov za spremembe izražanja genov v kontrolnih listih.	116
Sl. 51: Rezultati zmanjšanja prostora parametrov po prvem valu Bayesove linearne metodologije.	118
Sl. 52: Približanje krivulj simulacij dejanskemu obnašanju molekule za ERF v drugem valu.	119

KAZALO PRILOG

Priloga A: Seznam vseh reakcij nadgrajenega modela rastlinske imunske signalizacije (PIS).

Priloga B: Seznam novih komponent v nadgrajenem modelu rastlinske imunske signalizacije (PIS).

SLOVARČEK (POJMI, OKRAJŠAVE, SIMBOLI)

ABA	abscizinska kislina (<i>abscisic acid</i>)
AUX	avksini (<i>auxins</i>)
BS	brazinosteroidi (<i>brassinosteroids</i>)
CDS	kodirajoče zaporedje gena (<i>coding DNA sequence</i>)
CK	citokinini (<i>cytokinins</i>)
CSML	format zapisa bioloških poti CSML (<i>Cell System Markup Language</i>)
DAMP	gostiteljeve molekule povezane poškodbe (<i>damage associated molecular patterns</i>)
DE	diferencialna enačba (<i>differential equation</i>)
DIG	diferencialno izražen gen (<i>differentially expressed gene</i>)
DM	dvojni monoploid (<i>doubled-monoploid</i>)
dsRNA	dvojno-verižna molekula RNA (<i>double-stranded RNA</i>)
EST	oznaka izraženega zaporedja (<i>expressed sequence tag</i>)
ET	etilen (<i>ethylene</i>)
ETI	od efektorjev sprožen obrambni odgovor (<i>effector-triggered immunity</i>)
FDR	delež napačnih uvrstitev (<i>false discovery rate</i>)
GA	giberelinska kislina (<i>gibberellic acid</i>)
GCN	gensko korelacijsko omrežje (<i>gene correlation network</i>)
GO	genska ontologija (<i>Gene Ontology</i>)
HR	hipersenzitivni odgovor (<i>hypersensitive response</i>)
ITAG	mednarodna skupina za gensko anotacijo paradižnika (<i>International Tomato Annotation Group</i>)
JA	jasmonska kislina (<i>jasmonic acid</i>)
LH	latinska hiperkocka (<i>latin hypercube</i>)
LHS	vzorčenje z latinsko hiperkocko (<i>latin hypercube sampling</i>)
LLC	največja povezana komponenta (<i>largest connected component</i>)
logFC	logaritem razmerja spremembe (<i>log fold change</i>)
MAMP	z mikroboom povezani molekularni vzorci (<i>microbe-associated molecular patterns</i>)
MAPK	od mitogena aktivirana proteinska kinaza (<i>mitogen-activated protein kinase</i>)
MCL	algoritma razvrščanja po Markovu (<i>Markov Cluster algorithm</i>)
NGS	sekvenciranje naslednje generacije (<i>next generation sequencing</i>)
ODE	navadna diferencialna enačba (<i>ordinary differential equation</i>)
PAMP	s patogenom povezani molekularni vzorci (<i>pathogen-associated molecular patterns</i>)
PGSC	konzorcij določanja zaporedja genoma krompirja (<i>Potato Genome Sequencing Consortium</i>)
PO	rastlinska ontologija PO (<i>Plant Ontology</i>)

POCI	pobuda za pripravo mikromreže krompirja (<i>Potato Oligo Chip Initiative</i>)
PR-protein	protein povezan s patogenezo (<i>pathogenesis-related protein</i>)
PRR	membranski rastlinski receptor (<i>plant pattern recognition receptor</i>)
PTI	od PAMP sprožen obrambni odgovor (PAMP-triggered immunity)
PTNRD	obročasta nekroza gomoljev (<i>potato tuber necrotic ringspot disease</i>)
PUT	sestavljeni unikatni transkripti skupine PlantGDB (<i>PlantGDB-assembled unique transcripts</i>)
PVY	krompirjev virus Y (<i>potato virus Y</i>)
ROS	reakтивne kisikove vrste (<i>reactive oxygen species</i>)
R-protein	receptorju podobni protein (<i>receptor-like protein</i>)
RSD	algoritem recipročne najmanjše razdalje (<i>reciprocal smallest distance</i>)
RSN	robustna normalizacija (<i>robust spline normalization</i>)
RSS	supresorji utišanja RNA (<i>RNA silencing suppressors</i>)
SA	salicilna kislina (<i>salicylic acid</i>)
SBML	format zapisa bioloških poti SBML (<i>Systems Biology Markup Language</i>)
SGN	podatkovna baza SOL omrežja (<i>SOL Genomics Network</i>)
siRNA	mala interferenčna molekula RNA (<i>small-interfering RNA</i>)
StGI	genski indeks transkriptov krompirja Inštituta Dana Faber (<i>Solanum tuberosum Gene Index</i>)
TA	genski indeks transkriptov krompirja J. Craig Venter Inštituta (<i>TIGR plant transcript assemblies</i>)
TuMV	virus mozaika repe (<i>turnip mosaic virus</i>)
VSN	normalizacija s stabilizacijo variance (<i>variance stabilising normalisation</i>)
Y2H	dvohibridni sistem kvasovke (<i>yeast two hybrid system</i>)

1 UVOD

Krompir (*Solanum tuberosum* L.) spada poleg riža in pšenice med tri najpomembnejše poljščine na svetu (FAOSTAT, 2013). Poljščina je občutljiva tako na abiotiski kot biotski stres, kjer je pri slednjem od širokega razpona biotskih stresnih dejavnikov krompirjev virus Y (PVY) iz družine Potyviridae eden ekonomsko najbolj pomembnih iz vidika povzročanja škode na pridelku (Scholthof in sod., 2011). Odziv rastline na okužbo z virusom je zelo kompleksen, kljub relativni preprostosti njihove genetske strukture. Kompleksnost se odraža v aktivaciji različnih signalnih poti, katerih natančni mehanizmi še niso dobro definirani.

Na Oddelku za biotehnologijo in sistemsko biologijo Nacionalnega inštituta za biologijo potekajo raziskave o interakciji med krompirjem in PVY na morfološkem in biokemičnem nivoju ter na nivoju izražanja genov. Znano je, da je ustrezен obrambni odgovor pogojen s številnimi medsebojno deluječimi komponentami, vključno s hitrimi spremembami redoks potenciala (Milavec in sod., 2008), sočasnega delovanja številnih rastlinskih hormonov (Kovač in sod., 2009; Krecic-Stres in sod., 2005) in obsežnih sprememb na nivoju izražanja genov (Baebler in sod., 2009; Baebler in sod., 2011; Baebler in sod., 2014; Kogovšek in sod., 2010; Pompe-Novak in sod., 2006).

Spremembra v dogmi znanstvenega biološkega in biomedicinskega raziskovanja, t.j. prestopa iz tradicionalnih reduktionističnih pristopov proti bolj holističnem pogledu na razumevanje delovanja organizmov, je vodila v razmah razvoja novih visoko zmogljivih tehnologij pridobivanja podatkov na vseh nivojih biološke organizacije (omske tehnologije genomike, transkriptomike, proteomike in metabolomike). S ciljem razvoja celovitega modela rastlinskega obrambnega odgovora potrebujemo združevanje vsega obstoječega znanja. Za zagotavljanje novih pogledov v razumevanje kompleksnih bioloških procesov so najbolj uspešni predvsem pristopi sistemsko biologije, ki združujejo znanje večih nivojev hkrati (Keurentjes in sod., 2011; Kitano, 2002). Iz drugačne perspektive lahko z uporabo računalniškega modeliranja za opis in analizo bioloških sistemov simuliramo kompleksno vedenje celičnih signalizacijskih omrežij, kar nam da informacijo o mehanizmih, ki vodijo do razvoja kompleksnih bolezni (Bauer-Mehren in sod., 2009; Courtot in sod., 2011).

Priprave modela rastlinske obrambe so se Miljkovic in sod. (2012) lotili na osnovi znanstvenih objav in eksperimentnega znanja. Model vključuje povezave med biosintetskimi in signalnimi potmi treh hormonov pomembnih v rastlinskem obrambnem odgovoru: jasmonske kisline (JA), salicilne kisline (SA) ter etilena (ET). Na podoben način so nekoliko širši, a manj podroben model rastlinske signalizacije zgradili Naseem in sod. (2012). Dodatne vire znanja predstavljajo rezultati visoko zmogljivih tehnologij, npr. proteinskih interakcij, interakcij med virusnimi in rastlinskimi komponentami ter interakcij med virusu lastnimi komponentami (Arabidopsis Interactome Mapping Consortium, 2011; Elena in sod., 2011; Mukhtar in sod., 2011). Priprava omrežja, ki bo združevalo vse obstoječe znanje, nam bo omogočila iskanje novih lastnosti imunske signalizacije, ki ga sedanje znanje ekspertov ne zajema, na primer preko analize prenosa informacije po omrežju (Elena in Rodrigo, 2012) ali preko iskanja novih poti in podomrežij (González in sod., 2012).

Vzporedno z rastjo razvoja novih tehnologij pridobivanja podatkov, njihovih analiz ter združevanja v formalne matematične opise, raste tudi uporaba ontologij. Ontologije tvori harmoniziran slovar strokovnih izrazov (pojmov), ki na strukturiran način opisujejo domeno interesa (Hartung in sod., 2012). Pripravi in uporablja se jih lahko na vseh nivojih: za opise funkcij genov in razjasnjevanje delovanje najnižjega nivoja biološke organizacije, za opisovanje matematične strukture na formalno tehničen način, pa vse do opisov fiziološko fenotipskih značilnosti organizma. V moderni rastlinski biologiji se za opis funkcije genov najpogosteje uporablja Gene Ontology (GO) (The Gene Ontology Consortium, 2000), FunCat (Ruepp in sod., 2004) in rastlinsko-specifično ontologijo MapMan (Thimm in sod., 2004). Ontologiji GO in FunCat že od začetka obstajata tudi v obliki spletne aplikacije, kar pa ne drži za funkcionalne opise ontologije MapMan. Le-ti so obstajali kot neuskrajena zbirka posameznih datotek (t.i. mapiranj) za vsako rastlinsko vrsto posebej (npr. na oddelku pripravljeni mapiranji za krompir (Rotter in sod., 2007) ali vinsko trto (Rotter in sod., 2009)). To je predstavljalo motivacijo za pripravo spletne aplikacije, ki bo omogočala lažji vpogled v zbrano informacijo za končnega uporabnika ter olajšala izboljšave funkcionalnih pripisov različnih rastlinskih vrst.

1.1 CILJI IN HIPOTEZE DOKTORSKE NALOGE

Namen doktorskega dela je bil poglobiti razumevanje odgovora rastlin na okužbo z virusom, za kar smo se poslužili različnih pristopov pod okriljem sistemsko biološke doktrine. Hkrati smo želeli na osnovi ontologije MapMan razviti aplikacijo GoMapMan, ki bi omogočala lažjo sledljivost in nadgradnjo biološkega znanja, prenos biološkega znanja iz modelne rastline navadnega repnjakovca (*Arabidopsis thaliana*) na agronomsko pomembne poljščine ter bolj kvalitetne meta-analize podatkov visoko zmogljivih tehnologij.

Postavili smo si naslednji hipotezi s povezanimi cilji:

1. Spletna aplikacija osnovana na ontologiji MapMan ter funkcionalnost relacijske baze bosta vodili k bolj natančnim meta-analizam podatkov visoko zmogljivih tehnologij ter preko vključenosti širšega znanstvenega kroga omogočila izboljšavo genskih funkcionalnih pripisov za izbrane rastlinske vrste. Anotacija vsaj ene metabolne ali signalizacijske poti bo s pomočjo tako razvite aplikacije omogočila bolj natančno analizo podatkov visoko zmogljivostnih tehnologij ter bila v pomoč pri razlagi le-teh.
- Zaradi razpršenosti podatkov o krompirjevih genih, predvsem zaporedij, bomo pripravili pan-genom krompirja, ki bo le-ta združil v enotno celoto.
- Pripravljen pan-genom krompirja, gene nekaterih izbranih rastlin in njihove funkcionalne pripise bomo povezali preko informacije vsebovane v ortolognih skupinah različnih algoritmov. Informacijo bomo povezali v spletni aplikaciji.
- Zasnova pripravljene aplikacije bo dovolj prilagodljiva, da bo omogočala tako ročno kot avtomatsko popravljanje funkcionalnih pripisov MapMan za vsebovane gene ter tudi samo popravljanje ontologije MapMan.
- Pripravljena aplikacija bo vodila v bolj natančne meta-analize podatkov visoko zmogljivih tehnologij preko prenosa informacije iz aplikacije v primerno obliko za javno uporabo.

2. Nadgradili bomo že obstoječi model rastlinskega odgovora na okužbo z virusom družine Potyvirus za navadni repnjakovec. Pričakujemo, da bo prenos znanja iz te modelne rastline na krompir ter hkratna integracija z eksperimentalnimi podatki krompirja privedla do robustnega modela, sposobnega generacije novih znanstvenih hipotez. Vsaj eno od tako pridobljenih hipotez bomo v okviru raziskovalne skupine preverili ali v laboratoriju ali *in silico*, na tej podlagi bomo pa lahko vsaj grobo ocenili napovedno vrednost ter kvaliteto zgrajenega modela.
- Komponente in povezave v obstoječega modela rastlinskega odgovora na okužbo bomo primerno dodatno opisali. Ob tej nadgradnji bomo dovolili širjenje modela z novo dognanimi komponentami in povezavami.
 - Trenutno obstoječe znanje v literaturi, predvsem eksperimentalne rezultate visoko zmogljivih tehnologij, bomo primerno obdelali in iz njih pripravili omrežje povezav.
 - Omrežje javno dostopnega znanja bomo združili z lastnimi eksperimentalnimi podatki transkriptomike z uporabo vsaj enega izmed pristopov inference omrežij. Za to bomo predhodno transkriptomske podatke po potrebi primerno statistično obdelali.
 - Pridobljena omrežja bomo analizirali z različnimi pristopi analize omrežij ter eno od generiranih hipotez preizkusili ali *in silico* ali v laboratoriju.

2 PREGLED OBJAV

2.1 OPAZOVANI BIOLOŠKI SISTEM

2.1.1 Krompir (*Solanum tuberosum* L.)

Krompir (*Solanum tuberosum* L.) predstavlja svetovno najbolj pomembno gomoljnico ter tretjo najbolj pomembno poljščino za rižem in pšenico (Khoury in sod., 2014). Kot ugoden vir ogljikovih hidratov, vitaminov in mineralov je pomemben za kmetijstvo zmernih in tropskih podnebnih področij (Karasev in Gray, 2013), pa tudi v industrijskih aplikacijah (Kraak, 1992). Poleg paradižnika, paprike, jajčevca, petunije in tobaka spada v splošno ekonomsko pomembno družino razhudnikovk (*Solanaceae*).

Visok obseg proizvodnje, ki je leta 2013 obsegal 370 milijonov ton, pa ne dosega ocenjenega fiziološkega potenciala, t.j. 120 ton/hektar (FAOSTAT, 2013). Izboljšave krompirja preko uporabe klasičnih pristopov žlahtnenja omejujejo njegova visoka heterozigotnost, avtotetraploidnost ter vegetativni način razmnoževanja (The Potato Genome Sequencing Consortium, 2011). Dodatno je imel (z izjemo njegovega naravnega geografskega razpona v južni Ameriki) kultiviran krompir razmeroma nizek vnos raznolikosti na nivoju genoma (Spooner in Hijmans, 2001).

Poljščina je občutljiva tako na abiotski kot biotski stres. Kljub učinkoviti porabi vode je zaradi skopega in plitvega koreninskega sistema občutljiva na sušo (Yuan in sod., 2003) ter kaže znake zmerne občutljivosti na sol (Katerji in sod., 2000); oboje je še posebej pomembno z obzirom na podnebne spremembe. Od širokega razpona biotskih stresnih dejavnikov so iz vidika izgube pridelka najbolj pomembni koloradski hrošč (*Leptinotarsa decemlineata* L.), oomiceta *Phytophthora infestans* ter krompirjev virus Y (ang. *potato virus Y*; PVY).

Poznavanje genoma krompirja je do leta 2011 temeljilo na informaciji iz transkriptomskih oznak izraženih zaporedij (ang. *expressed sequence tag*; EST). Zaporedja so bila preko uporabe različnih algoritmov združena v genske skupke, shranjene v različnih bazah podatkov, npr. skupki SGN podatkovne baze SOL omrežja (Fernandez-Pozo in sod., 2014), skupki PUT skupine PlantGDB (Duvick in sod., 2008), genski indeks transkriptov TA J. Craig Vender Inštituta (Childs in sod., 2007), genski indeks transkriptov StGI Inštituta Dana Faber (Quackenbush in sod., 2001) ali pa pobude za pripravo mikromreže krompirja (ang. *Potato Oligo Chip Initiative*, POCI) (Kloosterman in sod., 2008).

Visoko kvaliteten osnutek genoma krompirja je leta 2011 pripravil konzorcij določanja zaporedja genoma krompirja (ang. *Potato Genome Sequencing Consortium*, PGSC) (The Potato Genome Sequencing Consortium, 2011). Kot biološki material je bila uporabljena homozigotna sorta krompirja genotipa *S. tuberosum* Phureja, t.i. dvojni monoploid (ang. *doubled monoploid*, DM). Rezultat določanja zapredij genskega materiala dvojnega monoploida je 12 psevdomolekul predstavlajoč delna zaporedja vseh kromosomov krompirja, pokrivajo 86 % celotnega 844-Mb velikega genoma ter več kot 95 % vseh obstoječih EST zaporedij. Na podlagi psevdomolekul konzorcija PGSC sta bila napovedana dva genomska modela – genomski model konzorcija PGSC ter genomski model konzorcija mednarodne skupine za gensko anotacijo paradižnika (ang. *International Tomato Annotation Group*, ITAG) (The Tomato Genome Consortium, 2012). Genski model PGSC šteje 39,031 napovedanih genov (oz. 56,218 transkriptov ob upoštevanju vseh spojitvenih oblik), genski model ITAG pa 35,004 genov (tu je število spojitvenih oblik enako številu genov).

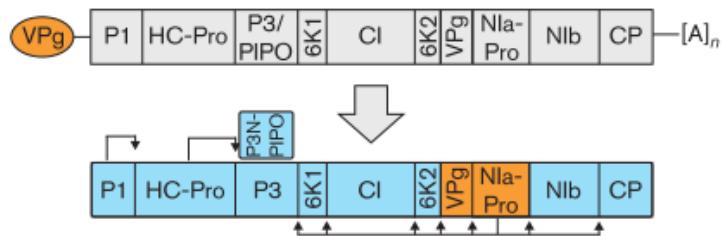
Genomsko zaporedje krompirja lahko tako žlahtiteljem kot znanstvenim skupinam pomaga pri razlagi in reševanju različnih bioloških vprašanj, na primer posledice nizkega vnosa variacije v populacije in s tem povezane depresije, obrazloži evolucijo razvoja krompirju svojstvenega gomolja (sorodne rastline družine razhudnikovk niso razvile sposobnosti tvorbe gomoljev) ter pripomore pri iskanju in karakterizaciji genov udeleženih v odgovoru na različne stresne dejavnike.

2.1.2 Krompirjev virus Y (PVY)

Krompirjev virus Y (PVY) je eden najstarejših znanih rastlinskih virusov, uvrščen v skupino desetih ekonomsko najpomembnejših rastlinskih virusov. V svetovni proizvodnji krompirja lahko povzroči tudi do 90 % izgube pridelka (Scholthof in sod., 2011). Spada v poljedelsko in ekonomsko pomembno družino Potyviridae, ki vključno s Potyvirusi šteje 8 različnih rodov (Ivanov in sod., 2014).

Razpon možnih gostiteljev za PVY je precej širok – šteje devet botaničnih družin, od tega veliko rodov družine razhudnikovk (rastline krompirja, paradižnika, tobaka in paprike), okrasnih rastlin (dalije in petunije) ter plevelov (Quenouille in sod., 2013). Dodatno se PVY v obilju nahaja tudi v divjini (Roossinck, 2011). Okužbo s PVY lahko prenaša več kot 40 vrstami listnih uši družine *Aphididae* (Piron, 1986), prenaša pa se tudi z vegetativno propagacijo gomoljev okuženega krompirja ter v manjši meri z mehanskim prenosom (Quenouille in sod., 2013). Najbolj učinkovit vektor prenosa iz vidika uspešnosti okužbe predstavlja siva breskova uš *Myzus persicae* (Boquel in sod., 2011).

Genom Potyvirus družine (Slika 1) tvori pozitivna (ang. *positive-sense*) enoverižna RNA (Adams in sod., 2005), ki se prevede v poliprotein, iz katerega nastane 10 od vseh 11ih zrelih proteinov (Rajamäki in Valkonen, 2009): tri proteaze (P1, HC-Pro, NIa-Pro), helikaza z ATPazno aktivnostjo (CI), od RNA odvisna RNA polimeraza (NIb), protein kapside (CP) ter蛋白 VPg, P3, 6K1 in 6K2 (Elena in Rodrigo, 2012). Zadnji protein (P3N-PIPO) se posebej tvori iz manjšega odprtrega bralnega okvirja (Chung in sod., 2008). Pet od teh proteinov ima sposobnost translokacije v jedro okužene celice: poleg NIa in NIb sta bila v jedru zaznana tudi proteina CI ter HC-Pro (Riedel in sod., 1998), prisotnost proteina P3 v jedru pa je bila zaznana v primeru okužbe celic z virusom razjed tobaka (ang. *tobacco etch virus*, TEV) (Langenberg in Zhang, 1997).



Slika 1: Shema Potyvirusnega genoma in prevedenih proteinov (Ivanov in sod., 2014: 3)

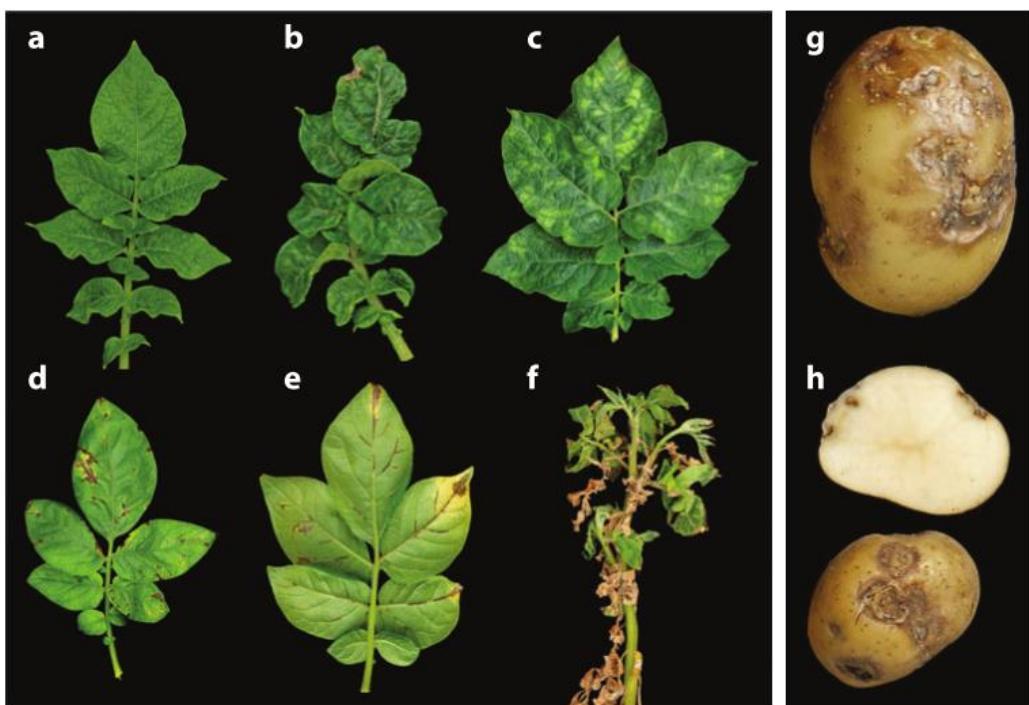
Genom Potyvirusa predstavlja pozitivna enoverižna molekula RNA, kovalentno povezana s proteinom VPg na 5' koncu ter poliadenzilirana na 3' koncu. Prevede se v večji poliprotein (ki se procesira v 10 posameznih proteinov) in en sam P3N-PIPO protein iz manjšega ločenega odprtrega bralnega okvirja. Virusna RNA je prikazana s sivo, prevedni proteini pa s svetlo modro oz. oranžno barvo.

Figure 1: Schematic of the Potyviral genome and its translated proteins (Ivanov in sod., 2014: 3)

Potyviral genome is represented by a positive-sense ssRNA molecule, covalently linked to VPg at its 5' end and polyadenylated at the 3' end. It is translated into a large polyprotein (processed into 10 individual proteins) and a single P3N-PIPO protein from a separate small ORF. Viral RNA is shown in grey and translated proteins in light blue.

Virusne komponente, tako nukleinske kisline kot nastali proteini, tvorijo množične in kompleksne interakcije ne samo med seboj (Guo in sod., 2001; Lin in sod., 2009), ampak tudi s komponentami gostiteljske celice (Bailer in Haas, 2009; Elena in Rodrigo, 2012; Tan in sod., 2007). Virusne okužbe tipično spremenijo fiziologijo gostitelja preko aktivne supresije obrambnih odgovorov gostitelja ter preko sprememb znotrajceličnega okolja v optimum za lastno translacijo, replikacijo, sestavljanje ter širitev (Dodds in Rathjen, 2010; Peng in sod., 2009). Na okužbo gostitelj odgovori s spremembami v izražanju komponent celičnih poti in vklopa specifičnih mehanizmov za zajezitev okužbe. Družno te spremembe določajo tip in jakost simptomov ter virulenco okužbe (Rodrigo in sod., 2012).

Okužba s PVY v krompirju povzroča širok razpon simptomov, ki jih lahko ločimo na bolezni listov in gomoljev (Slika 2). Izgled in jakost simptomov bolezni je odvisna od kultivarja krompirja, seva PVY, okoljskih pogojev ter tipa okužbe (Karasev in Gray, 2013). Bolezenska znamenja na primarno okuženih listih vključujejo nastanek lis, mozaikov in črtastih nekroz ter njihovo gubanje. Na sistemsko okuženih listih, ki se razvijejo nad okuženimi, pride do pojava rumenega mozaika in kodranja listne površine. Nekateri sevi PVY povzročajo bolezen obročkaste nekroze gomoljev (ang. *potato tuber necrotic ringspot disease*, PTNRD), pri kateri prihaja do tvorbe izbočenih obročev ali nabreklin temnejše barve, ki se lahko pri bolj občutljivih sortah širijo tudi pod povrhnjico gomolja (Beczner in sod., 1984).



Slika 2: Simptomi okužbe, ki jih na rastlini krompirja lahko povzroča virus PVY (Karasev in Gray, 2013: 2)

Na levi so predstavljeni simptomi opazni na listih, desno pa bolezen obročkaste nekroze gomoljev (PTNRD). (a) tvorba mozaika, gubanje; (b) raskav mozaik; (c) mozaik s klorozami; sistemsko nekrotske lezije in nekroza žil vidna na (d) zgornji in (e) spodnji površini lista; (f) sistemski nekroza zgornjih listov ter sindrom uvelosti spodnjih listov; (g) napredna PTRND s suhimi udrtimi nekrotičnimi področji; (h) zelo pozna PTRND, kjer so udrta nekrotična področja opazna tudi pod povrnjico gomolja.

Figure 2: Symptoms induced by the PVY virus in potato (Karasev and Gray, 2013: 2)

On the left leaf related symptoms are shown, on the right potato tuber necrotic ringspot disease (PTNRD). (a) mosaic and crinkling; (b) rugose mosaic; (c) mosaic with chlorotic spots; systemic necrotic lesions and vein necrosis of the (d) upper and (e) lower leaf surface; (f) systemic necrosis on top leaves and leaf-drop syndrome on lower leaves; (g) advanced stage of PTNRD with dry sunken necrotic rings; (h) very late stage of PTNRD where the sunken necrotic rings are visible under the tuber skin.

Sorte krompirja lahko ločimo glede na njihovo odpornost na PVY, in sicer občutljive ter odporne. Občutljive sorte dalje lahko ločimo na preobčutljive (sorti Igor in Nadine), kjer pride do razvoja simptomov tako na listih kot gomoljih, ter tolerantne (sorti Pentland squire in Désirée), kjer se virus širi po rastlini, a so simptomi manj opazni (Quenouille in sod., 2013). Med odpornimi sortami ločimo med sortami, pri katerih pride do preobčutljivostnega odgovora (ang. *hypersensitive response*, HR) ali do razvoja ekstremne odpornosti (ang. *extreme resistance*, ER). Oba odgovora zaustavita širjenje virusa v druge dele rastline (Hinrichs in sod., 1998), kjer so rezultat HR odgovora nekroze listov (sorta Rywal), medtem ko v ER odgovoru ne pride do razvoja bolezenskih simptomov (sorte PW363, Sante, Carlingford).

2.1.3 Rastlinski obrambni sistem

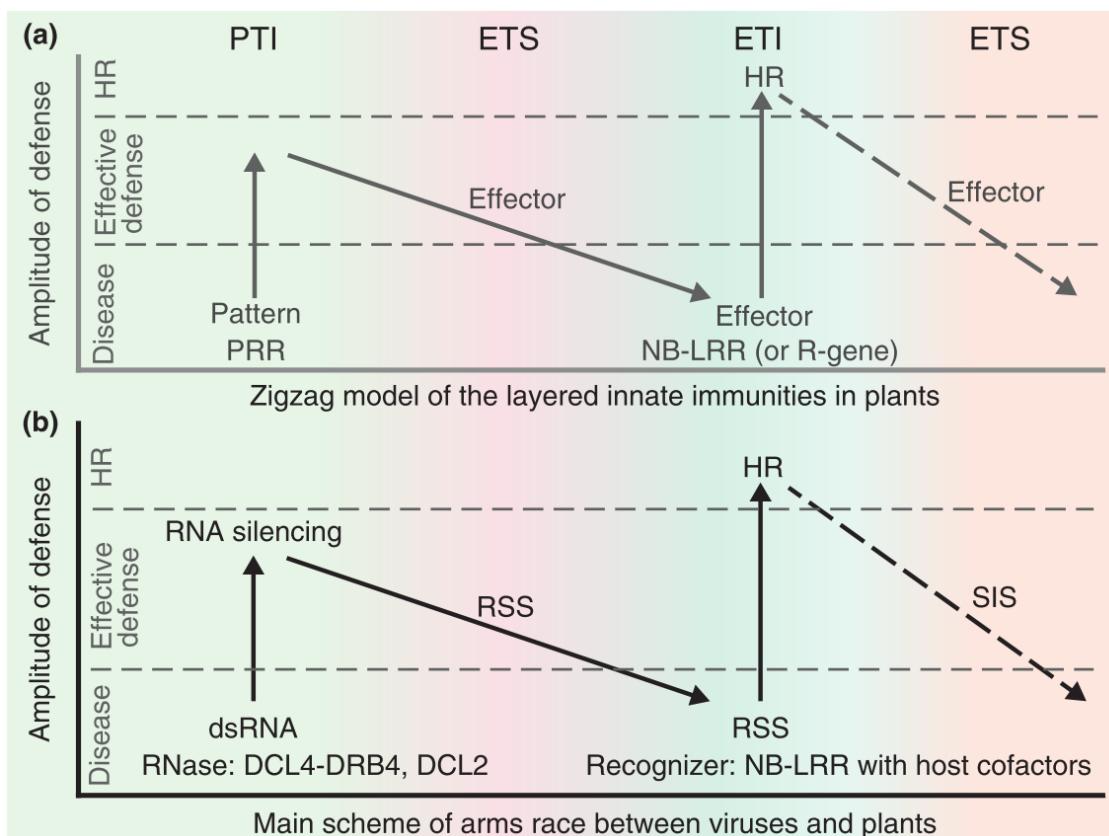
Rastline so ves čas pod vplivom različnih biotskih in abiotskih stresnih dejavnikov. Njihov pritrjeni način življenja je vodil v razvoj natančnega in prefinjenega mehanizma obrambe – obsežnega imunskega signalizacijskega omrežja. V njem se procesi detekcije, prenosa informacije in odgovora večinoma dogajajo v jedru, dodatno pa je omrežje v preseku z ostalimi ključnimi razvojnimi in regulatornimi procesi v celici (Motion in sod., 2015).

Razvoj bolezni je bolj izjema kot pogost pojav, saj so v večini primerov rastline sposobne preprečiti škodljive učinke okužbe. Za varovanje pred patogeni so razvile dve glavni strategiji obrambe: strategijo z receptorjem podobnimi proteini (ang. *receptor-like proteins*, R-proteini) ter strategijo preko utišanja RNA (ang. *RNA silencing*). Prva strategija je učinkovita proti večjemu razponu patogenov (glivam, bakterijam in virusom) in je opisana s t.i. zig-zag modelom (Slika 3a), medtem ko druga predstavlja najpomembnejši protivirusni mehanizem obrambe (Nakahara in Masuta, 2014).

V osnovnem zig-zag modelu (Jones in Dangl, 2006) membranski rastlinski receptorji (ang. *plant pattern recognition receptors*, PRR) prepoznaajo ali s patogenom povezane molekularne vzorce (ang. *pathogen-associated molecular patterns*, PAMP) oz. gostiteljeve molekule poškodbe (ang. *damage-associated molecules*, DAMP) (Schwessinger in Ronald, 2012). Ta od PAMP sprožen obrambni odgovor (ang. *PAMP-triggered immunity*, PTI) omeji, ne pa popolnoma ustavi patogenove rasti. Uspešni patogeni so se na PTI evolucijsko prilagodili preko razvoja efektorjev, na kar so rastline odgovorile z razvojem specifičnih nadzornih sistemov. R-proteini so sposobni prepoznavate in nadzora aktivnosti efektorjev, kar vodi v od efektorjev sprožen obrambni odgovor (ang. *effector-triggered immune response*, ETI) (Mandadi in Scholthof, 2013).

Za razliko od PTI koraka v prirojeni imunosti prvi nivo obrambe proti virusom predstavlja utišanje RNA. To obrambo stimulirajo dvojno-verižne molekule RNA (ang. *double-stranded RNA*, dsRNA), ki igrajo analogno vlogo PAMP/DAMP molekulam iz zig-zag modela obrambe. Mehanizmi gostitelja zaznane dsRNA procesirajo v male moteče RNA (ang. *small interfering RNAs*, siRNA), katere nato vodijo razgradnjo virusnih elementov (Heinlein, 2015; Mäkinen in Hafrén, 2014). V odgovor temu mehanizmu virusi namesto

razvoja efektorjev izražajo supresorje utišanja RNA (ang. *RNA silencing suppressors*, RSS), ki ovirajo biološko pot mehanizma utišanja. R-proteini nato prepoznavajo te RSS, enako kot pri klasičnem zig-zag modelu obrambe.



Slika 3: Modela tekmovanja med patogeni/virusi in rastlinami (Nakahara in Masuta, 2014: 3)

(a) Standardni zig-zag model, ki opisuje interakcijo med patogenom in rastlino, kjer membranski rastlinski receptorji (PRR) zaznajo molekularne vzorce, kar vodi v od PAMPa sprožen obrambni odgovor (PTI) in nato razvoj efektorjev uspešnih patogenih (ETS). Te efektorje prepozna receptorjem podobni proteini (R-proteini), kar vodi v od efektorjev sprožen imunski odgovor (ETI). (b) Model molekularne interakcije med virusom in gostiteljem, kjer prvi nivo obrambnega sistema sprožijo virusne dvojnoverižne molekule RNA (dsRNA) – t.i. utišanje RNA. Namesto razvoja efektorjev virusi izražajo supresorje utišanja RNA (RSS), ki jih prepozna R-proteini.

Figure 3: Arms race models between pathogens/viruses and plants (Nakahara and Masuta, 2014: 3)

(a) The standard zig-zag model describing pathogen and plant interaction where plant pattern recognition receptors (PRRs) detect molecular patterns, which leads to PAMP-triggered immunity (PTI) and then development of effectors by the successful pathogens (ETS). These are then recognized via receptor-like proteins (R-proteins), which leads to effector triggered immunity (ETI). (b) Molecular virus-host interactions model where viral double stranded RNAs (dsRNA) trigger the first immunity layer – RNA silencing. Here in place of effectors, viral expressed RNA silencing suppressors (RSS) are recognized by R-proteins.

2.1.3.1 Splošne signalne poti odziva rastline na okužbo

Spremembe, ki jih povzroča patogen v gostiteljski rastlini se delijo na zgodnje (v roku 1-30 minut po kontaktu s patogenom) ter pozne odgovore (ure oz. dnevi po kontaktu s patogenom) (Boller in Felix, 2009). Zgodnji odgovori na metabolnem nivoju povzročijo spremembe v toku ionov (Li in Ding, 2006), proizvodnjo in akumulacijo reaktivnih kisikovih vrst (ang. *reactive oxygen species*, ROS) ter spremembe v homeostazi kalcijevih (Ca^{2+}) ionov (Kenichi Tsuda in Somssich, 2015). Na celičnem nivoju to povroča spremembe v membranskem potencialu in prepustnosti membrane (Mur in sod., 2008), posttranslacijske modifikacije (Alcaide-Loridan in Jupin, 2012) in dvig koncentracije hormonov (Abe in sod., 2012).

Kasnejše spremembe zajemajo transkripcijsko reprogramiranje izražanja genov povezanih z obrambo (Zipfel in sod., 2006), kjer lahko pride do sprememb v izražanju tudi do 20 % celotnega gostiteljevega genoma (Padmanabhan in Dinesh-Kumar, 2010). Te spremembe so rezultat medsebojnega pogovora (ang. *cross-talk*) med potmi prenosa signala, kateri vključuje fosforilacijske kaskade od mitogena aktiviranih proteinskih kinaz (ang. *mitogen-activated protein kinase*, MAPK) (Asai in sod., 2002), proizvodnjo salicilne kisline (Kenichi Tsuda in Somssich, 2015) ter spremembe v poteh ostalih rastlinskih hormonov (Kenichi Tsuda in sod., 2009).

Aktivacija teh primarnih signalizacijskih kaskad vodi v spremembo transkripcijske aktivnosti genov v večih družinah proteinov povezanih s patogenezo proteinov (ang. *pathogenesis-related protein*, PR-protein) (Mur in sod., 2008), kot tudi nekaterih genov povezanih s signalizacijo in regulacijo transkripcije (Pacheco in sod., 2012; Postnikova in Nemchinov, 2012), kar povzroči aktivacijo sekundarnih signalizacijskih kaskad. V poznih odgovorih na okužbo prihaja tudi do sprememb v celični steni, kjer pride do nalaganja kaloze (Gómez-Gómez in sod., 1999), ali do sprememb v razvoju rastline, z zmanjšano fotosintetsko aktivnostjo (Bilgin in sod., 2010) in inhibicijo rasti poganjkov ali korenin (Zipfel in sod., 2006).

Obstoj razvjetenega omrežja povezav med signalizacijskimi potmi rastlinskih hormonov v odgovoru na okužbo je dobro znan (Koornneef in Pieterse, 2008). Poleg etilena (ET), jasmonske kisline (JA) in salicilne kisline (SA) (Palukaitis in sod., 2013), pa se vse bolj

kaže vključenost ostalih rastlinskih hormonov, katerih osnovna funkcija sicer ni obramba, med drugim abscizinske kisline (ABA), giberelinov (GA), avksinov (AUX) (Palukaitis in sod., 2013) ter citokininov (CK) (Naseem in sod., 2012). Če poleg rastlinskega medsebojnega pogovora upoštevamo tudi zunanje dejavnike (npr. patogen in vektor s katerim se prenaša), kompleksnost tega že tako večdimenzionalnega pogovora hitro naraste (Palukaitis in sod., 2013).

2.1.3.2 Signalne poti odziva rastline ob virusni okužbi

Odnos med virusom in rastlinsko imunostjo je relativno slabo raziskan, saj se je zgodovinsko rastlinska molekularna virologija ukvarjala bolj z raziskovanjem samega patogena (Zvereva in Pooggin, 2012). Odgovor rastline na virusno okužbo ponavadi ni označen kot PTI odgovor, saj virus ne izraža molekul PAMP/MAMP ali efektorjev. V rastlinah primarni protivirusni sistem obrambe predstavlja utišanje RNA (Ding in Voinnet, 2007), a so protivirusni imunski koncepti običajno izključeni iz modelov rastlinskega obrambnega odgovora (Mandadi in Scholthof, 2013).

Za uspešno virusno okužbo je potrebna uspešna ugrabitev gostiteljevega sistema (ang. *hijacking*), kjer se virusna RNA najprej sprosti in prevede po vstopu v rastlinsko celico, nato pa tvori replikacijske komplekse s faktorji gostitelja. Virus tako začne proces širjenja po rastlini preko okužb sosednjih celic. Interakcije med virusno RNA in faktorji gostitelja pospešujejo hitrost virusne okužbe in pripomorejo k vzpostavitvi optimalnih pogojev. Najbolj ugodne pogoje virus doseže predvsem preko motenj gostiteljevega obrambnega sistema in preusmerjanjem celičnih virov v proizvodnjo sebi lastnih komponent (Dodds in Rathjen, 2010; Ivanov in sod., 2014).

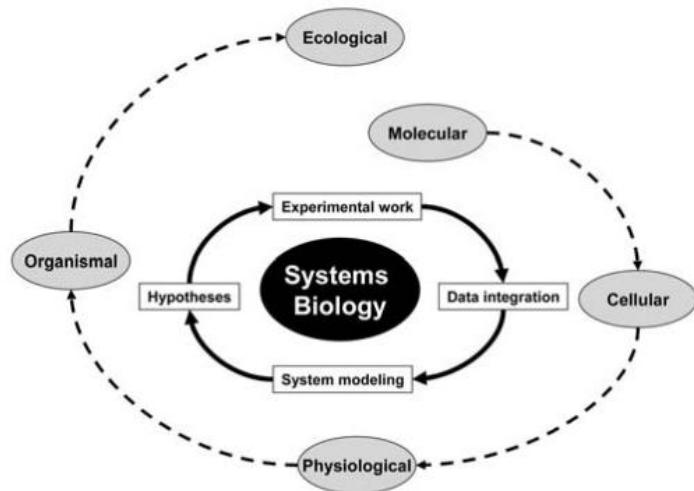
Analiza transkriptoma je pokazala tako povečano kot zmanjšano izražanje različnih genskih skupin gostiteljevih transkriptov, npr. genov vključenih v signalizacijo ali obrambo na biotski stres ter nekaterih transkripcijskih faktorjev (pregledni članek Mandadi in Scholthof, 2013). Poleg transkriptomske analize obstoje meritve sprememb na proteomskem nivoju po okužbi, katere pa niso tako izrazite kot na transkriptomskem (Di Carli in sod., 2012). Spremembe so bile zaznane v proteinih sinteze in razkroja, proteinov vključenih v prenos signala, obrambo pred boleznijo in regulacijo porabe energije na metabolnem nivoju (Yang in sod., 2011).

Obstoječi podatki še niso uspeli dobro ločiti učinkov virusne okužbe od procesov rastlinske obrambe. Znanje o mehanizmu interakcije med virusom in gostiteljsko rastlino je tako še precej pomanjkljivo: katere signalizacijske poti in transkripcijski faktorji se sprožijo v odgovoru, kateri viralni proteini so nujni pri transkripcijskem reprogramiraju (ter preko tega virusni ugrabitevi gostiteljevih poti), opis časovnega poteka virusne okužbe, ipd. (Ivanov in sod., 2014).

2.2 PRISTOPI SISTEMSKE BIOLOGIJE

Zaradi močne medsebojne povezanosti procesov nam obravnavanje živih organizmov le po posameznih komponentah omejuje razumevanje sistema kot celote. Vpogled v nešteta različna stanja sistemov nam omogočajo metodologije post-genomskega obdobja preko meritev podatkov v velikem obsegu (Gutiérrez in sod., 2005). Integracija teh podatkov nam nudi boljše razumevanje delovanja celotnega organizma, preko vpliva tako individualnih molekul kot tudi njihovih medsebojnih dinamik (Mazzocchi, 2012). Naloga združevanja heterogenih tipov informacije pa je hitro vodila v povezovanje različnih znanj, npr. biologije, fizike, matematike, tehnike/inženirstva in računalništva. V presečišču omenjenih disciplin se je tako razvila nova disciplina, sistemská biologija, ki poskuša določiti kako nelinearne interakcije med gradniki sistema privedejo do bioloških organizacijskih enot (Mazzocchi, 2012).

Področje sistemské biologije ločimo na dve večji skupini: pragmatično sistemsko biologijo ter sistemsko-teoretično biologijo (O’Malley in Dupré, 2005). Pristop raziskovanja je pri obeh enak, sestavljen iz korakov zbiranja podatkov, njihove integracije, formalnega matematičnega modeliranja sistema, temeljite računske analize in izvajanja poskusov (Ge in sod., 2003). Analiza rezultatov vpliva perturbacij na živi sistem nato vodi v oblikovanje novih bioloških hipotez, ki ta ponavljajoč se proces vrnejo na začetno točko (Slika 4).



Slika 4: Pristop sistemске biologije (Gutiérrez in sod., 2005: 2)

Sistemsko biologijo tvori ponavljajoč se proces poskusov na globalnem nivoju, integracije podatkov, modeliranja sistema in tvorbe novih hipotez (notranji krog). Končni cilj področja je generiranje modela rastline kot celote, ki opisuje procese preko vseh slojev biološke organizacije (zunanja spirala).

Figure 4: Systems biology approach (Gutiérrez in sod., 2005: 2)

Systems biology embodies an iterative process of experimentation at the global level, data integration, modeling of the system and hypothesis forming (inner circle). The ultimate goal of the domain is the generation of a whole plant model, which would describe processes across all layers of biological organization.

Razlike med stališčema pragmatične in sistemsko-teoretične smeri področja izhajajo iz samega pojmovanja celotnega sistema. Pripadniki pragmatične sistemski miselnosti izkoriščajo razvoj novih tehnologij za zbiranje in analizo velikih količin podatkov ter te podatke združujejo v različne modele (Auffray in sod., 2003), a biološke sisteme vedno vidijo le kot zbirke različnih (sicer večjih) delov in ne kot celoto.

Po drugi strani pripadniki sistemsko-teoretične smeri biološke sisteme vidijo kot rezultat kompleksnosti osnovnih organizacijskih principov in tako več pozornosti posvečajo analizi le-teh, a je opis opazovanega sistema tu zelo splošen. Dodatno so pristopi splošne sistemski teorije preveč abstraktni glede na cilje sistemski biologije. Kljub temu so nekateri rezultati sistemsko-teoretične biologije privedli do razvoja področja analize omrežij (Roy, 2014). Le-te se poslužuje vse več sistemskih biologov, saj so rezultati primerni za pojasnjevanje kompleksnosti interakcij med različnimi komponentami, ki izhaja iz večsmernosti znotrajceličnih signalizacijskih omrežij in njihovimi povratnimi mehanizmi. S tem je izkazala tudi pomembnost razvoja novih postopkov vizualizacije in robustnih statističnih metod za odkrivanje novih bioloških konceptov (Gutiérrez in sod., 2005).

Vzporedno z obema smernicama področja sta se razvila dva pristopa, in sicer od spodaj navzgor (ang. *bottom-up*; Noble, 2002) ter od zgoraj navzdol (ang. *top-down*; Bray, 2003). Pristopi pragmatične sistemsko biologije se najraje poslužujejo strategije od spodaj navzgor, kjer se postopek začne z natančnim opisom nižjih nivojev komponent (npr. DNA ali proteinov), nakar se model širi proti opisom njihovih dinamik na višjih nivojih (Windram in sod., 2014). Pri tem pristopu nepoznane ter nevključene komponente sistema povzročajo veliko odstopanje rezultatov simulacije v modelu od eksperimentalnih podatkov. Pristope od zgoraj navzdol uporabljajo sistemsko-teoretični sistemski biologi. Če je funkcija omrežja znana se začne z opisi višjih nivojev, nakar se lastnosti nižjih nivojev glede na višje nivoje. Zaradi teoretičnih in praktičnih omejitev obeh strategij se je razvil tudi pristop začetka iz sredine (ang. *middle-out*). Tu se nivo opazovanja prilagodi izbranemu znanstvenemu problemu, nato pa se nižje nivoje uporablja za podrobno obrazložitev delovanja, višje pa za celostni pogled na sistem (Mazzocchi, 2012; Noble, 2006).

2.2.1 Analiza bioloških omrežij

Pravilno delovanje celice potrebuje usklajeno delovanje tisočih genov, ki informacijo iz okolja ustrezno predelajo in opravljajo informaciji primerne biološke naloge. S tem namenom so celice razvile genska omrežja, ki omogočajo kroženje informacije (Elena in Rodrigo, 2012) in zagotavljajo pravilno izvajanje celičnih funkcij (Carrera in sod., 2009). V genskih omrežjih tako pogosto prihaja do oblikovanja specifičnih funkcionalnih modulov (Heyndrickx in Vandepoele, 2012). Močno orodje za proučevanje kompleksnih lastnosti bioloških omrežij predstavlja pristopi analize omrežij, ki eksperimentalne podatke združujejo z računalniškim modeliranjem (Albert in Barabási, 2002).

Omrežje oz. graf je umetno ogrodje uporabno za strukturirano opisovanje kompleksnih multivariatnih odnosov, s čimer je omogočena sistematična matematična analiza ter intuitivna vizualna predstavitev gradnikov oz. komponent (točk) in njihovih medsebojnih povezav. Preko analize gradnikov omrežja, njihovih medsebojnih razmerij in splošnih topoloških lastnosti nam analiza omrežij omogoča boljši vpogled v hierarhično organizacijo kompleksnih sistemov (Tully in sod., 2014).

Na biološkem področju so točke omrežja največkrat molekule, njihova medsebojna razmerja (povezave) pa lahko predstavljajo vse od fizičnih interakcij do regulatornih in funkcionalnih odnosov med njimi (Barabasi in Albert, 1999; Subramanian in sod., 2015). Najpogosteje je bil pristop analize omrežij uporabljen za opise na makromolekularnem nivoju, kjer v odvisnosti od pomena točke in povezave ločimo več tipov omrežij (Preglednica 1).

Preglednica 1: Tipi bioloških omrežij in pomen točk ali povezav v njih.

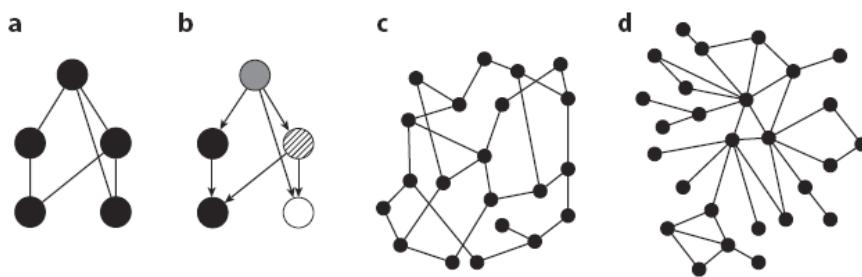
Table 1: Biological network types and the meaning of nodes and connections in them.

Tip biološkega omrežja	Pomen točke	Pomen povezave
metabolno	metabolit	udeleženost v isti metabolni reakciji
gensko regulatorno	transkripcijski faktor, mala RNA ali tarča	aktivacija, inibicija
proteinske interakcije	protein	medsebojna vezava
genetske interakcije	gen	indirektna (epistatična) interakcija
signalizacijsko	<i>raznolik</i>	<i>raznolik</i>

Signalizacijska omrežja so arhitekturno najbolj podobna električnim vezjem, z jasno ločenimi sestavnimi deli zadolženimi za zaznavo, procesiranje, filtriranje in prenos signala (Hyduke in Palsson, 2010). Narava komponente v signalizacijskem omrežju je zelo raznolika, dokler lahko vpliva na prenos signala od začetne točke (najpogosteje receptorja) do konca signalizacijske kaskade. Najbolj pogosto komponente predstavljajo proteini ali proteinski kompleksi, lahko pa so tudi geni, RNA, spojine. Tudi povezave v signalizacijskem omrežju so raznolike narave, npr. fizična vezava, fosforilacija, ubikvitinacija, glikozilacija, transkripcijska regulacija (Subramanian in sod., 2015).

2.2.1.1 Topološka analiza omrežij

Tvorba smiselnih zaključkov iz analiz velikih omrežij v današnji znanosti predstavlja velik izziv. Večina dejanskih omrežij (ang. *real-world network*), tudi bioloških, vzdržuje nekatere topološke lastnosti v okviru strukture in organizacije, ki se značilno razlikujejo od naključnih omrežij (Albert, 2005; Vidal in sod., 2011). Tako so se kot močno orodje za analizo izkazale študije topologije omrežij, katerih cilj je iskanje vzorcev interakcij v omrežju in analiza njihove splošne organizacije (Slika 5). Že samo integracija topoloških lastnosti z biološkimi opisi (npr. ontologijami) lahko privede do kvalitetnih hipotez, ki so primerne za preizkušanje v laboratoriju (Subramanian in sod., 2015).



Slika 5: Topološka organizacija omrežij (Windram in sod., 2014: 4)

Topologije omrežij: (a) neusmerjeno omrežje; (b) usmerjeno omrežje; (c) omrežje malega sveta, kjer ima večina komponent podobno število povezav; (d) brezlestvično omrežje, s povezavami od visoko povezanih žarišč komponent z eno samo povezavo.

Figure 5: Network topology types (Windram in sod., 2014: 4)

Network topologies: (a) an undirected network; (b) a directed network; (c) a small-world network, where most nodes have a similar number of connections; (d) a scale-free network, with highly connected hubs and down to nodes with single links.

Prva pomembna lastnost dejanskih omrežij je t.i. lastnost malega sveta (ang. *small-world network*, Slika 5c). V takem omrežju je povprečna najkrajša pot med vsemi točkami majhna oz. so medsebojne razdalje med vsemi točkami omrežja kratke (Barabási in Oltvai, 2004; Humphries in Gurney, 2008). Pot v omrežju je razdalja med dvema točkama oz. število povezav v najkrajši poti celotnega omrežja. Zaradi te lastnosti se dejanska omrežja tudi pogosto organizirajo v gosto povezane skupine (ang. *clusters*), katerih koeficient razvrščanja je višji od naključno pričakovanega (Mao in sod., 2009). Koeficient razvrščanja se izračuna glede na najkrajšo pot omrežja in dostopnost (ang. *closeness centrality*), katera predstavlja inverzno vsoto vseh najkrajših poti od žarišča (ang. *hub*) do vseh ostalih točk znotraj skupine (Watts in Strogatz, 1998). Te gosto povezane skupine iz biološkega vidika pogosto pomenijo funkcionalne module in so velikega pomena za razumevanje principov celične organizacije in pripadajočih funkcij (Heyndrickx in Vandepoele, 2012).

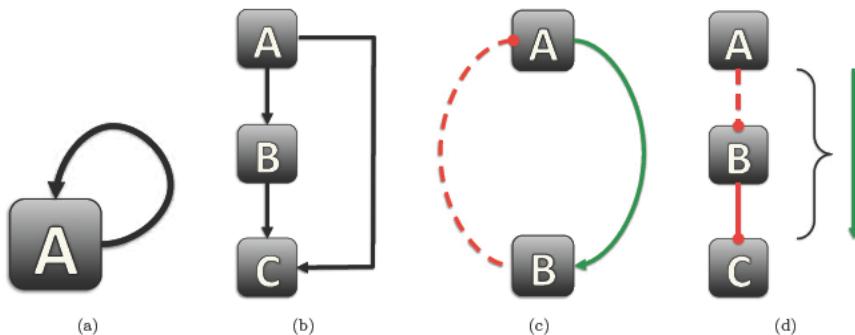
Druga lastnost dejanskih omrežij je brezlestvičnost (ang. *scale-free network*, Slika 5d), lastnost povezana s porazdelitvijo stopenj v omrežju (ang. *degree*). Stopnja točke v omrežju pomeni število vseh njenih povezav z ostalimi točkami omrežja, v usmerjenem omrežju pa jo ločimo še na vhodno (število povezav do te točke) ter izhodno (število povezav izhajajoč iz te točke). Ker veliko dejanskih omrežij tako iz vidika števila točk kot povezav pogosto sestavlja eno večje in več manjših podomrežij (Batada in sod., 2006), se brezlestvičnost preverja le na največji povezani komponenti (ang. *largest connected component*, LLC). V veliko primerih dejanskih omrežij porazdelitev stopenj v LLC nima zvonaste oblike, ampak sledi eksponentne razgradnje (Albert in Barabási, 2002). Praktično to pomeni da imajo nekatere točke izjemno visoko stopnjo, preostala večina pa nizko (Newman, 2004).

Tako lastnost malega sveta kot brezlestvičnost večata robustnost omrežja in manjšata vpliv naključnih perturbacij. Cena take robustnosti je na račun močne občutljivosti v žariščih z visoko stopnjo (Albert in sod., 2000), katerih perturbacije povzročajo velike spremembe delovanja omrežja, tudi popolno prenehanje delovanja (Barabasi in Albert, 1999). Za preživetje in razvoj fenotipskih lastnosti so pomembne tako žarišča največje povezane komponente (Gehan in sod., 2015), kot tudi žarišča, ki povezujejo različne funkcionalne module (Rodrigo in sod., 2012).

Topološke analize rastlinskih omrežij so nakazale občutljivost žarišč in vpliv njihovih perturbacij na pretok informacij po omrežju (Albert in sod., 2000). Ta lastnost žarišč se je v evoluciji odnosa med rastlino in patogeni pokazala kot pomembna v signalizaciji imunskega odziva rastline, saj proteini patogenov ciljajo ravno ta žarišča in s tem učinkovito onesposobijo obrambni sistem rastline (Mukhtar in sod., 2011; Rodrigo in sod., 2012). Na podlagi izračuna vmesnosti vsake točke (števila najkrajših poti med vsakim parom genov; ang. *betweenness centrality*) so Tully in sod. (2014) potrdili številne znane imunske regulatorje ter z obrambo povezane proteine iz različnih signalnih poti. Dodatno proteinsko omrežje proteinov Potyvirusa v osnovi nima lastnosti malega sveta in ni brezlestvično – to lastnost pridobi šele ob integraciji s proteinskim omrežjem gostitelja (Rodrigo in sod., 2012).

2.2.1.2 Pogosto najdeni motivi signalizacijskih omrežij

Biološka omrežja vsebujejo nekatere pogosto se pojavljajoče vzorce oz. motive (Slika 6). Nekateri od dobro proučenih motivov so avtoregulacija, usmerjena zveza (ang. *feedforward motif*), povratna zanka ter derepresijski motiv (Alon, 2007). Motivi vplivajo na ostale komponente omrežja in tako vodijo različnih kvalitativnih dinamik. Klasična orodja za analizo takih motivov so del sistemsko-teoretske analize, katere ključen koncept je naravna lastnost sistemov, da imajo omejeno število stabilnih stanj, proti katerim sistem teži (Muraro in sod., 2012). Zato je si moramo kot osnovne gradnike signalizacijskih in genetskih omrežij namesto posameznih komponent (bioloških molekul) predstavljeni te motive.



Slika 6: Pogosti motivi v omrežjih (Muraro in sod., 2012: 9)

Shema prikazuje štiri pogosto najdene motive signalizacijskih omrežij: (a) avtoregulacija; (b) usmerjena zveza; (c) povratna zanka; (d) derepresija. Zelene puščice predstavljajo aktivacijo, rdeče represijo, črne pa splošno regulacijo (aktivacija ali represija). Črtkane črte predstavljajo transkripcijsko regulacijo, polne črte pa proteinske interakcije.

Figure 6: Common network motifs (Muraro in sod., 2012: 9)

The scheme displays common signalling network motifs: (a) auto-regulation; (b) feedforward loop motif; (c) feedback loop motif; (d) de-repression motif. Green arrows represent activation, red repression, black general regulation (either activation or repression). Dashed lines represent transcriptional regulation, while solid lines represent protein-protein interactions.

Do avtoregulacije (Slika 6a) pride, kadar transkripcijski faktor aktivira ali inhibira svojo lastno transkripcijo. Negativna avtoregulacija skrajša čas odgovora (nivo ekspresije se hitro stabilizira) in zmanjša stohastično medcelično variabilnost, pozitivna avtoregulacija pa podaljša čas odgovora in s tem variabilnost poveča (Muraro in sod., 2012). Usmerjena zveza (Slika 6b) opisuje medsebojne odnose treh genov, ki so lahko tako direktni kot indirektni. Ker lahko vsaka od treh povezav pomeni ali aktivacijo ali inhibicijo, definiramo osem oblik usmerjenih zvez, iz katerih izhajajo različne oblike kvalitativne

dinamike, npr. zamude v odgovoru, pulzov v ekspresiji genov ali pospešenega odgovora (Muraro in sod., 2012). Derepresija oz. inaktivacija represorskega proteina (Slika 6d) preko razgradnje ali ubikvitinacije je pogost motiv večine rastlinskih signalizacijskih poti in je pogosto hitrejša od direktne aktivacije (Rosenfeld in sod., 2002).

Eden najbolj zanimivih motivov signalizacijskih omrežij predstavljajo povratne zanke, ki ohranjajo nivo ekspresije genov v določenem razponu. Pri tem motivu gre za prenašanje informacije od končne točke nazaj v izvorno točko prvotnega signala (vsaka reakcija ima povratno reakcijo). Glede na usmeritev povezav v motivu lahko zopet definiramo različne oblike povratnih zank, negativno povratno zanko (kombinacija pozitivne in negativne regulacije), dvojno pozitivno ter dvojno negativno zanko. Negativne povratne zanke lahko vodijo ali v homeostazo ali v oscilacije sistema (Baier in Sahle, 1998; Tyson in sod., 2001). Dvojno pozitivne zanke ter dvojno negativne zanke pa obstoje v dvojnostabilnih oblikah: rezultat dvojno pozitivnih zank je lahko ekspresija obeh naenkrat ali nobenega od transkripcijskih faktorjev, v primeru dvojno negativne zanke pa vedno lahko obstoji le eden izmed obeh.

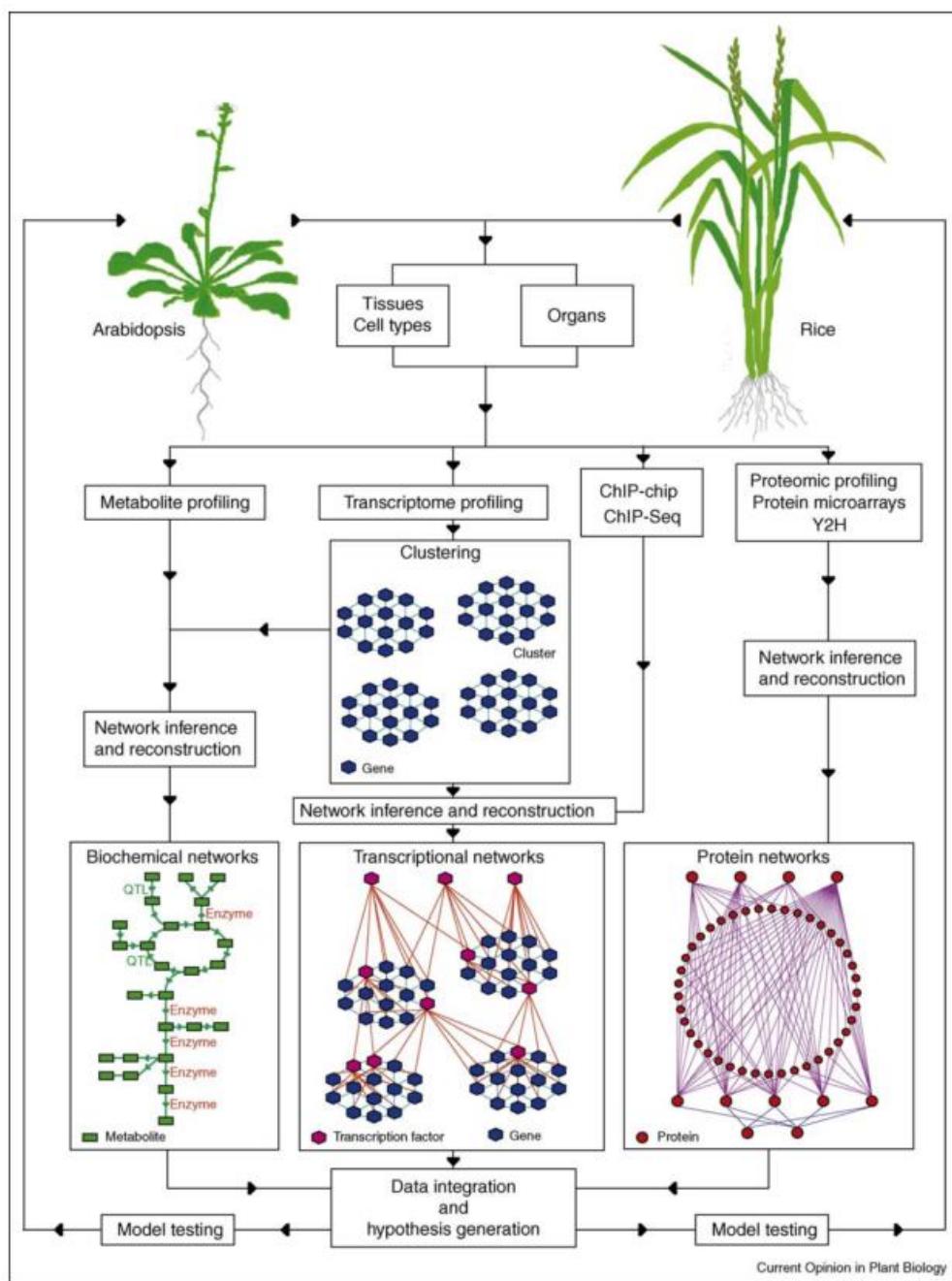
Različni motivi se lahko organizirajo v ugnezdenе oblike in s tem tvorijo nove množice možnih stabilnih stanj, kar dodatno oteži razumevanje končnih celičnih procesov, ki so lahko povezani tako preko povratnih kontrolnih točk kot medsebojne regulacije. Dodatno na podlagi motivov obnašanja sistema ni možno napovedat, saj so kvalitativne dinamike močno odvisne tudi od točnih parametrov kemijske kinetike (Bhalla in sod., 2002). Sistemska analiza sicer lahko poda možen razpon dinamičnih lastnosti sistema, kar je potrebno integrirati z eksperimentalno in simulacijsko analizo – dinamičnim modeliranjem sistemov (Bhalla, 2003).

Vsi omenjeni motivi signalizacijskih omrežij so že bili identificirani kot pomembni regulatorni mehanizmi rastlinskih procesov. Tako je direktna pozitivna avtoregulacija bistvena pri vzdrževanju apikalnega meristema poganjkov (Katsutoshi Tsuda in sod., 2011), usmerjena zveza nadzoruje časovno izražanje preklopa iz vegetativnega v reproduktivni razvoj (Saddic in sod., 2006), skupina negativnih povratnih zank regulira cirkadiano uro (Locke in sod., 2006), pozitivna povratna zanka odloča o diferenciaciji koreninskih epidermalnih celic (Kang in sod., 2009), motiv deregulacije pa omogoča hiter odgovor v signalizacijskem omrežju avksina (Farcot in sod., 2015).

2.2.2 Integracija predhodnega znanja z eksperimentalnimi podatki

Integracija podatkov, v okviru enega samega omskega pristopa, še bolj pa v okviru integracije med različnimi tipi omskih podatkov, predstavlja močno orodje za spoznavanje delovanja bioloških sistemov (Slika 7). Ker je večina rastlinskih imunskih odgovorov pod transkripcijskim nadzorom, so pristopi opazovanja vzorcev izražanja posebej primerni za proučevanje aktivacije ali inhibicije specifičnih regulatornih in metabolnih poti (Wise in sod., 2007). Kadar je o sistemu malo znanega ali kadar je sistem zelo velik in vsebuje velik deležem neznanih komponent, se pogosto uporabi pristope inference od zgoraj navzdol (ang. *reverse-engineering*). Tu se zanimive lastnosti sistema, npr. topologijo omrežja ali parametre, ki vodijo v kvalitativne dinamike (De Backer in sod., 2010; Dhar, 2007), najpogosteje opredeli na podlagi podatkov mikromrež (Bansal in sod., 2007).

Teoretičnih metod, algoritmov in orodij za inferenco omrežij je bilo razvitih že veliko (Krouk in sod., 2013; Marbach in sod., 2010; Penfold in Wild, 2011). Raznolikost obstoječih metod je posledica težavnosti same naloge in lastnosti podatkov, saj so le-ti omejeni v smislu velikega števila genov glede na število neodvisnih meritev (Breeze in sod., 2011) ter naključnega šuma prisotnega v meritvah (Godsey, 2013). Nasprotno imajo rezultati pristopov od zgoraj navzdol veliko vsebnost lažno pozitivnih in lažno negativnih povezav. Ta problem se lahko reši ali preko zmanjšanja genov z nezanimivimi vzorci izražanja ali izbora genov, ki se razlikujejo med različnimi eksperimentalnimi tretmaji (Kiddle in sod., 2009; Stegle in sod., 2010).



Slika 7: Pristopi inference funkcionalnih interakcij in rekonstrukcije omrežij (Moreno-Risueno in sod., 2010: 2)

Meritve transkriptov, proteinov in metabolitov skupaj z ostalimi pristopi (kromatinska imunoprecipitacija, mikromrežna hibridizacija, določanje zaporedij) se uporablja pri inferenci funkcionalnih interakcij med glavnimi komponentami. Interpretacija teh interakcij nato vodi do formulacije novih hipotez in pripomore pri novih napovedi funkcij genov.

Figure 7: Approaches to functional network inference and reconstruction (Moreno-Risueno in sod., 2010: 2)

Measurements of transcripts, proteins and metabolites coupled with other approaches (chromatin immunoprecipitation, hybridization to microarrays, sequencing) are used to infer functional interactions between main components. Interpretation of these interactions then leads to new formed hypotheses and assists in novel gene function prediction.

Ena izmed možnih metod inference je ko-ekspresija, ki iz transkriptomskih ekspresijskih podatkih izlušči skupine genov (module) s podobnimi časovnimi vzorci izražanja (Ma in Bohnert, 2007). Ko-ekspresijsko analizo se lahko združi z iskanjem skupnih vezavnih mest transkripcijskih faktorjev (*cis*-regulatornih elementov). Oba pristopa družno pripomoreta h karakterizaciji transkripcijskih regulatornih omrežij (Tully in sod., 2014) in iskanju novih genov (Windram in sod., 2014).

Rezultat ko-ekspresijskih metod so genska korelacijska omrežja (ang. *gene correlation network*, GCN). Omrežja se lahko dodatno uteži glede na jakost korelacije med geni, glede na izbrano pražno vrednost (Horvath in Dong, 2008; Langfelder in Horvath, 2008) ali preko uporabe linearnih modelov, ki utež določijo glede na drugo že obstoječe omrežje (Galbraith in sod., 2006; Liao in sod., 2003). Dobljena GCN se v odvisnosti od njihove velikosti in obstoječega znanja o delovanju komponent lahko analizira z različnimi orodji, (Muraro in sod., 2012). Ker imajo obe lastnosti dejanskih omrežij (lastnosti malega sveta in brezlestvičnost), so primerna za topološke analize (Jordan in sod., 2004; Lee in sod., 2010), ne pa tudi v dinamičnem modeliranju sistemov (Muraro in sod., 2012).

Analize genskih korelacijskih omrežij so se uporabile za identifikacijo in iskanje novih kandidatnih genov (Cai in sod., 2014; Mustroph in sod., 2010), iskanje ključnih modulatorjev imunskih odgovorov rastline (Tully in sod., 2014; Zheng in Zhao, 2013), rekonstrukcijo regulatornih poti (Ma in sod., 2007) ter opazovanje časovno-prostorsko specifičnih vzorcev ekspresije v tkivu (Brady in sod., 2007). Veliko moč odkrivanja podatkov imajo tudi medvrstne primerjalne ko-ekspresijske študije med različnimi vrstami, kjer so opazovali ohranjenost genskih modulov tako med bližnje sorodnimi vrstami (Movahedi in sod., 2011; Street in sod., 2008) kot tudi med filogenetsko bolj oddaljenimi (Ma in sod., 2005).

Geni iste ko-ekspresijske skupine večinoma vsebujejo medsebojno funkcionalno povezane gene, sodelujejo v podobnih bioloških procesih in so zato pogosto opisani s podobnimi funkcionalnimi pripisi (Massa in sod., 2013). V nekaterih primerih so tudi del iste biološke poti (Lee in Sonnhammer, 2003; Ma in sod., 2007). Tu se pokaže pomen integracije samih omrežij tudi z ostalimi tipi podatkov, npr. ontologijami (Dutkowski in sod., 2013), ki okrepijo informacijo v genskem korelacijskem omrežju.

Močen pristop, ki omogoča prioritizacijo specifičnih genov, predstavlja npr. fuzija podatkov, kjer se z uporabo pristopov strojnega učenja analizira heterogene podatke in poda seznam prioritetnih genov za proučevan biološki sistem (Moreau in Tranchevent, 2012; Žitnik in Zupan, 2014). Pristop je primeren za organizme, kjer je na voljo veliko podatkov, zato trenutno ni najbolj primeren za uporabo na redkih podatkih rastlinskega sveta. Alternativo za rastline predstavlja orodje MORPH (Tzfadia in sod., 2012), ki združuje ko-eksprejsiska omrežja ter omrežja medsebojnih povezav med komponentami s seznamom genov, zanimivih za neko proučevano pot.

Integracijo omskih podatkov v rastlinskem svetu se uporablja za različne raziskovalne cilje, od zasnove rastlinskih metabolnih poti preko integracije genomskeh in metabolomskeh podatkov (Oksman-Caldentey in Saito, 2005), za anotacijo in analizo novih genomov preko integracije proteomskeh in metabolomskeh podatkov (May in sod., 2008), pa do iskanja novih aktivnih rastlinskih poti z integracijo transkriptomike in metabolomike (Jauhainen in sod., 2012). Integracija omskih podatkov s topološko informacijo v omrežju se je izkazala za uspešno pri iskanju ohranjenih skupin (Mutwil in sod., 2010), iskanju neznanih transkripcijskih faktorjev (Hirai in sod., 2007) ali pri iskanju novih malih RNA (Zhou in sod., 2007). Bolj tarčni poskusi so priveli do genskega omrežja riževih genov vključenih v stresni odgovor (Cooper in sod., 2003), definirali regulatorne točke sistemsko pridobljene odpornosti (Wang in sod., 2006) ter vodili v razvoj orodij, ki raziskovalcem omogočajo lastno analizo, npr. VirtualPlant (Gutiérrez in sod., 2007).

2.2.3 Dinamično modeliranje sistemov

Tkvno specifična diferenciacija, rast ter odgovori rastlinskih celic na različne okoljske signale so pod nadzorom hormonov (Ubeda-Tomás in sod., 2012). Obstoje mehanizmov medsebojnega pogovora med signalizacijskimi potmi različnih hormonov je dobro poznano in omogoča širok razpon odgovorov na okoljske stimule – sinergistične, aditivne ali antagonistične (Pieterse in sod., 2009). Heterogene porazdelitve koncentracij hormonov po različnih tkivih, način transporta med tkivi in prepustnost samih tkiv predstavljajo izziv v razumevanju končnih učinkov teh kompleksnih interakcij (Windram in sod., 2014). Identifikacija in razumevanje dinamik interakcij ni intuitivna ter potrebuje posluževanje pristopov dinamičnega modeliranja na večih nivojih (Muraro in sod., 2012).

Za boljše razumevanje medsebojnih interakcij med rastlino in patogenom je pomembno ustvariti zanesljiv shematičen model omrežnega pogleda na rastlinsko imunsko signalizacijo, ki bo omogočal razčlenitev kritičnih procesov, njihovo modeliranje in kasneje, manipulacijo rastlinske imunosti (Motion in sod., 2015). Pojem modela na področju matematičnega modeliranja pomeni podrobен opis sistema glede na njegove sestavne enote in njihove medsebojne odnose. Modele se prilagodi v odvisnosti od zadanega cilja raziskovanja in so odvisni od nivoja biološkega sistema. Ker je opredelitev nivoja v bioloških znanostih lahko precej široka, so modeli lahko stukturno molekulski, celični, makroskopski ter populacijski. Na stukturno molekulskem nivoju se raziskuje evolucijo fizičnih in biokemijskih stanj ene same celice – analiza genov, proteinov in signalov; celični nivo predstavlja skupine različno vzajemno delujočih celic in molekul; makroskopski nivo vključuje tkiva, organe in sisteme. Najvišji organizacijski nivo opisuje dinamiko populacij glede na eno ali več lastnosti (Motta in Pappalardo, 2013).

Tako kot se razlikujejo posamezni nivoji biološke organizacije, se razlikujejo tudi pristopi modeliranja. Modeliranje na znotraj-celičnem nivoju je oteženo zaradi nepopolnega znanja o medsebojnih odnosih komponent celice, kar se lahko rešuje z uporabo t.i. zaprtih modelov (ang. *black box model*). Tkiva so ponavadi opisana z navadnimi in delnimi diferencialnimi enačbami (ang. *differential equation*, DE), za organe pa za vsak tip tkiva in biološke funkcije potrebujemo posamičen model. Modeli populacij (npr. epidemiološki modeli) poleg svojih lastnih kompleksnosti podedujejo tudi vse kompleknosti prej omenjenih nivojev (Motta in Pappalardo, 2013). Področje se počasi razvija tudi v hkratno modeliranje večih nivojev, kjer se osnovni model zaradi trenutnih težav z neizračunljivostjo razdeli v več podmodelov in obdeluje ločeno (Alemani in sod., 2012).

Dinamični modeli nadrobno opisujejo specifične celične procese in predstavljajo primerno okolje za postavljanje in preizkušanje novih hipotez z *in silico* poskusi (Gutiérrez in sod., 2005). Izhajajoč iz splošne definicije modela lahko izpeljemo tri glavne namene uporabe dinamičnih modelov v znanosti: (1) sinteza oz. odkrivanje novih znanj ob primerjavi z rezultati modela; (2) analiza različnih vplivov na sistem in kasneje napovedi odgovorov sistema; (3) zasnova alternativnih sistemov, sposobnih oponašanja delovanja izvornega sistema.

Primere uporabe dinamičnega modeliranja rastlinskih sistemov predstavljajo tako manjši modeli, npr. modeli povratnih zanke udeleženih v transkripcijski in post-transkripcijski regulaciji (Pokhilko in sod., 2012) ali transkripcijska stikala stresnega odgovora (Penfold in sod., 2012), kot tudi večje študije proučevanja metabolizma (Williams in sod., 2010), odgovora na okužbo (Windram in sod., 2012), kompleksnih medsebojnih odnosov med rastlinskimi hormoni (Li in sod., 2006; Likhoshvai in sod., 2007; Naseem in sod., 2012) ali modeliranja celotne cirkadiane ure rastline navadnega repnjakovca (Zeilinger in sod., 2006).

2.2.3.1 Tipi dinamičnih modelov signalizacijskih omrežij

Predstavitev delovanja signalizacijskega omrežja preko uporabe diagramov in pripovednih modelov je v biologiji pogosto uporabljen pristop (Endy in Brent, 2001). Dinamično obnašanje velikih skupin vzajemno deluječih komponent se lahko zaobjame z uporabo matematičnih tehnik s pravilnim formalnim opisom sistema (Wolkenhauer, 2001). Trenutno je dinamično modeliranje zaradi matematične kompleksnosti še neprimerno za analizo celih bioloških sistemov, lahko pa se ga uporabi na manjših podenotah sistema. Pristope za modeliranje malih signalizacijskih omrežij (velikosti od deset do nekaj sto genov) lahko razdelimo v tri glavne skupine glede na pristop: logične, deterministične in stohastične.

Logični modeli omogočajo kvalitativno razumevanje dinamike celotnega omrežja, njihova enostavnost pa dovoljuje analize tudi na večjem številu komponent (Karlebach in Shamir, 2008), npr. preko Bayesove formulacije modela ali Boolejevih logičnih modelov. Bayesovo omrežje je verjetnostni model v obliki usmerjenega acikličnega grafa, kjer so povezave med komponentami pogojno odvisne (ang. *conditional dependency*), odsotnost usmerjene povezave v grafu pa je zapisana kot pogojna neodvisnost (ang. *conditional independency*). Zaradi take oblike osnovno Bayesovo omrežje ni zmožno zajeti dinamike sistema, saj se nivo genskega izražanja v omrežju ne more spremenjati s časom, a se jih lahko nadgradi v npr. diskretna Bayesova omrežja (Muraro in sod., 2012). V Boolejevih logičnih modelih je vsak gen v nekem koraku lahko v aktivnem ali neaktivnem stanju (zapisano s stanji 0 in 1). Omrežje se podobno kot v diskretnih Bayesovih omrežjih časovno razvija preko diskretnih korakov, v odvisnosti od funkcij, ki so zapisane z logičnimi pravili. Kljub enostavnosti logičnih modelov, ki nam omogoča analizo

relativno velikih omrežij, težave pri analizi predstavljajo omejena podrobnost sistema ter nedoslednosti oz. nepravilnosti v modelu.

Drugo skupino modelov, zvezne deterministične modele, se uporablja za podrobno analizo dinamik v sistemu. Ti modeli so ponavadi zapisani v obliki navadnih diferencialnih enačb (ang. *ordinary differential equations*, ODE), preko katerih opisujejo hitrost sprememb v času, prostoru ali drugi zvezni spremenljivki in so izraženi v obliki derivatov diferencialnih enačb. Formalizem ODE omogoča delo z zelo podrobnnimi modeli, in tako omogoča razlikovanje med različnimi tipi procesov, npr. hitrostjo translacije, transkripcije, tvorbe bioloških kompleksov ali razgradnje bioloških molekul (Windram in sod., 2014). Taki tipi modelov omogočajo kvantitativne in bolj direktne primerjave z biološkimi poskusi, potrebujejo pa podrobne kinetične podatke in dobre parametre procesov modela, ki pogosto niso znani. Težave povzročajo sistemi, kjer koncentracije vseh komponent modela ne ohranjajo konstantnih koncentracij, v takih primerih se lahko poslužimo uporabe delnih diferencialnih enačb (ang. *partial differential equations*, PDE; (Muraro in sod., 2012).

Zadnjo skupino predstavljajo stohastični modeli. Na gensko ekspresijo ima zaradi slučajnosti vezav med proteini ali geni ter razpada proteinov stohastičnost velik vpliv, posebej za majhna števila molekul. Eksponentno hitro naraščanje računske kompleksnosti stohastičnih modelov z rastjo sistema je vodilo v razvoj manj natančnih algoritmov na račun večje učinkovitosti, npr. hibridnih pristopov s kombinacijo navadnih diferencialnih enačb (Gillespie, 2001) ali diskretnih skokov z naključno izbranimi koraki (Kiehl in sod., 2004).

2.2.3.2 Izgradnja dinamičnega matematičnega modela

Ob dobrem razumevanju sistema lahko le-tega zapišemo v obliki matematičnega modela (Windram in sod., 2014). V prvem koraku je potreben izbor najbolj primerenega formalizma zapisa modela z obzirom na opazovani problem. Za modeliranje bioloških sistemov obstaja veliko različnih formalizmov (nekateri opisani v prejšnjem poglavju), delno zaradi raznolikosti pojavov, delno pa zaradi same multidisciplinarnosti raziskovalnih skupin. Na podlagi postavljenega modela se lahko lotimo kvantitativnih analiz, npr. preko *in silico* matematičnih simulacij.

Določanje parametrov diferencialnih enačb (inferenca) predstavlja drug korak v pripravi matematičnega modela, t.i. kalibracijo. Po formulaciji modela vsebujejo enačbe, ki povezujejo komponente modela in so jih izbrali strokovnjaki področja (npr. masni prenos, Michaelis-Menten kinetika, Hillova enačba), neznane parametre. Kompleksni modeli, posebej prostorski, pogosto vsebujejo veliko število neznanih parametrov; dodatno tudi znane vrednosti kinetičnih parametrov spremelja velika negotovost. Zelo majhen delež modelov se lahko reši analitično preko algebrske derivacije lastnosti sistema (Motta in Pappalardo, 2013), za ostale pa potrebujemo računske približke.

Inženirska stroka te težave združuje in rešuje pod sklopom t.i. problema identifikacije sistemov (ang. *system identification problem*). V primeru, da imajo diferencialne enačbe analitično rešitev, zadostuje uporaba metode nelinearne regresije (Chatterjee in Simonoff, 2012), sicer pa lahko uporabimo alternativne rešitve tega problema, ki so jih razvili na inženirskem področju (Chen in Wu, 2008; De Boor in Swartz, 1973; Himmelblau in sod., 1967; Huang in sod., 2006; Ramsay in sod., 2007). Poleg razvoja v inženirski stroki se metodologije za ocenjevanje parameterske strukture modelov razvijajo tudi na področju sistemsko biologije (Äijö in Lähdesmäki, 2009; Gao in sod., 2008; Kirk in Stumpf, 2009; Rogers in sod., 2007; Wang in sod., 2007). Večina naštetih metod preko metode optimizacije privede do končnega rezultata, ki je ponavadi podan kot ena sama kombinacija parametrov, ki pa ni nujno tudi obrazložljiva iz biološkega vidika. Alternativo predstavlja Bayesova linearna metoda, ki kot rezultat namesto ene kombinacije poda prostor možnih rešitev danega sistema, ki je bila uporabljena za zelo različnih področij raziskovanja (Craig in sod., 1997; Craig in sod., 2001; Vernon in sod., 2010; Williamson in sod., 2012).

Zadnji korak modelirnega procesa predstavlja validacija modela, kar v praksi pomeni primerjavo rezultatov simulacije modela z eksperimentalnimi vrednostmi. Matematični modeli morajo čim bolje preslikati rezultate eksperimentalnega dela in opazovanj (opisni modeli) ali biti sposobni napovedovanja novih lastnosti ter predlagati najbolj smiselen eksperimentalni načrt (napovedni modeli) (Motta in Pappalardo, 2013). Rezultati modelov, ki niso v dovoljenem okvirju odstopanja, so lahko osnova za izboljšave osnovnega modela (strukture, enačb ali pa parametrov enačb), rezultati modela skladni z eksperimentalnimi podatki pa vodijo v potrjevanje novih hipotez. Iz tega vidika ciklični pristop ponavljanja korakov med modeliranjem, simulacijo, napovedmi in eksperimentalnim preizkušanjem vodi v odkrivanje novega znanja (Windram in sod., 2014).

2.3 JAVNO DOSTOPNO ZNANJE V RASTLINSKI BIOLOGIJI

2.3.1 Eksperimentalni podatki

Hiter razvoj tehnologij različnih omskih pristopov je omogočil podrobljeno preiskovanje globalnih lastnosti kompleksnih bioloških sistemov. To je pospešilo razvoj različnih podatkovnih virov in spodbudilo razvoj translacijskega (ang. *translational*) raziskovanja z integracijo znanja v različnih rastlinskih vrstah (Mochida in Shinozaki, 2011). Omske pristope pridobivanja podatkov lahko ločimo na štiri osnovne (genomika, transkriptomika, proteomika, metabolomika). Iz pridobljenih podatkov so se razvili integratori pristopi variomike (vsa genetska variacija populacije), epigenomike (vsi elementi, ki nadzorujejo izražanje genov in niso zapisani v genskem kodu), interaktomike (vse molekularne interakcije sistema) ter fluksomike (časovno odvisne dinamike malih molekul) (Baker, 2013).

Uporaba tehnologije sekvenciranja naslednje generacije (ang. *next generation sequencing*, NGS) je vodila v ponovno določanje zaporedij genomov nekaterih rastlin ter do *de novo* sestavljanja genomov. Število novih rastlin katerih z določenim genomskeim zaporedjem narašča (pregledni članek Mochida in Shinozaki, 2011). Poleg osnovnega zaporedja genoma je pomembna tudi njegova variabilnost – identifikacija mutacij in alelov povezanih s fenotipsko raznolikostjo (DePristo in sod., 2011), kjer je pomemben

projekt določanja zaporedij genomov 1001 rastlin navadnega repnjakovca (Cao in sod., 2011). Pomembne podatkovne baze z združeno informacijo o genomih in variacijo različnih rastlinskih vrst so PlantGDB (Duvick in sod., 2008), Gramene (Liang in sod., 2008), Phytozome (Goodstein in sod., 2012), Ensembl Plants (Kersey in sod., 2014) ter PLAZA (Proost in sod., 2014).

Tudi na področju analize transkriptov poteka veliko projektov na osnovi tehnologije NGS, za namen iskanja alternativnih transkriptov izbranih rastlin, kvantifikacije transkriptov ali *de novo* sestavljanja zaporedij transkriptomov. Kot vira različnih surovih podatkov poskusov so pomembne baze ArrayExpress (Kolesnikov in sod., 2014), Gene Expression Omnibus (Edgar in sod., 2002) ter Sequence Read Archive baze NCBI. Pomembno spletno aplikacijo predstavlja Bio-Analytic Resource for Plant Biology, BAR (Patel in sod., 2012), ki omogoča vpogled in analizo genske ekspresije med različnimi rastlinami, rastlinskimi tkivi ter ogled npr. molekularnih markerjev. Epigenomika sledi različnim post-transkripcijskim spremembam, rezultati področja so prav tako zbrani v svojih bazah podatkov, npr. SIGnAL (Lister in sod., 2008).

Informacije o skupinah proteinov celotnih genomov so za proteoma rastlin navadnega repnjakovca in koruze zbrani v rastlinski proteinski bazi, PPDB (Sun in sod., 2009), za riž pa v OryzaPG-DB (Helmy in sod., 2012). Analize interaktomov, ki jih izluščimo z rezultati interakcij med proteini oz. proteini in DNA omogočajo natančno razumevanje delovanja proteina v celici (Cusick in sod., 2005). Za navadnega repnjakovca obstojejo splošne povezave proteinskih interakcij večjega obsega (Arabidopsis Interactome Mapping Consortium, 2011) ali podatki proteinskih interakcij specifičnih za okužbo s patogenom (Mukhtar in sod., 2011), za riž pa interaktom odgovora na stresne dejavnike (Seo in sod., 2011). Obstaja tudi že nekaj baz podatkov, v katere so množice eksperimentalnih podatkov vnesli ročno (Kerrien in sod., 2012; Stark in sod., 2006), pa tudi baze z računalniško napovedanimi interakcijami (Cui in sod., 2008; De Bodt in sod., 2010; Gu in sod., 2011).

V metabolomiki, zadnjem od osnovnih omskih pristopov, veliko doprinesejo netarčne analize, npr. vzorci izražanja znanih in neznanih metabolitov v različnih življenjskih ciklih navadnega repnjakovca (Carroll in sod., 2010; Matsuda in sod., 2010; Sakurai in sod., 2013).

2.3.2 Ontologije in ortologne skupine

Pri analizi podatkov visoko zmogljivih tehnologij je potrebna integracija večih tipov informacije iz raznovrstnih virov, pri čemer so za funkcionalen opis in razjasnitev pojmov pomembne ontologije (Azé in sod., 2008). Te omogočajo sistematski opis znanja preko kontroliranega slovarja pojmov z definirano strukturo, v kateri so ontološki pojmi med seboj povezani z definiranimi odnosi (Klie in Nikoloski, 2012; Schulz in sod., 2011). Vsak gen je povezan z nizom pojmov, ki opisujejo njegovo funkcionalnost – dodatno pa je zaradi definiranih odnosov povezan tudi z vsemi starševskimi pojmi. Pogosto uporabljane in pomembne ontologije rastlinskega področja opisujejo specifične funkcije encimov (Hu in sod., 2012), fenotipske značilnosti rastlin (Arnaud in sod., 2012; Shrestha in sod., 2012) ali rastno okolje rastlin (Buttigieg in sod., 2013).

Za analizo funkcije rastlinskih genov so pomembne ontologije Gene Ontology (GO) (The Gene Ontology Consortium, 2013), FunCat (Ruepp in sod., 2004) in MapMan (Thimm in sod., 2004). Medtem ko sta bili GO in FunCat sprva razviti z namenom karakterizacije mikrobnih sistemov in kasneje postali bolj vrstno nespecifični, je bila ontologija MapMan zasnovana specifično za navadnji repnjakovec (Thimm in sod., 2004). Organizacija pojmov v GO je v obliki treh ne-prekrivajočih se ontologij, ki opisujejo molekularno funkcijo, biološki proces in celično komponento (The Gene Ontology Consortium, 2013). MapMan je za razliko od GO organiziran kot ena sama ontologija (tako kot FunCat), kjer so odnosi med koncepti organizirani v obliki hierarhičnega drevesa in ne acikličnem grafu. Najvišji (najbolj splošni) nivoji ontologije MapMan so podobni biološkim procesom GO, najnižji (najbolj specifični) opisi pa so podobni molekularni funkciji GO.

Pojme standardiziranih terminologij se lahko prenaša med rastlinami, saj so evolucijsko sorodni geni ohranili funkcijo med različnimi organizmi (Du Plessis in sod., 2011). Najbolj zanesljive anotacije določijo strokovnjaki področja (skrbniki, ang. *curators*), iz eksperimentalnih dokazov v osnovni literaturi ali drugih virov (podobnost zaporedij, vnosi v bazah podatkov, pregledni članki). Baumgartner in sod. (2007) so nakazali, da ročne anotacije genov niso zmožne shajati s hitrostjo pridobivanja novih podatkov. Funkcija genov navadnega repnjakovca je s poskusi dokazana le pri 13 % celega genskega modela, ostale funkcije pa so bile določene preko primerjave zaporedij (Lamesch in sod., 2012). Od cele množice genov (skupno okoli 27,000), je približno 70 % vseh proteinsko

kodirajočih genov anotiranih z nekim funkcionalnim pripisom (GO ali MapMan). Za ostale rastline katerih genomi so večji in (večkrat) podvojeni je delež anotiranih genov manjši (Jiao in sod., 2011). Anotacijski proces je zato potrebno avtomatizirati (Azé in sod., 2008) in razviti metode (predvsem) tekstovnega rudarjenja, ki so na podlagi ortologne informacije sposobne določanja funkcije novo odkritih genov in genskih produktov (Wittkop in sod., 2013).

Pomen prenosa znanja med modelnimi rastlinami in agronomsko pomembnimi rastlinami je velik. Ortogni proteini (t.j. homologni proteini, ki so se razvili iz istega prednika ob razvoju nove vrste; ang. *speciation event*) ohranjajo podobno strukturo funkcionalnih domen in ohranjajo isto funkcionalno nišo (Chen in sod., 2007). Informacija o ortolognih skupinah tako predstavlja pomemben vir za prevajanje funkcionalnega znanja med vrstami (Yang in Zhang, 2008). Tak način propagacije znanja med filogenetsko povezanimi vrstami je še vedno eden najboljših pristopov, zaradi dobrega prenosa v večje merilo in zanesljivosti (Rentzsch in Orengo, 2009). Razvrstitev genov v ortologne skupine je odvisna od nabora rastlin na katerem se izvede algoritmom identifikacije ortolognih skupin. Širši kot je izbran nabor rastlinskih vrst (filogenetsko oddaljene rastline) za določanje ortolognih skupin, večje in manj natančne so ortologne skupine. Velja tudi obratno, ožji kot je nabor rastlinskih vrst (filogenetsko bližje rastline), manjše in bolj natančne so skupine (Powell in sod., 2014). Težavo povzročajo tudi paralogni proteini (homologni proteini, ki so se razvili iz istega prednika preko podvojitev; ang. *duplication event*), katerih ločba od ortologov je pomembna za relevantnost rezultatov (Chen in sod., 2007).

Večina skupin, ki je razvila raznolike algoritme določanja filogenetskih razmerij med geni in ortognimi skupinami, je svoje rezultate predstavila tudi v obliki spletnih aplikacij s podporo podatkovnih baz. Tako obstoje baze za rezultate OrthoMCL, InPARANOID, Cluster of Orthologous Group, COG (Tatusov in sod., 2003), i-ADHoRE (Proost in sod., 2012), ipd. Dodatno so tudi rastlinsko specifične baze ali vključile ali pripravile lastne ortologe, npr. Phytozome na podlagi primerjave zaporedij (Goodstein in sod., 2012), Gramene s podporo filogenetskih dreves (Liang in sod., 2008), PlantGDB (Duvick in sod., 2008) ter PLAZA (Proost in sod., 2014). Slednja je še posebej uporabna, saj poleg informacije o filogenetskih odnosih vključuje tudi druge tipe informacij (druge ontologije, funkcionalne pripise delovanja genov).

3 MATERIALI IN METODE

3.1 PRIPRAVA TRANSKRIPTOMSKIH PODATKOV

Dinamiko tolerančne interakcije med rastlino krompirja in PVY smo proučevali na slepo ali virusno inokuliranih (spodnjih) ter ne-inokuliranih (zgornjih) listih kultivarja krompirja Désirée ter transgene rastline nesposobne akumulacije salicilne kisline (NahG-Désirée). Serija vzorcev inokuliranih spodnjih listov šteje 5 časovnih točk (1, 3, 4, 5 in 7 dni po inokulaciji), serija vzorcev ne-inokuliranih zgornjih listov pa 8 (še 9, 10 in 11 dni po inokulaciji). Dodatno so vzorci pobrani še na dan inokulacije (dan 0) za vsakega od bioloških tipov (Désirée inokulirani, NahG-Désirée inokulirani in Désirée ne-inokulirani). Vsi vzorci so pobrani v treh bioloških ponovitvah, z izjemo enega, ki je ponovljen štirikrat – skupno torej 118 vzorcev (Slika 8a). Poskus je bil zastavljen v okviru doktorskega dela Tjaše Stare (Stare in sod., 2015). Dodatno smo v kasnejših korakih integracije uporabili (a ne tudi statistično analizirali) tudi podatke pridobljene v okviru drugega projekta (Baebler in sod., 2014) za kultivarje krompirja PW363, Rywal ter transgeni Rywal (NahG-Rywal) (Slika 8b).

a) Tip vzorca	Status	Časovne točke											b) Tip vzorca	Status	Časovne točke		
		0	1	3	4	5	7	9	10	11	1	3			1	3	6
Désirée inokulirani listi	kontrola			3x	3x	3x	3x	3x	3x		Rywal inokulirani listi	kontrola	4x	4x	4x		
	virus PVY	3x	3x	3x	3x	3x	3x	3x	3x				4x	4x	4x		
NahG-Désirée inokulirani listi	kontrola			3x	3x	3x	3x	3x	3x		NahG-Rywal inokulirani listi	kontrola	4x	4x	4x		
	virus PVY	3x	3x	3x	3x	3x	3x	4x		4x		4x	4x				
Désirée ne-inokulirani listi	kontrola			3x	PW363 inokulirani listi	kontrola	4x	4x	4x								
	virus PVY	3x	3x	3x	3x	3x	3x	3x	3x				4x	4x	4x		

Slika 8: Eksperimentalni načrt proučevanja dinamike tolerančne interakcije med krompirjem in virusom.

Figure 8: Experimental design used to analyse the dynamics of tolerant interaction between potato and the virus.

Hibridizacija ter branje signalov vzorcev RNA so bili opravljeni po postopku vzpostavljenem v IMGM Laboratories GmbH (Nemčija). Enobarvni protokol hibridizacije (s Cy-3) je bil opravljen na mikromreži POCI (AMADID 015425) (Kloosterman in sod., 2008) proizvajalca Agilent, ki pokriva 42,034 biološko relevantnih zaporedij krompirja.

3.1.1.1 Predprocesiranje podatkov

Surovi podatki intenzitet fluorescentnih signalov časovnega odgovora kultivarjev Désirée ter NahG-Désirée na okužbo z virusom so bili analizirani v programskem okolju R (R Project for Statistical Computing; R Development Core Team, 2011; version 2.13.2) z uporabo knjižnic Agi4x44PreProcess (Lopez-Romero, 2012) in Limma (Smyth, 2004).

Za osnovno oceno kvalitete podatkov smo pregledali nekatere izmed standardnih načinov prikazovanja mikromrežnih podatkov: izris intenzitete signala na vsaki mikromreži posebej, izris okvirjev z ročaji (ang. *boxplot*) ter izris gostote medianih intenzitet signala (*gMedianSignal*) oz. ozadja (*gBGMedianSignal*). S tem smo preverili, da ni prišlo do tehničnih napak in od tod izhajajočih artefaktov, kot npr. večjih predelov na mikromrežah, kjer je bilo spiranje slabo ter da ni prišlo do šaržnega efekta. Poleg vizualnega pregleda stanja podatkov smo se poslužili tudi bolj objektivnih merit ocenjevanja, ki jih poda program proizvajalca Agilent, uporabljen za pretvorbo iz slik v tabele intenzitet fluorescentnih signalov. Te kontrolne oznake opisujejo različne kvantificirane napake signala in se jih lahko uporabi za filtriranje signalov, ki niso dosegli minimalnih kriterijev kvalitete (Lopez-Romero, 2012). Vse oznake kvalitete so Boolejeve oblike in imajo lahko vrednost 0 ali 1 (pomen vrednosti pa je odvisen od oznake).

Prvi tip oznak se navezuje na ohranjanje vrednosti intenzitet v dinamičnem območju čitalca. Temu služijo oznake *IsFound* (sonda je najdena), *IsSaturated* (sonda je nasičena) ter *IsWellAboveBG* (sonda je nad intenziteto lokalnega signala ozadja). Drug tip oznak se navezuje na kvaliteto samih sond, kjer uporabljamo oznaki *IsFeatNonUnifOL* (sonda je neenotna – njen šum presega arbitrarno določene vrednosti enotne sonde) ter *IsFeatPopnOL* (sonda je osamelec populacije - njen signal je nižji od spodnjega pragu ali višji od zgornjega določenega pragu, določenega preko interkvartilnega razpona populacije signalov). Na podlagi teh petih oznak smo pripravili matriko uteži, kjer je vsaka sonda nekega vzorca prejela utež 0, če je katera od teh oznak kazala na slabo kvaliteto, drugače pa je bila utežena z vrednostjo 1. Če je bila utež za izbrano mikromrežno sondo enaka 0 pri več kot 90 % vzorcev, je bila slednja izločena iz nadaljnih analiz (Bourgon in sod., 2010).

Naslednji korak pri obdelavi rezultatov mikromrež je normalizacija signalov med vzorci. Večina metod normalizacije predpostavlja, da ob primerjavi različnih vzorcev pri večini genov ni razlik ter da obstaja simetrija med številom induciranih ter represiranih genov. Naši začetni poskusi so ti predpostavki potrdili, zato smo podatke normalizirali ter s tem popravili morebitne razlike tehničnih vzrokov med vzorci. Preizkusili smo tri metode normalizacije: kvantilno (Bolstad in sod., 2003), robustno (ang. *robust spline normalisation*, RSN; Lin in sod., 2008) ter normalizacijo s stabilizacijo variance (ang. *variance stabilising normalisation*, VSN; Huber in sod., 2002).

Kvantilno normalizacijo smo izbrali kot eno izmed najpogosteje uporabljenih metod v obdelavah mikromrežnih podatkov, RSN zaradi njene podobnosti kvantilni normalizaciji, a z manjšimi modifikacijami samih podatkov, VSN pa kot tretjo bolj oddaljeno metodo iz vidika samega delovanja. Kvantilna normalizacija deluje na osnovi kvantilov, njen končni rezultat pa so enake porazdelitve signalov med različnimi vzorci. Tako RSN in VSN metodi za razliko od kvantilne normalizacije kot vhodne podatke uporabljata transformirane podatke (npr. logaritmirane). RSN združuje lastnosti kvantilne in loess normalizacije, medtem ko VSN dinamično menja med uporabo naravnega logaritma (visoke intenzitete signalov), linearne transformacije (nizke intenzitete signalov) ter interpolacije med obema (za vmesni razpon).

Po normalizaciji brez odstevanja ozadja je ponovno sledila kontrola kvalitete v obliki izrisa okvirjev z ročaji ter gostote porazdelitev intenzitet signala pred ter po normalizaciji. Dodatno smo za začetni pregled podatkov vzorce uredili v skupine z metodo razvrščanja (ang. *hierarchical clustering*). Vzorce smo primerjali glede na razlike ocenjene z Evklidsko razdaljo izračunane med vsemi kombinacijami vzorcev. Rezultate metode razvrščanja smo prikazali v obliki toplotnih grafov (ang. *heatmap*) brez ali z urejanjem (preko izračuna Pearsonove korelacije).

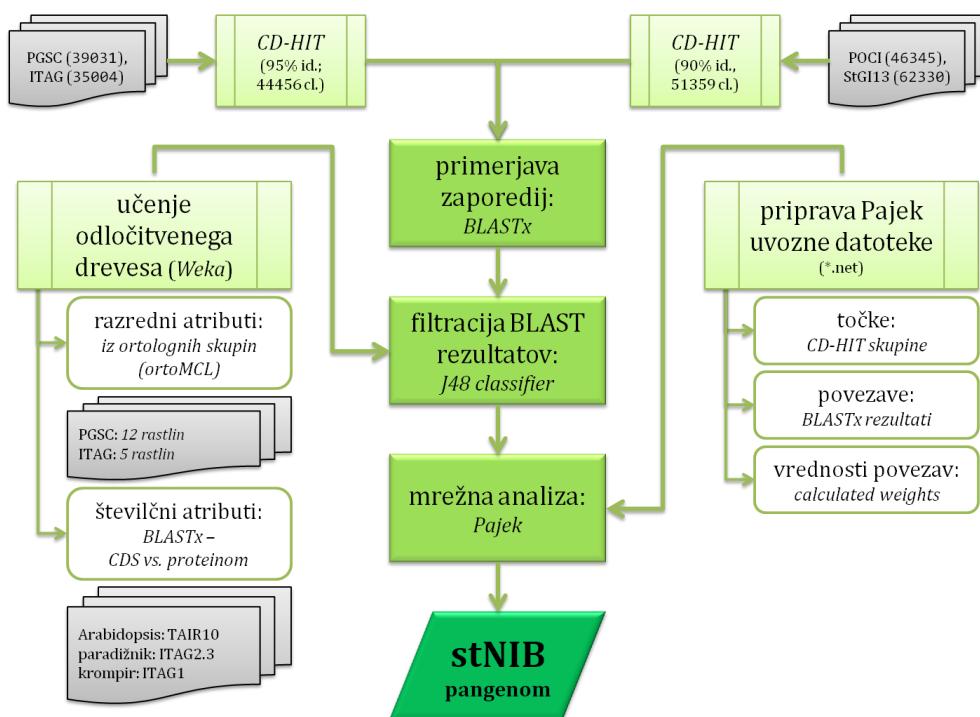
3.1.1.2 Priprava linearnega modela ter primerjave med različnimi pogoji

Za detekcijo statistično značilnih razlik v izražanju genov med različnimi pogoji smo uporabili empirično Bayesovo metodo (Smyth, 2004). Razlike smo opazovali med vzorci kontrolnimi in okuženimi rastlinami krompirja za vsako časovno točko in tip vzorca za vse tri tipe normalizacije. Za izbor značilno drugačno izraženih genov smo uporabili t-test (od povprečja virusno okuženih rastlin smo odštevali povprečje kontrol) s deležem napačnih uvrstitev (ang. *false discovery rate*, FDR; (Benjamini in Hochberg, 2009) pri prazni vrednosti FDR- $p < 0.05$ za vse primerjave in izrise. Želeli smo preveriti tudi razlike med različnimi tipi izbranih normalizacij, zato smo za vsako primerjavo pripravili izris v obliki Vennovih diagramov.

3.2 INTEGRACIJA ZAPOREDIJ V OKVIRU ENE RASTLINSKE VRSTE

3.2.1 Priprava pan-genoma krompirja

Za pripravo pan-genoma krompirja smo združili genska modela PGSC (The Potato Genome Sequencing Consortium, 2011) in ITAG (The Tomato Genome Consortium, 2012), katerima smo dodali znana zaporedja dveh množic unigene na osnovi tehnologije EST. Glede na datum zadnje posodobitve ter internih izkušenj s kvaliteto zaporedij smo izbrali zaporedja unigene StGI (Quackenbush in sod., 2001) ter POCI (Kloosterman in sod., 2008). Po korakih, prikazanih na Slika 9 in podrobneje opisanih v spodnjih štirih poglavjih, smo pridobili paralogne genske skupine krompirja.



Slika 9: Koraki združevanja večih različnih množic zaporedij v skupne paralogne genske skupine.
 Celoten postopek vsebuje štiri glavne korake – v prvem se zmanjša redundanco množic zaporedij po skupinah (genski modeli skupaj, množice EST skupaj; algoritem CD-HIT). Sledi primerjava zaporedij zmanjšanih množic zaporedij med skupinama s primernim algoritmom (algoritem BLASTx). V tretjem koraku se s pomočjo pristopa učenja odločitvenih dreves zmanjša količino rezultatov (klasifikator J48), nakar v četrtem koraku sledi analiza omrežij povezav BLASTx (program Pajek).

Figure 9: Steps required to combine various sequence sets into parologue gene groups.

The whole procedure is composed of four main steps – in the first one we reduce the dataset(s) redundancy separately for gene models and EST-derived datasets (e.g. using the CD-HIT algorithm). Sequence comparison of the reduced set follows with an appropriate sequence comparison algorithm (BLASTx in our case). The third step involves learning decision trees to reduce (filter-out) the result set (J48 classifier), and was followed by network analysis of the BLASTx connections (Pajek).

3.2.1.1 Zmanjšanje redundancije v množicah podatkov

Da bi povezali oba genska modela ter zmanjšali redundanco v obeh množicah zaporedij unigene, smo uporabili algoritem CD-HIT (Fu in sod., 2012; Li in Godzik, 2006). Slednji uredi seznam danih zaporedij po njihovi dolžini, in primerja najdaljšega z ostalimi. Če je podobnost med zaporedjima enaka ali višja od uporabniško izbrane pražne vrednosti podobnosti, se zaporedji združi v skupino in najdaljše zaporedje postane reprezentativno zaporedje skupka. Ta postopek se nato rekurzivno ponovi za vse ostala zaporedja. Za potrebe združevanja genskih modelov smo uporabili pražno vrednost podobnosti 95 % (vsa zaporedja predstavljajo isti kultivar), za združevanje zaporedij unigene pa 90 % (zaporedja predstavljajo različne kultivarje).

3.2.1.2 Primerjava zmanjšanih množic zaporedij

Rezultate algoritma CD-HIT za genske modele in množici unigene smo v naslednjem koraku primerjali med seboj z uporabo algoritma BLAST (Altschul in sod., 1997). Ker sta za genske modele skupini PGSC in ITAG napovedali tudi zaporedja končnih proteinov, smo lahko uporabili algoritem BLASTx (verzija 2.2.27+). Le-ta primerja nukleotidna zaporedja prevedena v vseh šestih bralnih okvirjih (rezultati CD-HIT zaporedij EST) s proteinskimi zaporedji v bazi (rezultati algoritma CD-HIT na zaporedjih genskih modelov). Da bi pokrili kar največji spekter potencialno funkcionalno povezanih zaporedij, so bili kriteriji uporabljeni pri analizi z algoritmom BLAST zelo blagi: dovolili smo do 100 najboljših zadetkov za vsako iskalno zaporedje in pričakovano vrednost (ang. *expect value*, E-vrednost) nižjo od 0,1.

3.2.1.3 Zmanjšanje števila genskih skupkov preko učenja odločitvenega drevesa

Da bi se znebili lažno pozitivnih rezultatov algoritma BLAST, smo se poslužili učenja odločitvenih dreves z uporabo treh različnih algoritmov, J48 (Quinlan, 1994), ADTree (Freund in Mason, 1999) in JRip (Cohen, 1994), v programu WEKA (Hall in sod., 2009).

V prvem koraku je bilo potrebno vsem trem algoritmom pripraviti t.i. učno množico (ang. *training set*) za kar smo uporabili skupine algoritma orthoMCL dvanajstih rastlin konzorcija PGSC (The Potato Genome Sequencing Consortium, 2011) ter rezultate algoritma orthoMCL petih rastlin konzorcija ITAG (The Tomato Genome Consortium, 2012). Zadržali smo le tiste ortologne skupine, ki so vsebovale vsaj dva gena katerekoli od rastlin navadnega repnjakovca, krompirja ali paradižnika. Za obdržane gene smo pridobili kodirajoča zaporedja genov (ang. *coding DNA sequence*, CDS) ter njihova proteinska zaporedja. Ta množica proteinskih zaporedij je postala baza BLAST, proti kateri smo primerjali množico kodirajočih zaporedij z algoritmom BLASTx. Številčne vrednosti rezultatov primerjav, t.j. E-vrednost, odstotke identitete, podobnosti, vrzeli ter pokritja, smo definirali kot številčne attribute (ang. *numerical attributes*) učne množice.

Vrednosti razrednega atributa (ang. *class attribute*; je ortolog / ni ortolog), smo izpeljali iz samih ortolognih skupin orthoMCL. V okviru vsake ortologne skupine smo ustvarili vse možne kombinacije brez ponavljanja. Če je skupina vsebovala en gen rastline

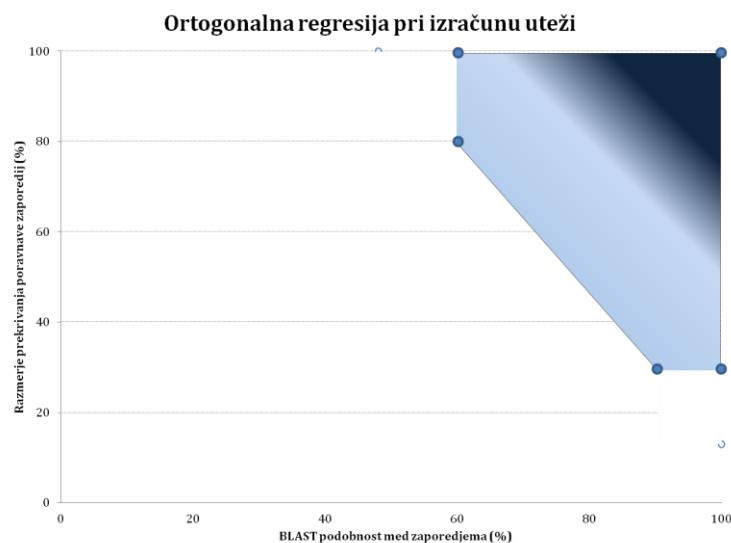
navadnega repnjakovca (*ath1*), en gen krompirja (*stu1*) ter en gen paradižnika (*sly1*), je bil rezultat kombinacij brez ponavljanja sledeč: *ath1-stu1*, *ath1-sly1*, *stu1-sly1*. Za pripravo učne množice smo nato združili rezultate BLASTx med izbranimi zaporedji CDS teh treh rastlin in kombinacij. Vsak zadetek BLASTx iz seznama naših kombinacij je dobil razredni atribut 'je ortolog'.

Učna množica za vse tri algoritme je vsebovala 310,179 povezav BLASTx med zaporedjem sonde in tarče, oznako razrednega atributa povezave ter vse številčne attribute. Porazdelitev razrednega atributa je bila 30 % za 'je ortolog' ter preostanek za 'ni ortolog'. Podatke je algoritem v WEKI sam razdelil na učno (66 %) in testno množico (33 %). Odločitveno drevo algoritma J48 smo prevedli v Perl skripto, ki je bila uporabljeni za filtriranje BLASTx lažno pozitivnih rezultatov med skupinami CD-HIT genskih modelov ter zaporedij EST iz poglavja 3.2.1.2. Odločitveni drevesi algoritmov ADTree in JRip sta bili manj obsežni in tako vmesnega koraka prevoda nista potrebovali, dodatno pa nista bili uporabljeni v kasnejših korakih filtriranja rezultatov.

3.2.1.4 Priprava končnih paralognih skupin v programu Pajek

Preostali zadetki BLASTx so bili prenešeni v korak analize omrežij v programu Pajek (Batagelj in Mrvar, 1998). Vsak rezultat BLASTx je bil pretvorjen v neusmerjeno povezavo med zaporedjem sonde in tarče. Dodatno smo za vsako povezavo izračunali njeno vrednost uteži, za kar smo uporabili odstotek podobnosti med dvema zaporedjima ter odstotek pokritosti poravnanih zaporedij glede na krajše od obeh zaporedij. Obenem smo želeli tudi, da je utež odvisna od razmerja med pokritjem in podobnostjo zaporedij: za visoko pokritost poravnave smo dovolili manjšo podobnost med zaporedjima in obratno (Slika 10). Na podlagi treh določenih točk, A (100 % podobnost, 100 % pokritje, utež 1), B (60 % podobnost, 80 % pokritje, utež 0) ter C (90 % podobnost, 30 % pokritje, utež 1) smo izračunali enačbo, na podlagi katere smo lahko izračunali vrednost uteži za vse povezave tri-dimenzionalnega prostora.

$$\text{Utež} = -2,1 + 0,02 \times \text{Podobnost \%} + 0,01 \times \text{Pokritje \%} \quad \dots (1)$$



Slika 10: Izbrane točke uporabljeni v enačbi izračuna uteži omrežja genov krompirja.

Predstavljene so točke uporabljeni za izračun uteži v tri-dimenzionalnem prostoru. Če je imel zadetek BLASTx 100 % podobnost zaporedij ter 100 % pokritost zaporedij, je bila vrednost uteži nastavljena na 1. Zahtevano je bilo tudi, da je odstotek podobnosti zaporedij vedno večji kot 60 % ter odstotek pokritosti manjšega od obeh zaporedij vedno večji od 30 %. To je na sliki prikazano z odtenki modre barve na grafu, temnejša kot je, višja je vrednost izračunane uteži.

Figure 10: Selected points used to calculate weights in the potato gene network.

The points used for fitting the plane using orthogonal distance regression planes in the three-dimensional space are represented. If a BLASTx match had 100 % sequence similarity and 100 % sequence coverage, then the value of the weight was set to 1. In general, sequence similarity percentage was set to be larger than 60 % at all times, and the smaller sequence coverage larger than 30 %. This is indicated by the shading of the blue colour on the graph, the darker the colour, the higher the assigned weight.

V omrežje smo vključili le povezave s pozitivnimi vrednostmi uteži. Uteži so pri analizi omrežij služile pri rekurzivni dekompoziciji; z odstranjevanjem vseh povezav pod nekim pragom smo zaporedno izolirali t.i. paralogne genske skupine krompirja. Vsaka paralogna skupina je vsebovala neko kombinacijo identifikatorjev maksimalno štirih tipov (ITAG, PGSC, POCI ter StGI13) in vsaki smo dodelili reprezentativen identifikator po naslednji hierarhiji odločitev:

- (1) Če je skupina vsebovala točno eno zaporedje ITAG, je to postalo reprezentativno.
- (2) Če je skupina vsebovala več kot eno zaporedje ITAG, je reprezentativno postalo tisto z najdaljšim zaporedjem CDS (oz. zaporedjem EST v kasnejših korakih).
- (3) Če skupina ni vsebovala zaporedij ITAG, pa je vsebovala zaporedje PGSC se je koraka 1. in 2. ponovilo za skupino PGSC.
- (4) Če skupina ni vsebovala zaporedij ITAG ali PGSC, se je odločalo na podlagi zaporedij POCI, tudi s korakoma 1. in 2.
- (5) V zadnjem koraku se je odločalo na podlagi zaporedij StGI13.

3.2.2 Priprava pan-transkriptoma tobaka in povezovanje z mikromrežnimi zasnovami

Preden je bilo določeno zaporedje genoma tobaka (Sierro in sod., 2014) smo leta 2011 za potrebe projekta žeeli pripraviti skupen seznam tobakovih unigenov. Združili smo zaporedja iz različnih virov: podatkovne baze GenBank in na njeni osnovi zgrajene baze NCBI UniGene v17 (Benson in sod., 2014), rezultate Evropskega projekta sekvenciranja tobaka na nivoju cDNA ESTobacco (Edwards in sod., 2010), podatkovne baze TIGR (Childs in sod., 2007) ter podatkovne baze Dana Farber NtGI 7.0 (Quackenbush in sod., 2001). Za večje pokritje obstoječega znanja in kasnejšo integracijo z eksperimentalnimi podatki, smo dodali še skupine zaporedij iz dveh komercialnih mikromrež: proizvajalcev Affymetrix (ArrayExpress A-AFFY-135; Edwards in sod., 2010) ter Agilent (ArrayExpress A-MEXP-1963; Soitamo in sod., 2011).

Za združevanje in zmanjševanje redundance v množicah zaporedij z algoritmom CD-HIT (Li in Godzik, 2006) smo v primeru baz GenBank, TIGR in DFCI uporabili zaporedja prisotna v preseku zaporedij med NCBI UniGene v17 in unijo zaporedij obih mikromrežnih zasnov. Reprezentativna zaporedja preostala po združevanju z algoritmom CD-HIT pri pragu 95 % podobnosti smo v naslednjem koraku primerjali proti proteomom štirih rastlinskih vrst (navadnega repnjakovca, krompirja, paradižnika in riža) z uporabo algoritma BLASTx (Altschul in sod., 1997). Namesto določanja podobnosti z uporabo naučenega drevesa J48 smo tu določili prazne vrednosti za nekatere številske vrednosti rezultatov BLASTx: BitScore večja od 50, E-vrednost nižja od 10^{-50} , odstotek identitet zaporedij višji od 80 % ter odstotek pokritja poravnave glede na krajše zaporedje vsaj 70 %. Veliko tobakovih zaporedij ni imelo človeku prijaznih opisov anotacij genov, zato smo preko najboljšega zadetka zaporedja tobakovega gena z eno izmed teh štirih rastlin tudi pripisali opis glede na vrednost BitScore (podoben pristop k opisom ima orodje Mercator (Lohse in sod., 2013)). Vsako zaporedje je podedovalo opis prvega najboljšega zadetka ter opis kvalitete podobnosti:

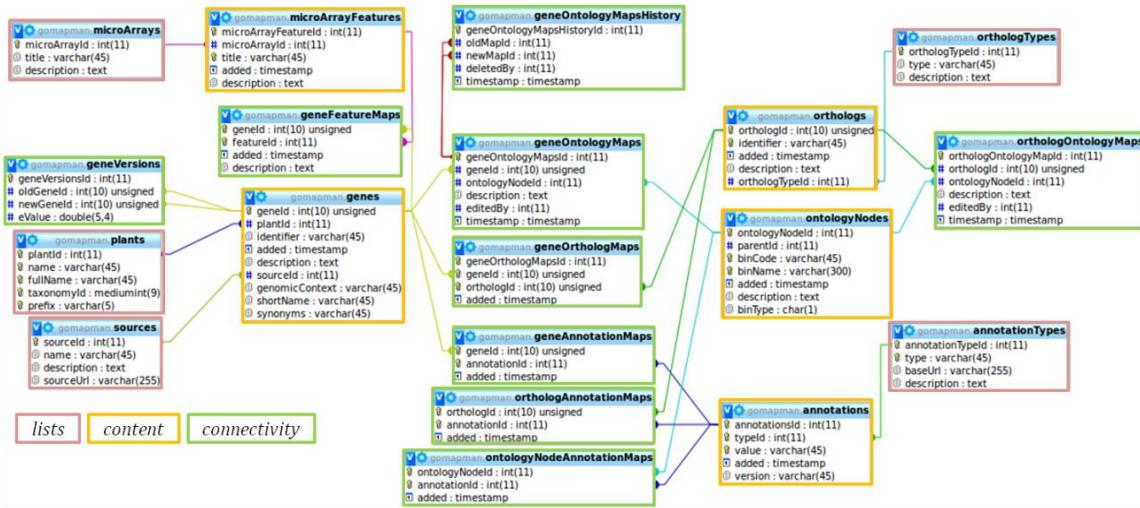
- (1) *zelo šibko podoben*: BitScore med 50 in 100
- (2) *šibko podoben*: med 101 in 200
- (3) *zmerno podoben*: med 201 in 500
- (4) *visoko podoben*: med 501 in 1000
- (5) *skoraj identičen*: BitScore večja od 1001

3.3 GOMAPMAN: POVEZOVANJE ZNANJA MED RASTLINSKIMI VRSTAMI

Aplikacija GoMapMan je postavljena na serverju Apache HTTP z operacijskim sistemom Fedora. Podatki so shranjeni v relacijskem sistemu za upravljanje s podatkovnimi bazami MySQL (podatkovna baza *gomapman*). Dodatno smo za potrebe dinamičnega spletnega vmesnika uporabili orodja Drupal, PHP in JavaScript. Za potrebe večjih posegov v podatke so implementirane t.i. skripte za paketno obdelavo, zasnovane v jeziku Bash.

Podatkovna baza *gomapman* vsebuje vso informacijo povezano z biološkimi podatki uporabljenimi za aplikacijo GoMapMan, dodatno pa aplikacija vsebuje se manjšo t.i. *gomapman_dr* bazo, ki vsebuje informacijo o uporabnikih – za sledljivost sprememb podatkov. Glavne tabele v bazi *gomapman* se lahko okvirno razdeli v tri večje razrede (Slika 11):

- (1) Seznami entitet: Te tabele povezujejo sezname glavnih vključenih entitet z originalnim virom, najpogosteje predstavljenim v obliki internetne povezave (npr. genski model navadnega repnjakovca TAIR9, mikromrežna zasnova krompirja POCI, anotacijska tipa GO ali PO). Te tabele so v bazi najmanjše (pod 100 vnosov) in se tudi redko spreminja, saj se nove entitete dodaja le nekajkrat na leto.
- (2) Vsebinske tabele: Vsebujejo sezname glavnih vključenih entitet (geni, ortologne skupine, ontološka vozlišča oz. razdelki (ang. *BIN*) ontologije MapMan, anotacije, mikromrežne sonde) in nekaterimi izbranimi informacijskimi predeli (npr. kratek simbol gena, opis mikromrežne sonde, opisi razdelkov ontologije MapMan s spremljajočim opisom, opisi pomena anotacije).
- (3) Povezujoče tabele: Med seboj povezujejo vsebinske tabele in se se dalje ločijo na tri večje odseke. V prvem odseku so tabele, ki povezujejo anotacije z geni ali ontološkimi razdelki; v drugem odseku so tabele, ki povezujejo ontološke razdelke z geni, ortolognimi skupinami ter zgodovino; v tretjem odseku pa so tabele, ki povezujejo gene z mikromrežnimi sondami, gene z ortologi ali gene z verzijo genskega modela.



Slika 11: Relacijski diagram (ER) za bazo gomapman.

Konceptualna shema aplikacije GoMapMan je sestavljena iz 19 ločenih tabel. Najmanjše tabele (rdeči okvir) povezujejo sezname glavnih vključenih entitet z originalnim virom. Vsebinske tabele (rumeni okvir) vsebujejo sezname z dodatnimi informacijami glavnih entitet. Povezujoče tabele (zeleni okvir) povezujejo vsebinske tabele med seboj.

Figure 11: Entity-relationship (ER) diagram for gomapman database.

The GoMapMan conceptual schema is composed of 19 separate tables. A purple frame marks the five content-wise smallest tables, which connect lists of main included entities with the original data source. Content tables which contain lists of main entities with additional information on them are marked with a yellow frame around them. Connecting tables (green) connect the content tables between themselves.

Povezujoče tabele so pogosto t.i. m×n tabele, saj ima lahko en gen ali ortolog več različnih položajev v ontologiji MapMan, več različnih anotacij ali en gen več mikromrežnih povezav ter obratno. Položaj v ontologiji je posebej določen le za gene in ortologe, ne pa tudi za mikromrežne identifikatorje. Slednji podedujejo informacijo o vozliščih MapMan iz starševskega gena (oz. genov).

3.3.1 Glavni gradniki podatkovne baze GoMapMan

3.3.1.1 Ontologija MapMan

Ontologija MapMan predstavlja sistem hierarhičnega mapiranja, ločen na 35 osnovnih t.i. funkcionalnih pripisov (ang. *BINs*), ki opisujejo najvišje nivoje različnih bioloških procesov. Vsako funkcionalni pripis se na hierarhičen način dalje deli v več podrazdelkov, ki predstavljajo bolj specifične opise funkcije. Za pripravo ontologije MapMan, ki smo jo vključili v aplikacijo GoMapMan, smo se poslužili javno dostopnih izvoznih datotek dostopnih na uradni strani aplikacije MapMan. Datoteke so ločene glede na rastlinsko vrsto ter verzijo genskega vira (za navadni repnjakovec glede na njegove genske modele TAIR9, TAIR10 ipd.). Uporabili smo vse izvozne datoteke, ki so bile dostopne do marca 2012, obenem pa dodali se izvozno datoteko mikromreže POCI, katera je bila v velikem obsegu ročno popravljena na oddelku (Rotter in sod., 2007). Ontološka drevesa smo sinhronizirali tako, da smo odstranili artefakte same priprave (razlike v imenih razdelkov) ter pripravili prevajalno datoteko za vsa zastarela imena.

3.3.1.2 Rastlinske vrste

V aplikacijo GoMapMan smo vključili genske modele petih rastlin, pan-genom krompirja in pan-transkriptom tobaka. Podatki o rastlinskih vrstah so naštetи v Preglednici 2.

Preglednica 2: Rastlinske vrste vključene v aplikaciji GoMapMan.

V tabeli so podane vse rastline, s predponami in latinskimimi imeni, ter ime genomskega modela. Genomi vseh rastlin, z izjemo krompirja in tobaka, so bili vnešeni v izvorni obliki, pan-genom krompirja (stNIB-v1) ter pan-transkriptom tobaka (ntaUG17) pa sta bila pripravljena po naših postopkih (glej Poglavlje 3.2).

Table 2: Plant species which are included in GoMapMan.

The table lists the plants, with their prefixes and latin names, and the genome model. Note that all genome but potato and tobacco are left unmodified as in the original form, while the potato pan-genome (stNIB-v1) and tobacco pan-transcriptome (ntaUG17) were prepared using our procedures (see Chapter 3.2).

Rastlina	Latinsko ime	Predpona	Genski model	Leto/mesec dodaje
navadni repnjakovec	<i>Arabidopsis thaliana</i>	ath	TAIR9	2012-09
riž	<i>Oryza sativa</i>	osa	MSU_v7	2012-10
krompir	<i>Solanum tuberosum</i>	stu	stNIB-v1	2013-03
paradižnik	<i>Solanum lycopersicum</i>	sly	SL2.40/ITAG2.3	2013-03
tobak	<i>Nicotiana tabacum</i>	nta	ntaUG17	2014-04
kakav	<i>Theobroma cacao</i>	tca	Phytozome9.1	2015-01
sladkorna pesa	<i>Beta vulgaris</i>	bvu	RefBeet1.1	2015-01

Poleg genomske informacije so v bazo vključene tudi razširitve za nekatere transkriptome ter povezave z nekaterimi izbranimi mikromrežnimi zasnovami (Preglednica 3). Za razliko od transkriptomov, kjer je tip povezave $1 \times n$ (en gen ima lahko enega ali več transkriptov, ne pa tudi obratno) je pri mikromrežah še dodaten zaplet v dejstvu, da so tu povezave $m \times n$ oblike (en gen ima lahko več mikromrežnih sond, obenem pa lahko sonda pokriva tudi več različnih genov). Tako transkripti kot sonde mikromrežnih zasnov nimajo lastnih ontoloških povezav z ontologijo, ampak informacijo podedujejo od starševskih genov.

Preglednica 3: Transkriptomi in mikromrežne zaslove v aplikaciji GoMapMan.

V tabeli so podane vse rastline (do junija 2015), z njihovimi predponami in latinskim imeni, dodatno pa se seznam vseh transkriptomov ali identifikatorjev mikromrežnih zasnov.

Table 3: Transcriptomes and microarray platforms included in GoMapMan.

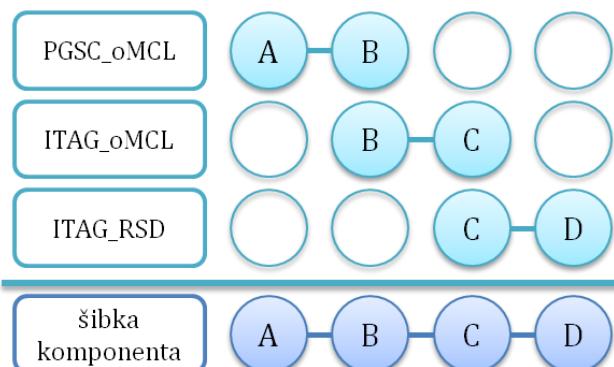
The table lists the plants (until June 2015), with their prefixes and latin names, while in addition listing the transcriptomes or platform design identifiers.

Rastlina	Latinsko ime	Predpona	Tip zaslove	Leto/mesec dodaje
krompir	<i>Solanum tuberosum</i>	stu	Agilent-015425 (POCI)	2013-03
tobak	<i>Nicotiana tabacum</i>	nta	A-AFFY-135	2014-04
tobak	<i>Nicotiana tabacum</i>	nta	A-MEXP-1963	2014-04
kakav	<i>Theobroma cacao</i>	tca	Phytozome9.1 transkriptom	2015-01
sladkorna pesa	<i>Beta vulgaris</i>	bvu	RefBeet1.1 transkriptom	2015-01

3.3.1.3 Ortologne skupine

Zadnji nujni gradnik baze predstavlja vključitev različnih tipov ortolognih skupin. Vsak različen algoritmom določanja ortolognih skupin deluje na sebi specifičen način in rezultati algoritmov (končne skupine napovedanih ortologov) so močno odvisni od izbranih rastlin in njihovih medsebojnih filogenetskih razdalj. S tem razlogom smo v bazo GoMapMan vključili različne tipe ortolognih skupin: dva niza skupin pridobljenih z uporabo orthoMCL algoritma (Li in sod., 2003), in sicer konzorcija PGSC na proteomih dvanajstih rastlin ter konzorcija ITAG na proteomih petih rastlin. Dodane so ortologne skupine pridobljene z uporabo metode recipročne-najmanjše razdalje RSD (Wall in Deluca, 2007) na proteomih štirih rastlinskih vrst. Kot zadnje so vključeni tudi širje ortologni tipi skupine PLAZA (Proost in sod., 2014), tako družine kot pod-družine eno- in dvokaličnic.

Ker vsak niz ortolognih skupkov vsebuje različne skupine genov, odvisne od rastlinskih vrst izbranih za pripravo množice ter njihovih medsebojnih evolucijskih razlik, smo dodatno tri implementirane ortologne nize (PGSC_oMCL, ITAG_oMCL in ITAG_RSD) združili s pomočjo analize omrežij v programu Pajek (Batagelj in Mrvar, 1998). V usmerjenih grafih lahko ločimo med močnimi in šibkimi komponentami: prve predstavljajo maksimalni podgraf v katerem obstoje poti med vsako kombinacijo točk grafa preko poti po vseh usmerjenih povezavah. Za razliko od močne komponente, je šibka komponenta maksimalni podgraf brez upoštevanja usmerjenosti povezav. V tem primeru je bil graf neusmerjen in so bile tako šibke in močne komponente enake. Princip združevanja različnih ortolognih tipov v šibke komponente je prikazan na Sliki 12.



Slika 12: Predstavitev tvorbe šibkih komponent iz treh nizov ortolognih skupkov.

Svetlo modri okvirji na sliki predstavljajo različne nize ortolognih skupin implementiranih v podatkovni bazi (zgornji trije) oz. skupine, ki so bile rezultat analize šibkih komponent. Črke predstavljajo gene, katere so različni algoritmi določanja ortolognih skupin in izbranih rastlinskih vrst uvrstili v drugačne ortologne skupine. Končni rezultat ene šibke komponente predstavlja združek vseh različnih povezav med geni, kar olajša popravljanje položajev v ontologiji.

Figure 12: Representation of the weak component creation from three types of orthologous groupings.

Light blue boxes on the image represent different orthologous groupings implemented in the database (top three) or the resulting grouping after weak component formation. The letters represent genes, for which different orthology detection algorithms and plant species selections resulted in the formation of different orthologous groups. The end result of a weak component is the merger of all these different connections, which enables easier correction of plant ontology assignments.

3.3.1.4 Povezovanje z znanjem v drugih javnih bazah podatkov

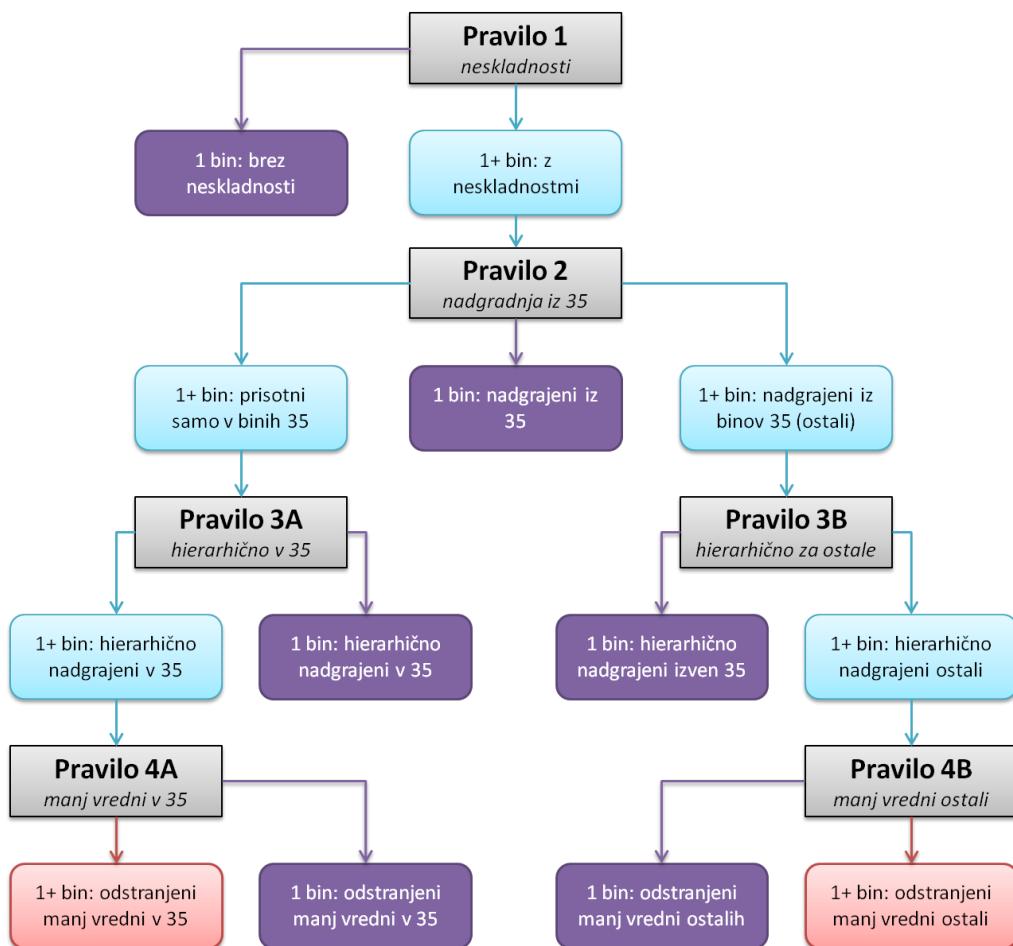
V bazi so kot vir zunanje informacije vključene dodatne informacije o genih (Entrez Gene), iz ostalih ontologij (Gene Ontology GO, Plant Ontology PO), literature (PubMed), baze KEGG ter precej različnih virov v zvezi s strukturo in funkcijo proteinov (EC, Gene3D, HAMAP, InterPro, PANTHER, PFAM, PIR, ProDom, PROSITE, SMART, SPRINT, SUPERFAMILY, TIGRFAM). V pomoč pri pridobivanju zunanjih anotacij so nam bile aplikacije, ki že same zbirajo gensko informacijo iz različnih virov, specifično BioMart (Smedley in sod., 2009) ter z njim povezan portal Ensembl Plants (Kersey in sod., 2014).

3.3.2 Funkcionalnosti aplikacije GoMapMan

3.3.2.1 Proces avtomatske konsolidacije genskih funkcionalnih pripisov

Po naselitvi osnovnih gradnikov podatkovne baze (geni, ortologi, ontološko drevo), je bilo potrebno vzpostaviti povezave med geni in ontološkim drevesom. Informacijo o teh povezavah smo pridobili iz izvornih izvoznih MapMan datotek: TAIR9 za navadni repnjakovec, MSU_v7 za riž, ITAG2.3 za paradižnik ter PGSC za krompir. Dodatno smo imeli za krompir lastne interne izvozne datoteke pripravljene za mikromreže POCI (Rotter in sod., 2007), kjer je bilo narejeno veliko popravkov v povezavi z biotskim stresom.

Za avtomatsko konsolidacijo funkcijskih pripisov genov pri različnih rastlinskih vrstah smo pripravili niz odločitvenih pravil, ki pripomorejo h prenosu znanja med rastlinskimi vrst preko uporabe znanja v ortolognih genskih skupinah. Odločitvena pravila preverijo možnost popravkov uvrstitve določene ortologne skupine v posamezen funkcionalni pripis glede na ontološko anotacijo genov v ortologni skupini. V predpripravah se združi ontološke pozicije za vsak gen neke ortologne skupine. Če je končno unikatno število razdelkov ortologne skupine enako 1, potem so geni označeni, da nimajo neskladnosti in ne vstopijo v naslednje korake konsolidacije, preostanek pa vstopi v postopek avtomatske konsolidacije (Slika 13).



Slika 13: Shema niza konsolidacijskih odločitvenih pravil za izboljšavo genskih funkcionalnih pripisov implementirana v aplikaciji GoMapMan.

Sivo obarvani okvirji slike predstavljajo odločitvena pravila, svetlo modri okvirji datoteke generirane med postopkom konsolidacije, vijolično so končni rezultati konsolidacije, v rdečih okvirjih pa konsolidirani rezultati, ki pa potrebujejo ročen pregled ekspertov področja. Geni brez funkcionalnega pripisa (razdelki skupine 35) predstavljajo eno od podskupin za pravili 3 in 4. Če točno ena od anotacij ni v skupini 35, potem je bila skupina nadgrajena iz 35 in zaključila s postopkom. Če je anotacija različna od 35 več, potem ta skupina nadaljuje v naslednje korake konsolidacije. Pravili 3 in 4 nato nadgradita ortologne skupine v nižje in bolj natančne hierarhične nivoje ter odstranita manjvredne razdelke (opisane kot 'mešano'). Ortologne skupine, ki se vedno vsebujejo več kot en sam končni funkcionalni pripis, se označi za ročni pregled (rdeče), ostale rezultate (vijolično) pa pretvori v paket ukazov, ki popravijo anotacije genov.

Figure 13: Consolidation scheme for improvement of gene functional annotations as implemented in GoMapMan.

Grey coloured boxes denote the decision rules, light blue boxes the intermediate files of the scheme, purple denote final results of the consolidation and red consolidated results, which however still require a manual check by an expert. If all the BINs are marked as unknown, not annotated or previously not described (assigned to BIN 35), genes have an unknown function and represent a subset for rules 3 and 4. If exactly one of the ontology positions does not equal BIN 35, then genes are deleted from BIN 35 and conclude with the procedure. If there are several non-group-35 BINs, then this group continues with further consolidation steps. Rules 3 and 4 then upgrade orthologous groups into lower fine grained hierarchical levels and remove BINs of lesser value (described as miscellaneous). Orthologous groups, that still contain more than one final BIN, are then marked for manual checks (red), while the remaining results (purple) are converted into batch commands which correct the gene annotations.

Odločitvena pravila lahko podajo enega izmed dveh možnih izidov: seznam ortolognih skupin, ki jih pripravljene skripte avtomatsko popravijo (označene vijolično na Sliki 13) ali pa seznam ortolognih skupin, katere mora znanstvenik področja preveriti ročno (označene rdeče na Sliki 13). Možni so trije različni izidi, ki odražajo verodostojnost in kvaliteto funkcionalnih pripisov gena oz. genov ortologne skupine:

- (1) Tip 1: Predstavljajo ga šibke komponente oz. ortologne skupine, v katerih so vsi geni skupine opisani z enim samim funkcionalnim opisom. Tudi vse šibke komponente, ki vsebujejo en sam gen nosijo to oznako. Pri tem tipu možnosti avtomatskih popravkov ontologije ni, saj si vsi geni delijo enake ontološke opise.
- (2) Tip 2: Vsebuje šibke komponente oz. ortologne skupine, ki so prisotne v več kot enem razdelku, ampak je prvi nivo vsakega od teh različnih razdelkov drugačen (prvi nivo razdelka se znotraj skupine nikoli ne ponovi večkrat). Te so potencialno pravilno anotirane in popravljene na nivoju vseh genov.
- (3) Tip 3: Šibke komponente oz. ortologne skupine, kjer se prvi nivo razdelkov ponovi večkrat pri vsaj enem izmed skupine razdelkov. Anotacije le-teh se ne konsolidirajo, ampak geni obdržijo anotacije glede na osnovne izvozne datoteke. Te komponente predstavljajo najmanjši delež glede na celoto, in jih je potrebno ročno pregledati (ekspert področja).

Avtomatske popravke se opravi ob vsaki novi nadgradnji baze, kot npr. dodaji novih rastlin ali še bolj pogosto, ortolognih skupin, saj slednje potencialno doprinesejo dodatne povezave med geni rastlin. Ko pride do potrebe po konsolidaciji se bazo zaklene vsem registriranim uporabnikom, pridobi najnovejše položaje genov v ontologiji (da se postopek opravi na najbolj relevantnih podatkih) in odklene, ko so premiki v bazi končani.

Vir funkcionalnega opisa gena ali ortologne skupine je nakazan z uporabo dokaznih oznak (ang. *evidence code*), katere so enake oznakam, ki jih uporablja GO. Dokaznih oznak je 21 (Preglednica 4), od teh je samo elektronska anotacija (ang. *inferred from electronic annotation*, IEA) avtomatsko dodeljena, vse ostale dodeli skrbnik. Dokazne oznake niso sinonim za kvaliteto anotacije, saj ima vsaka dokazna oznaka lahko več različnih tehnik pridobivanja anotacij, dodatno pa vpliva tudi sama interpretacija rezultata iz samega članka.

Preglednica 4: Seznam dokaznih oznak kot jih uporablja Gene Ontology (GO).
Vsega skupaj je 21 dokaznih oznak, kjer vse razen ene (IEA) dodeli skrbnik po pregledu podatkov.

Table 4: The list of evidence codes as are being used in Gene Ontology (GO).
There are a total of 21 evidence codes, all but one (IEA) derived from curator overview.

Dokazna oznaka	Opis	Skupina
EXP	poskus	rezultat poskusa
IDA	direktni test	rezultat poskusa
IPI	fizična interakcija	rezultat poskusa
IMP	mutantni fenotip	rezultat poskusa
IGI	genetska interakcija	rezultat poskusa
IEP	vzorec ekspresije	rezultat poskusa
ISS	podobnost zaporedja ali strukture	računsko izpeljana
ISO	ortologija zaporedij	računsko izpeljana
ISA	poravnava zaporedij	računsko izpeljana
ISM	genomski model	računsko izpeljana
IGC	genomski kontekst	računsko izpeljana
IBA	biološka lastnost prednika	računsko izpeljana
IBD	biološka lastnost potomca	računsko izpeljana
IKR	ključni aminokislinski ostanki	računsko izpeljana
IRD	hiter razhod od prednikovega zaporedja	računsko izpeljana
RCA	recenzirana računska analiza	računsko izpeljana
TAS	sledljiva izjava avtorja	izjava avtorja članka
NAS	ne-sledljiva izjava avtorja	izjava avtorja članka
IC	izjava skrbnika	izjava skrbnika
ND	izjava brez biološkega dokaza	izjava skrbnika
IEA	elektronska anotacija	avtomatska

3.3.2.2 Ročno popravljanje genskih funkcionalnih opisov

Poleg avtomatskih popravkov funkcionalnih opisov genov in ortognih skupin smo v aplikaciji GoMapMan implementirali tudi uporabniški vmesnik, ki omogoča bolj natančne ročne popravke. Dostop do vmesnika za ročno popravljanje dobi vsak avtoriziran in registriran uporabnik, ki je po obrazložitvi osnovnih konceptov ustrezno uveden v tehnične in biološke vidike popravkov.

3.3.2.3 Povezovanje z zunanjimi aplikacijami

Rezultati ročnih popravkov in avtomatske konsolidacije vodijo do izboljšav v povezavah med geni in ortognimi skupinami ter ontologijo MapMan. Za lažjo uporabo funkcionalnih opisov genov iz aplikacije GoMapMan smo implementirali sistem skript, ki podatke v bazi pretvorijo v obliko izvoznih datotek, primernih za uporabo v različnih orodjih za analizo podatkov. Datoteke so dostopne na glavni strani GoMapMan in

posodobljene ob vsaki večji spremembi v bazi (dodatek nove rastline, spremembe ontologije MapMan, konsolidacijske spremembe popravkov genskih anotacij) oz. na tri mesece za samo ročne popravke.

Tipe izvoznih datotek predstavljajo generične izvozne datoteke, izvoz ontologije v formatu odprtih biomedicinskih ontologij (ang. *Open Biomedical Ontologies*, OBO (Smith in sod., 2007), izvoza za orodji MapMan (Thimm in sod., 2004) in Gene Set Enrichment Analysis (Subramanian in sod., 2005) ter skupine večih posameznih datotek za orodji Biomine (Eronen in Toivonen, 2012) in SegMine (Podpečan in sod., 2011; Trajkovski in sod., 2008).

3.4 INTEGRACIJA HETEROGENIH TIPOV PODATKOV

V okviru Oddelka za biotehnologijo in sistemsko biologijo Nacionalnega inštituta za biologijo je bil na osnovi znanstvenih objav in znanja domenskih ekspertov leta 2012 zgrajen model rastlinske obrambe (Miljkovic in sod., 2012). Zgrajeno kvalitativno signalizacijsko omrežje je opisovalo korake biosinteze, prepozname hormonskega signala in prenos signala, korakov, ki vodijo do aktivacije efektorskih molekul glavnih rastlinskih hormonov udeleženih pri obrambnem odgovoru rastline – salicilne kisline (SA), jasmonske kisline (JA) in etilena (ET). Nadgradnja modela se loči na prvi del, kjer smo že obstoječi model združili z obstoječimi podatki v literaturi in javnih bazah podatkov, ter na drugi del, kjer smo dodali še rezultate lastnih transkriptomskih poskusov.

Prvi del sestoji iz treh korakov: (1) nadgradnje že obstoječega modela rastlinske obrambe z dodatnim znanjem iz znanstvenih objav in prenos v drugo orodje (CellDesigner), (2) priprave velikega omrežja iz javno dostopnih podatkov ter (3) izluščenje omrežja rastlinske imunske signalizacije (ang. *plant immune signalling*, PIS) preko kombinacije rezultatov v prvem in drugem koraku. Drugi del vsebuje koraka tvorbe dveh različnih ko-ekspresijskih omrežij (4, 5), ki sta bili združeni s omrežjem PIS.

Tako omrežje PIS iz prvega dela kot omrežje PIS v kombinaciji z ko-ekspresijskimi podatki sta bila analizirana z uporabo različnih pristopov analize omrežij.

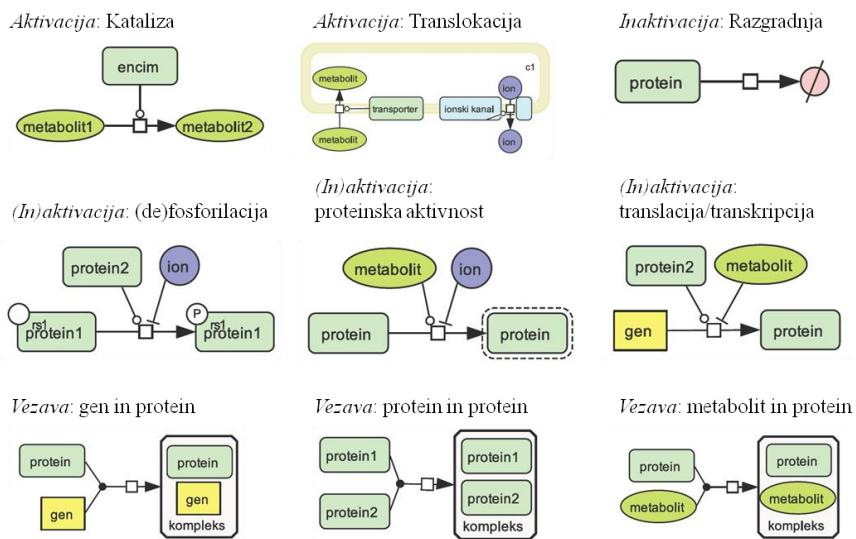
3.4.1 Nadgradnja strukturnega modela rastlinske obrambe

3.4.1.1 Nadgradnja modela imunske signalizacije rastline navadni repnjakovec

Osnovni model kot so ga pripravili Miljkovic in sod. (2012) smo prenesli iz orodja CellIllustrator (Nagasaki in sod., 2010) v orodje CellDesigner (Funahashi in sod., 2003).

Prednosti orodja CellDesigner v primerjavi z orodjem CellIllustrator predstavljata med drugim možnost bolj natančnega opisa biološkega reakcije ter povezava z orodjem za dinamično modeliranje COPASI (Hoops in sod., 2006). Z namenom priprave modela za dinamično modeliranje v prihodnosti (katerega metode so v orodju CellIllustrator slabo podprtne) smo za vsako že obstoječo reakcijo glede na podatke v literaturi določili podskupino reakcije, ki je ontologija reakcij po Miljkovic in sod. (2012) ni vsebovala. S tem smo poleg popravkov nekaterih že obstoječih reakcij ter vključitve novih reakcij iz literature nadgradili tudi že obstoječo ontologijo reakcij modela (Slika 14).

Iz druge strani smo model iz rastline navadnega repnjakovca prevedli na krompir, za kar smo vzporedno nadgradili taksonomijo komponent modela (Slika 15). Ob tem smo vsaki reakciji dodali nivo opisa procesa, torej v kateri biološki poti komponenta (protein, proteinski kompleks ali metabolit) sodeluje. Za vsako družino genov iz Miljkovic in sod. (2012) smo v orodju PLAZA (Proost in sod., 2014) poiskali ortologne skupine, po potrebi družino razširili z dodatnimi ortolognimi geni ter tako izboljšali pokritje genov v družini. Ker nam je povezava z ortolognimi skupini orodja PLAZA omogočila prevod na gene krompirja in s tem povezano z eksperimentalnimi podatki, smo zaradi opazne razlike med vzorci izražanja genov določili še pod-družine.

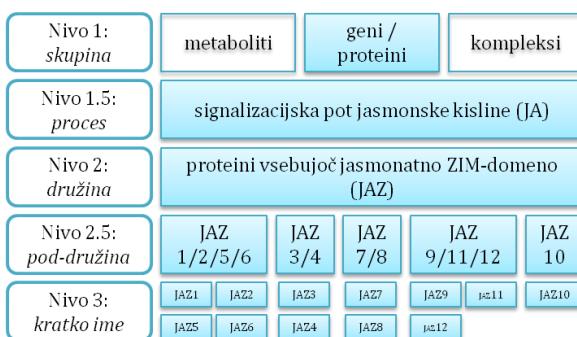


Slika 14: Ontologija reakcij modela rastlinske imunske signalizacije (PIS).

Definirane so tri glavne skupine glede na funkcionalnost reakcije: aktivacija, inaktivacija in vezava. Vse skupaj je definiranih 12 reakcijskih tipov, kjer so tri od teh lahko tako aktivacijski kot inaktivacijski, kar je odvisno od biološkega konteksta (aktivacija proteina, translacija/transkripcija, (de)fosforilacija). Preostaneta dve aktivacijski reakciji (kataliza in translokacija), ena inaktivacijska (razgradnja) ter tri vezavne reakcije, ki so odvisne od tipov komponent, ki se med seboj vežejo.

Figure 14: Reaction ontology of the plant immune signalling (PIS) model.

Three major groups according to reaction functionality are defined: activation, inactivation and binding. In total 12 possible reaction types are defined, where for three groups activation and inactivation depends on the biological context (protein activation, translation/transcription, (de)phosphorylation). Two activation reactions remain (catalysis, translocation), one inactivation (degradation) and three binding reactions dependent on the component types which bind.



Slika 15: Taksonomija komponent modela rastlinske imunske signalizacije (PIS).

Originalna shema Miljkovic in sod. (2012) je ločevala komponente v skupino, družino in gene, kjer smo naknadno dodali še opis procesa in pod-družine. Skupina opisuje tip komponente, proces opisuje poti v katerih je komponenta prisotna, družina predstavlja nadpomenko za vse gene, ki ji pripadajo (geni), medtem ko pod-družina ločuje gene neke družine glede na njihovo ortologno pripadnost.

Figure 15: Component taxonomy of the plant immune signalling (PIS) model.

Originally Miljkovic in sod. (2012) organised the components on the level of their group, family and genes, while we additionally added the process and clade levels. The group describes the type of component, process describes the pathways in which the component is involved, family is represented with the large umbrella term for all the genes belonging to it, and clades represent an orthologue dependent gene separation.

3.4.1.2 Integracija javno dostopnega znanja

Iz javnih virov podatkov smo združili eksperimentalno znanje o regulaciji aktivnosti genov (transkripcijski faktorji, male RNA) ter o proteinskih interakcijah. Dodatno smo imeli na voljo še znanstvene objave, ki še niso bile vključene v javne baze. Osredotočili smo se na članke vezane točno na komponente rastlinske imunske signalizacije. Podatkovni gradniki (objave, baze podatkov) za veliko omrežje znanja navadnega repnjakovca so naštetи in podrobnejše opisani v Preglednici 5.

Preglednica 5: Pregled podatkov iz literature v velikem omrežju znanja navadnega repnjakovca.
PPI pomeni proteinske interakcije, TF transkripcijsko regulacijo in miRNA regulacijo preko malih RNA. Vsak od podatkovnih skupov vsebuje tudi opis kvalitete podatkov (izvor iz tarčnih poskusov, visoko-zmogljivih študij ter *in silico* napovedanih podatkov).

Table 5: Overview of literature sources combined into the large knowledge network of *A. thaliana*.
PPI denotes protein-protein interactions, TF transcriptional regulation and miRNA micro-RNA regulation. Each of the sets also contains a descriptor of data quality (experimental data of targeted experiments, high-throughput studies data and *in silico* putative data).

Tip podatka	Opis
PPI <i>tarčni</i>	Iz baze AtPIN smo vključili le visoko kvalitetne ročno pregledane podatke iz literature. Napovedi proteinskih interakcij na osnovi ortologije nismo vključili (Brandão in sod., 2009).
PPI <i>tarčni</i>	Iz baze STRING smo uporabili le povezave pokrite v literaturi ter z vrednostjo povezave kot so jo ocenili v bazi večjo od 0.900 (Szklarczyk in sod., 2014).
PPI <i>visoko zmogljivi</i>	Velik poskus v dvohibridnem sistemu kvasovke (ang. <i>yeast two hybrid system</i> , Y2H) vključujoč preko 10.000 genov navadnega repnjakovca (Arabidopsis Interactome Mapping Consortium, 2011).
PPI <i>visoko zmogljivi</i>	Velik poskus za določanje vezave med membransko vezanimi proteini. Uporaba alternativne Y2H metode z ločevanjem ubikvitina (Jones in sod., 2014).
PPI <i>visoko zmogljivi</i>	Manjši Y2H poskus, ki je proučeval povezave med proteini navadnega repnjakovca in virusa TuMV, analoga našemu virusu PVY (Elena in sod., 2011).
PPI <i>visoko zmogljivi</i>	Manjši Y2H poskus povezav z virulenčnimi faktorji rastline (Mukhtar in sod., 2011).
PPI <i>napoved</i>	Interakcije napovedane na podlagi lokalizacije, znanja o bioloških procesih ter ko-ekspresijskih podatkov (De Bodt in sod., 2009).
TF <i>tarčni</i>	Baza AtRegNet vsebuje povezave, potrjene v literaturi, ter tudi samo napovedane (Palaniswamy in sod., 2006). Pod različno oznako kvalitete smo vključili obe množici.
TF <i>HT</i>	ChIP-Seq poskus povezan z časovnim odgovorom navadnega repnjakovca na tretiranje z etilenom (Chang in sod., 2013).
TF <i>napoved</i>	S ko-ekspresijo napovedane povezave med TF in tarčami (Srivastava in sod., 2010).
miRNA <i>tarčni</i>	Baza miRTarBase vsebuje ročno pregledane povezave iz literature, ki so bile validirane tudi eksperimentalno (Hsu in sod., 2014).
miRNA <i>visoko zmogljivi</i>	Baza PMRD združuje podatke iz raznih baz podatkov in jih združuje z eksperimentalnimi podatki (Zhang in sod., 2010).

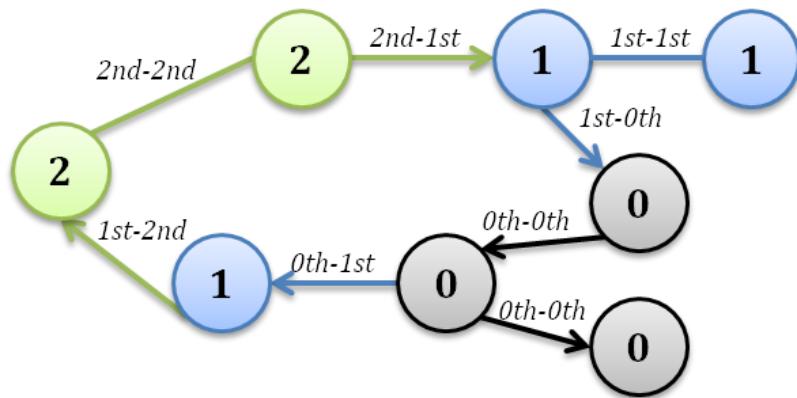
Uporabljeni viri podatkov so se glede na kakovost in zanesljivost vsebovane informacije razdelili na tri dele s padajočo stopnjo zanesljivosti: (1) podatki tarčnih eksperimentov z najvišjo stopnjo zanesljivosti, (2) podatki iz visoko zmogljivih študij ter *in silico* napovedi različnih algoritmov, kjer smo (3) v okviru znanstvene objave validiranim *in silico* napovedim pripisali rahlo višjo stopnjo zaupanja kot (4) ne-validiranim *in silico* napovedim z najnižjo pripisano stopnjo zanesljivosti.

V veliko omrežje smo dodali tudi povezave med geni reakcij iz modela, razširjenih iz družin na gene. Ker je veliko omrežje zgrajeno na podlagi povezav med geni in ne vključuje tudi metabolitov, smo pri prepisu reakcij v modelu le-te dodatno ločili. Od 642 povezav v modelu na nivoju genov smo 510 povezavam pripisali najvišjo stopnjo zaupanja (1). Preostalih 132 reakcij predstavlja reakcije sinteze s poenostavljivo z globalno najnižjo stopnjo zaupanja (5): namesto da bi v model vključili metabolite, smo med seboj povezali encim, ki opravi samo sintezo metabolita, z naslednjim encimom v seriji pretvorb sintezne poti.

3.4.1.3 Priprava omrežja imunske signalizacije krompirja

Veliko omrežje znanja navadnega repnjakovca je bilo potrebno prevesti na krompir, pri čemer smo se poslužili znanja vsebovanega v ortolognih tipih: ITAG_orthoMCL, ITAG_RSD, PGSC_orthoMCL, šibkih komponent ter družin in pod-družin dvokaličnic spletni aplikacije PLAZA (Proost in sod., 2014). Na podlagi ortolognih skupin s katerimi se povezujejo geni navadnega repnjakovca, smo omrežje nato prevedli na gene krompirja. Osnovni geni prevedenega omrežja krompirja so lahko predstavljeni v obliki identifikatorjev ortolognih skupin (abstrakcija omrežja), lahko pa omrežje razširimo na vse gene krompirja vsebovane v tisti ortologni skupini (razširitev omrežja).

Na podlagi seznama genov, ki so prisotni v modelu rastlinske imunske signalizacije, smo iz velikega omrežja znanja navadnega repnjakovca pripravili omrežje rastlinske imunske signalizacije (PIS). Od vseh 260 genov nadgrajenega modela PIS (Poglavlje 3.4.1.1) smo obdržali le gene vključene v poti etilena, jasmonske kisline ali salicilne kisline (214 genov navadnega repnjakovca). Vsakemu od teh smo v velikem omrežju znanja pripisali oznako ničte generacije. Nato smo v velikem omrežju znanju najprej določili vse prve sosedne genov ničte generacije ter s procesom nadaljevali dokler je bilo možno (Slika 16).



Slika 16: Priprava podomrežij rastlinske imunske signalizacije preko določanja bližnjih sosedov.
Genom, ki so tako v našem modelu kot v velikem omrežju, pripisemo rank ničte generacije in povezave med njimi so opisane kot ničte-ničte povezave. Vsem genom, ki se povezujejo z geni ničte generacije, pripisemo oznako prve generacije in povezavi oznako ali ničte-prve oz. prve-ničte (usmerjenost povezav) ali prve-prve. Proses razvrščanja genov se ponavlja, dokler niso pokriti vsi geni, ki so povezani v omrežju (nepovezani deli omrežja so izpuščeni).

Figure 16: Extraction of plant immune signalling subnetworks via gene ranking.

All the genes which are both in our model and the large network get the rank of 0th generation and connections between them are termed 0th-0th connections. All genes connecting with these genes are termed genes of the 1st generation and their connections either 0th-1st or 1st-0th (connection directionality) or 1st-1st. The process of gene ranking is repeated until all the connected genes are covered (unconnected parts of the network are ignored).

Iz velikega omrežja navadnega repnjakovca z razvrščenimi gene (glede na osnovni seznam genov pomembnih za obrambni odgovor rastline) lahko pripravimo več podomrežij različne velikosti. Podomrežja so odvisna od nivoja povezav pri katerem omrežje odrežemo, npr. omrežje AtPIS-g2 vsebuje vse povezave od genov z oznako ničte do druge generacije. Dobljena podomrežja se z uporabo PLAZA ortolognih skupin poddržin dvokaličnic nato prevede na podomrežja StPIS.

3.4.2 Priprava ko-ekspresijskih omrežij

3.4.2.1 Priprava ko-ekspresijskih omrežij v programu BioLayout

Za pripravo eksperimentalnih podatkov, ki smo jih uporabili v programu BioLayout smo uporabili StPIS-g1 podomrežje (omrežje, ki izhaja iz AtPIS-g1, torej vsebuje povezave genov ničte generacije in njih prvih sosedov). Z uporabo prevajalne tabele iz Poglavlja 3.2.1 smo te gene krompirja prevedli na mikromrežne sonde POCI. V podomrežju smo obdržali le tiste povezave, pri katerih sta oba interaktorja v povezavi imela pripisano vsaj eno mikromrežno sondu POCI.

Ker je po pripravi pan-genoma krompirja vsakemu genu v paralogni skupini lahko pripadalo več mikromrežnih sond, je bilo potrebno za vsak gen določiti reprezentativno sondi POCI. Za le-te smo se odločali na podlagi časovne serije odzivov rastline na virus v 18ih različnih pogojih iz Poglavlja 3.1.1.2. Odločitveno drevo je bilo sestavljeno iz treh pogojev:

- (1) Izbrana je bila sonda, ki je bila v največjem številu primerjav diferencialno izražena pri FDR popravljeni p-vrednosti manjši od 0,05.
- (2) Če je bilo sond z enakim številom pogojev več, smo izbrali tisto, ki je imela največje povprečje logaritma razmerja spremembe med pogoji (ang. *log fold change*, logFC), kjer je bila diferencialno izražena.
- (3) Če je bilo sond določenih v prejšnji točki več, smo izbrali sondi z največjo povprečno ekspresijo sond, ki so prešle kontrole kvalitete.

Seznam teh sond in njihove antilogaritmizirane normalizirane vrednosti izražanja iz poskusov opisanih v Poglavlju 3.1, povprečene po bioloških ponovitvah, smo uvozili v program BioLayout (Theocharidis in sod., 2009) in izračunali ko-ekspresijo teh genov z uporabo treh mejnih vrednosti Pearsonovega korelacijskega koeficienta (0,85/0,90/0,95).

3.4.2.2 Priprava tarčnih ko-ekspresijskih omrežij

Ko-ekspresijska omrežja smo pripravili tudi bolj tarčno, z uporabo razvite ko-ekspresijske metode (Itkin in sod., 2013), kateri se poda seznam vab – genov, za katere vemo da so udeleženi pri neki celični poti. Nato se za vsak gen iz seznama vab izračuna vrednost korelacijskega koeficienta r z mikromrežnimi sondami iz poskusov opisanih v Poglavlju 3.1.

V našem primeru so bile vabe specifične za etilenski del modela. Pripravili smo dva seznama vab (genov): oba seznama sta pokrivala celo signalizacijsko pot etilena, razlikovala pa sta se v pokritju sintezne poti, kjer je krajiši seznam (17 genov krompirja) vseboval le zadnji korak sinteze, daljši seznam (34 genov krompirja) pa celotno pot sinteze.

Oba seznama smo kot vabe uporabili pri pripravi tarčne ko-ekspresijske mreže na podatkih vseh mikromrežnih sond poskusa (~37 tisoč). Pripravili smo tri različne datoteke, ki so vsebovale različne skupine bioloških vzorcev: prva je vsebovala vseh 67 bioloških vzorcev; druga je vsebovala le vrednosti Désirée/NahG-Désirée inokuliranih listov; tretja pa je vsebovala le vrednosti Rywal/NahG-Rywal inokuliranih listov. Vse datoteke so vsebovale logaritmizirane (log2) povprečene vrednosti, v omrežju pa smo obdržali le povezave s korelacijsko vrednostjo r večjo od 0,9.

3.4.3 Povezovanje omrežja imunske signalizacije s podatki eksperimentov

V zadnjem koraku smo pripravljena omrežja imunske signalizacije pri krompirju, t.j. StPIS-g1 podomrežje (iz poglavja 3.4.1.3), ko-ekspresijska omrežja programa BioLayout (iz poglavja 3.4.2.1) ter tarčna ko-ekspresijska omrežja (iz poglavja 3.4.2.2) združili v eno samo. Od rezultatov programa BioLayout smo uporabili vsa izmed treh možnih omrežij izhajajoč iz različnih praznih vrednosti korelacijskega koeficienta (0,85/0,90/0,95). Od treh možnih tarčnih ko-ekspresijskih omrežij smo uporabili le omrežje pridobljeno s kratkim seznamom vab etilenske signalizacije (17 genov krompirja) na podatkih Désirée/NahG-Désirée inokuliranih listov. Končna omrežja smo poimenovali omrežja BTP (BTP-0,85, BTP-0,90 in BTP-0,95), saj predstavljajo integrirano ko-ekspresijsko omrežje iz programa BioLayout, metode tarčne ko-ekspresije ter omrežje StPIS-g1.

Za primerjavo uspešnosti napovedovanja novih povezav vsake od podskupin (veliko omrežje znanja navadnega repnjakovca in obe ko-ekspresijski omrežji) smo kot merilo vzeli ročno pregledane povezave vsebovane v modelu PIS. V predpripravah smo vse štiri enote preko PLAZA-v3 ortolognih skupin prevedli na ortologne skupine, saj nam je to omogočilo najlažjo primerjavo, brez dodatnih zapletov velikosti genskih družin v obeh rastlinah. Dodatno smo obdržali le tiste povezave, kjer sta se obe komponenti povezave prevedli na ortologno skupino, ki je vsebovala tako gen navadnega repnjakovca kot gen krompirja.

3.4.4 Analiza pripravljenih omrežij

3.4.4.1 Nenadzorovana analiza omrežij

Hierarhično združevanje je ena izmed metod iskanja strukture skupin v omrežju, kjer se na podlagi neke utežnostne funkcije komponente in povezave razdeli v hierarhijo skupin. V delu smo se poslužili algoritma razvrščanja po Markovu (ang. *Markov Cluster Algorithm*, MCL; Van Dongen, 2000) brez spremnjanja nastavitev. Algoritem smo uporabili na ko-ekspresijskih mrežah programa BioLayout za vse tri pražne vrednosti korelacijskega koeficienta r (0,85/0,90/0,95). Celotno BTP-0,90 omrežje smo analizirali tudi s algoritmoma MCODE (Bader in Hogue, 2003) in CentiScaPe (Scardoni in sod., 2009) v programu Cytoscape (Shannon in sod., 2003), ponovno brez spremnjanja nastavitev.

3.4.4.2 Izboljšava modela rastlinske obrambe

Za iskanje novih povezav v modelu rastlinske obrambe smo uporabili AtPIS-g0 (najmanjše) omrežje izluščeno iz velikega omrežja znanja navadnega repnjakovca na podlagi vseh genov modela rastlinske obrambe PIS (geni ničte generacije). Izluščene povezave smo nato primerjali z že obstoječimi v modelu PIS in iskali nove povezave. Dodatno smo se osredotočili na pregled vseh 11 virusnih proteinov in njihovo interakcijo s komponentami našega modela.

Iskanja novih povezav v imunski signalizaciji smo se lotili tudi v omrežju StPIS-g1, ki je vsebovalo tudi prve sosedne (in njih povezave) genov modela rastlinske obrambe. Uporabili smo pristop iskanja najkrajših poti ter sprehodov med dvema točkama v omrežju (sprehodom smo dovolili maksimalno dolžino 4). Za ta namen smo pripravili seznam ključnih genov imunske signalizacije (Preglednica 6), ki je za vsako signalizacijsko pot rastlinskega hormona vseboval (vsaj en) receptor in prenosnik (ang. *transmitter*). Iskanje kratkih poti in sprehodov smo opravili v programu Pajek (Batagelj in Mrvar, 1998).

Preglednica 6: Seznam izbranih genov za iskanje najkrajših poti in sprehodov.

Tabela predstavlja seznam receptorjev in prenosnikov za vsak rastlinski hormon, uporabljeni v analizi, najprej formalno kratko ime gena, kateremu sledi genski identifikator TAIR navadnega repnjakovca.

Table 6: The list of selected genes for shortest path and walks searching.

The table lists the receptor and transmitter genes for each plant hormone used in the analysis, first their officially defined short name, followed by TAIR gene identifier.

Rastlinski hormon	Receptor	Prenosnik
Etilen	ETR1 (AT1G66340)	EIN3 (AT3G20770) ERF1A (AT4G17500)
Salicilna kislina	NPR1 (AT1G64280)	TGA2 (AT5G06950)
Jasmonska kislina	COI1 (AT2G39940)	MYC2 (AT1G32640)
Abscizinska kislina	PYR1 (AT4G17870)	SNRK (AT5G66880)
Avksini	TIR1 (AT3G62980)	ARF2 (AT5G62000)
Citokinini	AHK2 (AT5G35750) AHK3 (AT1G27320) AHK4 (AT2G01830)	/
Brazinosteroidi	BRI1 (AT4G39400)	BES1 (AT1G19350)

Od izbranih 16-ih genov smo pripravili 120 kombinacij brez ponavljanja v obe smeri (torej skupno 240), saj smo kratke poti in sprehode iskali v usmerjenem omrežju in lahko pride do razlik v najdenih poti oz. sprehodih.

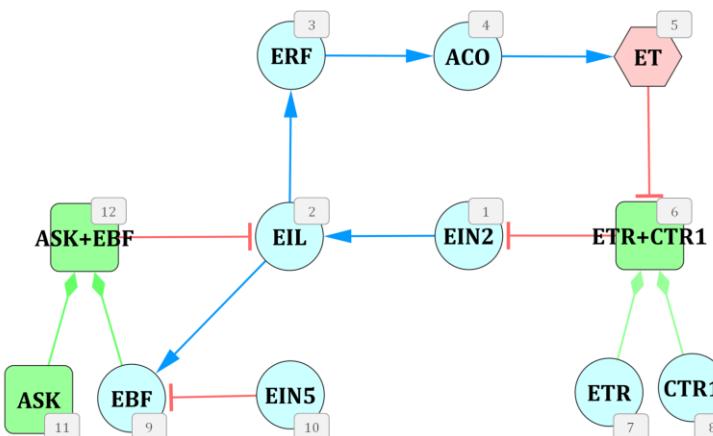
3.4.4.3 Analiza biološke hipoteze

Za podrobnejšo analizo biološke hipoteze smo se osredotočili na gen *npr1* (s kodo AT1G64280 v navadnem repnjakovcu). Glede na informacijo v petih ortolognih tipih (PGSC_oMCL, ITAG_oMCL, ITAG_RSD, PLAZA-v3 družine/poddružine dvokaličnic) smo gen lahko prevedli iz navadnega repnjakovca na krompir (Sotub07g011600.1.1). Za analizo promotorske regije, smo zaporedje gena krompirja poiskali v genomskem brskalniku SpudDB konzorcija PGSC (The Potato Genome Sequencing Consortium, 2011) ter izolirali promotorsko regijo (2000 nukleotidov pred začetkom odprtrega brialnega okvirja). Pridobljenemu zaporedju smo poiskali cis-regulatorne regije z uporabo aplikacij TRANSFAC (Matys in sod., 2006) in PlantCare (Lescot in sod., 2002).

3.5 RAZVOJ DINAMIČNEGA MODELA RASTLINSKE OBRAMBE

3.5.1 Predpriprava

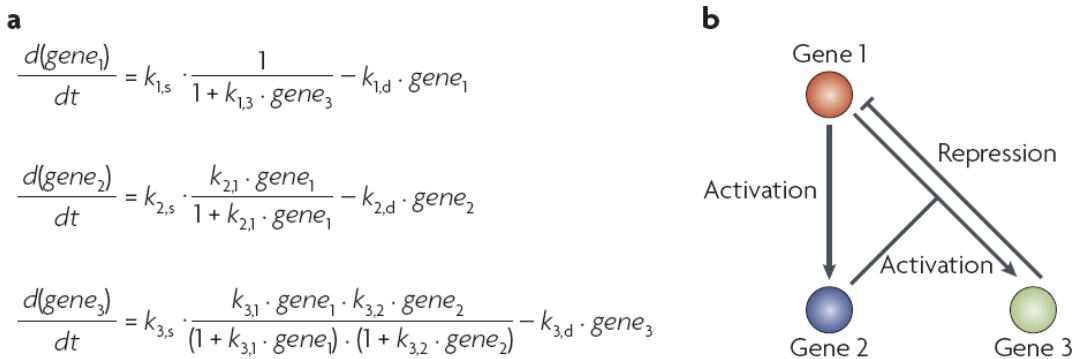
Bayesova linearna metodologija omogoča zmanjševanje razpona možnih kombinacij parametrov preko ocene verjetnosti oz. neverjetnosti (ang. *implausibility*; Vernon in sod., 2010). Za metodologijo potrebujemo pripravljen model, kjer ima vsaka komponenta določeno obliko navadne diferencialne enačbe ter pripisano eksperimentalno vrednost. Za dinamično modeliranje smo pripravili majhen podmodel etilenske signalizacije, ki je vseboval 12 komponent, od tega 8 proteinov, en metabolit in tri proteinske komplekse (Slika 17). Komponente je med seboj povezovalo 11 reakcij, od tega 5 z aktivacijskim delovanjem, 4 z inhibicijskim in 2 reakciji nastanka kompleksa.



Slika 17: Etilenski podmodel uporabljen v pristopu Bayesove linearne metodologije.
Podmodel sestavljajo različni tipi komponent – proteini (svetlo modre točke), metabolit (rdeča) in kompleksi (zeleni), ki so povezani s tremi tipi reakcij – aktivacijskimi (modre povezave), inhibicijskimi (rdeče) ali reakcije tvorbe kompleksa (zelene).

Figure 17: The ethylene submodel used with the Bayesian linear methodology approach.
The submodel is composed of protein (light blue nodes), metabolite (red) and complex (green) components, which are connected with three reaction types, activation (blue connection), inhibition (red) or complex formation (green).

Kot eksperimentalne podatke smo uporabili časovno serijo (iz Poglavlja 3.1) dolžine 5-ih točk kontrolnih listov kultivarja krompirja Désirée. Izbrani sistem dvanajstih komponent smo predstavili s formalizmom navadnih diferencialnih enačb. Kot je predstavljeno v Karlebach in Shamir (2008) ter prikazano na Sliki 18 smo vsaki komponenti dodelili parameter hitrosti sinteze ter hitrosti razgradnje, ter dodatno v odvisnosti od tipa reakcije, hitrost aktivacije, inhibicije oz. vezave (Slika 19). Sintezo obeh kompleksov smo predstavili v obliki enačb masnega prenosa.



Slika 18: Primer priprave navadnih diferencialnih enačb za preprost sistem, ki vključuje aktivacijo in represijo genov (Karlebach in Shamir, 2008: 2)

(a) Model navadnih diferencialnih enačb z (b) grafično predstavljivjo regulatornih odnosov. V prvi enačbi je gen 1 inhibiran z genom 3, v drugi enačbi je gen 2 aktiviran z genom 1, v zadnji enačbi pa sta zajeti tako aktivacija kot inhibicija (gen 3 je aktiviran z genoma 1 in 3; gen 1 inhibiran z genom 3). Konstante reakcij so označene s 'k', kjer $k_{l,s}$ pomeni sintezo gena 1, $k_{l,d}$ razgradnjo gena 1 ter $k_{l,2}$ vpliv (aktivacijo oz. inhibicijo gena 1 z genom 2).

Figure 18: An example of ODE equation set up for a simple system including gene activation and repression (Karlebach and Shamir, 2008: 2)

(a) A model of ordinary differential equations with the (b) graphical depiction of regulatory relations. The first equation represents the repression of gene 1 by gene 3, the second activation of gene 2 by gene 1 and the last a combination of both activation and inhibition (gene 3 is activated by genes 1 and 3; gene 1 is repressed by gene 3). Reaction rate constants are denoted by 'k', where $k_{l,s}$ denotes gene 1 synthesis, $k_{l,d}$ gene 1 degradation and $k_{l,2}$ influence (activation or inhibition) of gene 1 by gene 2.

```

EIN2 df1 <- x_s1*(1/(1+(x_i1.6*f[6]))) - x_d1*f[1]
EIL df2 <- x_s2*((x_a2.1*f[1])/(1+(x_a2.1*f[1])))*(1/(1+x_i2.12*f[12])) - x_d2*f[2]
ERF df3 <- x_s3*((x_a3.2*f[2])/(1+(x_a3.2*f[2]))) - x_d3*f[3]
ACO df4 <- x_s4*((x_a4.3*f[3])/(1+(x_a4.3*f[3]))) - x_d4*f[4]
Et df5 <- x_s5*((x_a5.4)*f[4]/(1+(x_a5.4*f[4]))) - x_d5*f[5]
ETR-CTR1 df6 <- (x_s6*f[7]*f[8])*(1/(1+(x_i6.5*f[5]))) - x_d6*f[6]
ETR df7 <- x_s7 - x_s6*f[7]*f[8] - x_d7*f[7]
CTR1 df8 <- x_s8 - x_s6*f[7]*f[8] - x_d8*f[8]
df9 <- x_s9*(1/(1+x_i9.10*f[10]))*(x_a9.2*f[2])/(1+x_a9.2*f[2]) - 
    (x_s12*f[9]*f[11]) - x_d9*f[9]
EBF (x_s12*f[9]*f[11]) - x_d9*f[9]
EIN5 df10 = x_s10 - x_d10*f[10]
ASK df11 = x_s11 - x_s12*f[9]*f[11] - x_d11*f[11]
ASK-EBF df12 = x_s12*f[9]*f[11] - x_d12*f[12]

```

Slika 19: Sistem definiranih navadnih diferencialnih enačb za etilenski podmodel.

Vsaki komponenti je bila dodeljena diferencialna enačba, sestavljena iz parametrov sinteze in degradacije in z dodatnimi parametri vpliva drugih genov v odvisnosti od tipa reakcije (aktivacija, inhibicija). S tem smo definirali 33 parametrov, katerih vrednosti smo ocenjevali v kasnejših korakih metodologije: 12 sinteznih, 12 degradacijskih, 5 aktivacijskih in 4 inhibicijskih. Reakcije vezave kompleksov niso imele definiranih dodatnih parametrov, saj so bile predstavljene kot enačba masnega prenosa med dvema komponentama.

Figure 19: The defined system of ordinary differential equations for the ethylene submodel.

Each component had an ordinary differential equation assigned to it, combined of synthesis and degradation parameters, with additional parameters of other gene influence depending on reaction types (activation, inhibition). This resulted in a total of 33 parameters for the methodology to determine: 12 synthesis, 12 degradation, 5 activation and 4 inhibition parameters. The binding reactions were excluded, as they were represented as a mass reaction equation between two components.

3.5.2 Koraki v valu Bayesove linearne metodologije parametrizacije

En val Bayesove linearne metodologije za določanje parametrov v sistemu diferencialnih enačb tvorijo trije koraki: (1) priprava vzorčnega prostora parametrov z numerično integracijo diferencialnih enačb na manjši podskupini, (2) izgradnje linearnih modelov ter (3) izračun mere neverjetnosti, na podlagi katere nato izločimo predele vzorčnega prostora parametrov, ki niso možne rešitve sistema enačb. S tem se zaključi val in ohranjene kombinacije parametrov lahko vodimo v naslednji val.

3.5.2.1 Vzorčenje z latinsko hiperkocko in numerična integracija

V prvem koraku z uporabo statistične metode vzorčenja z latinsko hiperkocko (ang. *latin hypercube sampling*, LHS) generiramo verodostojne zbirke vrednosti parametrov iz večdimenzionalne distribucije. Koncept LHS vzorči spremenljivke (v našem primeru parametre diferencialnih enačb) na enakomeren način glede na definirane razpone možnih vrednosti, nato pa naključno združi spremenljivke v različne kombinacije.

Za vsako komponento smo najprej določili njeno začetno vrednost, kar smo povzeli kar po intenziteti izražanja v dnevu 0 (na prvi dan, pred tretmajem rastlin z vodo oz. virusom). Dvema komponentama proteinskega kompleksa in metabolitu (5, 6 in 12 na Sliki 17) nismo mogli določiti reprezentativne mikromrežne sonde, zato smo tem komponentam določili arbitrarno vrednost 8000 (Preglednica 7).

Preglednica 7: Začetne vrednosti za 12 komponent etilenskega podmodela.

Devet komponent je imelo pripisano svojo lastno mikromrežno sondu, preostalim trem (dva kompleksa in etilen) pa smo določili začetno vrednost 8000.

Table 7: Starting values for the 12 components of the ethylene submodel.

Nine components were assigned with its own microarray probe, the remaining three (two complexes and ethylene) had the starting value of 8000 assigned.

Komponenta	Začetna vrednost	Komponenta	Začetna vrednost
EIN2	2041	ETR1	896
EIL	5765	CTR1	936
ERF	183	EBF	2257
ACO	25978	EIN5	102
Et	8000	ASK	36891
ETR-CTR1	8000	ASK-EBF	8000

Začetne vrednosti vseh 33ih parametrov smo nastavili na vrednost 1 in jih nato povečali oz. zmanjšali za $5000\times$, s čimer smo vsakemu parametru razširili možen prostor. Iz tako definiranih spodnjih in zgornjih meja vrednosti parametrov je bila nato zgrajena majhna latinska hiperkocka (ang. *latin hypercube*, LH) za 400 kombinacij vrednosti parametrov. Te kombinacije vrednosti parametrov majhne LH smo uporabili za reševanje diferencialnih enačb (numerične integracije) s funkcijo *lsoda* iz paketa deSolve (Soetaert in sod., 2010) v programskem okolju R (R Project for Statistical Computing; R Development Core Team, 2011; version 2.13.2).

3.5.2.2 Priprava linearnih modelov in emulatorja

V naslednjem koraku smo primerjali rezultate numerične integracije glede na zastavljenе začetne parametre iz majhne latinske hiperkocke (LH) z eksperimentalnimi vrednosti mikromrež v petih časovnih točkah. Za vsako komponento pri vsaki časovni točki smo zgradili linearni regresijski model glede na 400 simulacij naenkrat (torej 45 linearnih modelov za 9 merjenih komponent pri petih časovnih točkah). Vrednosti v prvi izmerjeni časovni točki ob začetku eksperimenta smo uporabili za nastavitev začetnih vrednosti naših komponent.

Preko analize zgrajenih modelov (njihovih standardnih napak in prilagojenega R^2) smo identificirali najbolj vplivne parametre našega sistema. Linearne modele le-teh najbolj vplivnih parametrov smo nato uporabili za pripravo emulatorja, s katerim zaobidemo problem integracije reševanja velikega števila diferencialnih enačb. Nato smo pripravili veliko LH v velikosti 40,000 kombinacij za 33 parametrov in te kombinacije smo uporabili za nove simulacije, sedaj z emulatorjem in ne numerično integracijo.

3.5.2.3 Izračun neverjetnosti in izbor ne-neverjetnih kombinacij parametrov

Za vsak linearji model komponente pri neki časovni točki smo nato izračunali vrednost neverjetnosti za vsako izmed 40,000 kombinacij velike LH. Visoka vrednost izračunane neverjetnosti pomeni, da ta kombinacija parametrov (kljub vsem upoštevanim nezanesljivostim problema) ne predstavlja dobrega približka med vrednostmi modela in opazovanimi eksperimentalnimi vrednostmi. Vrednost neverjetnosti se izračuna na podlagi uporabe t.i. emulatorja, statističnega konstrukta linearnih modelov, po naslednji enačbi,

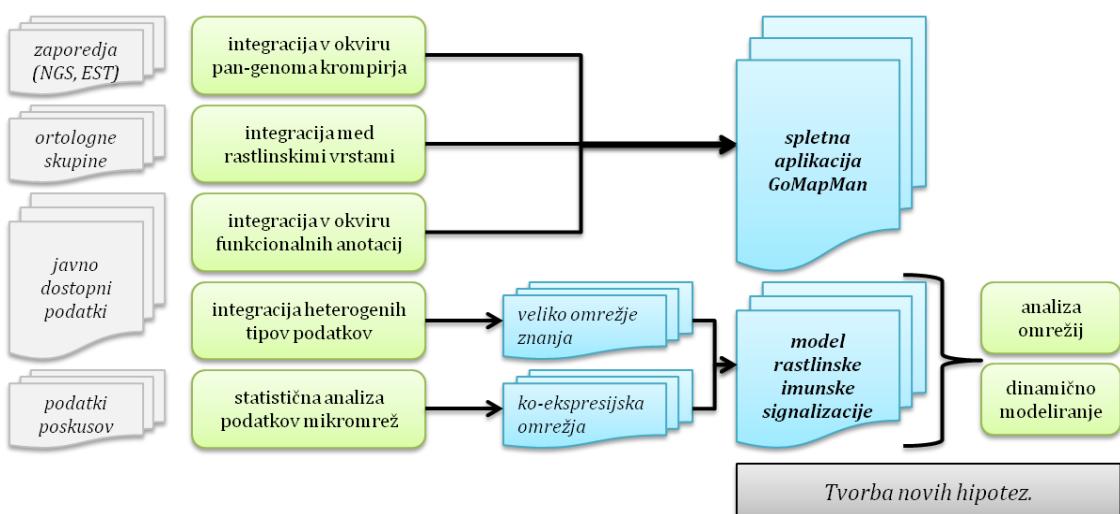
$$I_{(i)}^2(x) = \frac{|E_{D_i}(f_i(x)) - z_i|^2}{\left(\text{Var}_{D_i}(f_i(x))\right) + \text{Var}(\epsilon_{md,i}) + \text{Var}(\epsilon_{obs,i})} \quad \dots (2)$$

kjer je vrednost neverjetnosti ocenjevanega parametra $I_{(i)}^2(x)$ odvisna od razlike med izračunanimi vrednostmi emulatorja $E_{D_i}(f_i(x))$, t.j. rezultatov linearnih modelov, ter opazovanih vrednosti izražanja v časovnih točkah z_i . Napako emulatorja predstavlja kombinacija napak linearnih modelov $\text{Var}_{D_i}(f_i(x))$, napak izhajajočih iz nepopolnosti postavljenega modela $\text{Var}(\epsilon_{md,i})$ ter napak samih opazovanih vrednosti izražanja $\text{Var}(\epsilon_{obs,i})$.

Z zmanjšanim naborom kombinacij parametrov na podlagi vrednosti neverjetnosti lahko nato vstopimo v naslednji zaporedni val metodologije, ali pa zaključimo s procesom ocenjevanja vrednosti parametrov, če so dobljene neverjetnosti nizke.

4 REZULTATI

V prvem delu doktorske naloge smo obdelali lastne podatke transkriptomskega poskusa, združili javno znana zaporedja v pan-genom krompirja ter pripravili in združili znanje razpršeno v različnih virih. Na podlagi tega smo pripravili ko-eksprejsiska omrežja in veliko omrežje znanja v rastlini, in vzporedno razvili spletno aplikacijo GoMapMan. Ti rezultati prvega dela so nato v drugem delu omogočali analizo omrežij, nadgradnjo modela imunske signalizacije in tvorbo novih hipotez, ter vodili v začetek dinamičnega modeliranja (Slika 20).



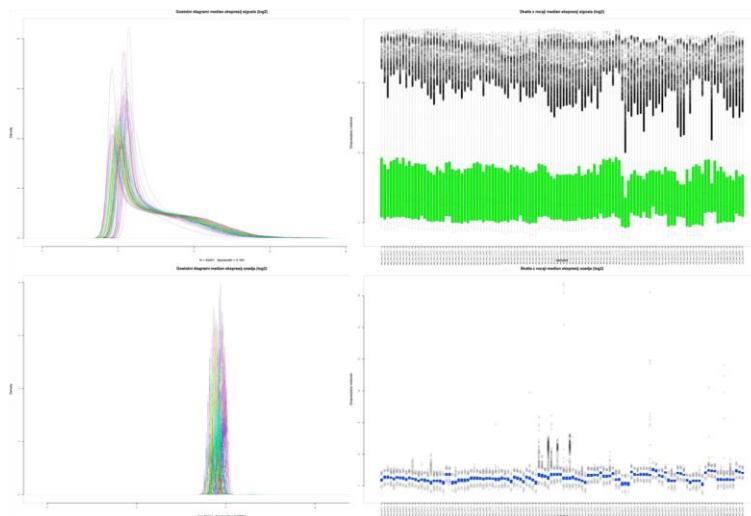
Slika 20: Shema dela predstavljačjoč naloge opravljene v doktorskem delu.

Figure 20: Scheme representing the undertaken tasks of the doctoral work.

4.1 STATISTIČNA ANALIZA TRANSKRIPTOMSKIH PODATKOV

Preden smo se lotili izločanja točk na mikromrežah slabše kakovosti v odvisnosti od oznak kakovosti proizvajalca Agilent smo za vrednosti intenzitet signalov in ozadij vseh mikromrež pregledali izrise škatel z okvirji in gostotne dijagrame (Slika 21). Na podlagi teh izrisov smo tako ocenili, da korekcija ozadja ni potrebna.

Za kasnejše analize med vzorci smo želeli ohraniti le kvantitativne in kvalitetne sonde, zato smo z uporabo petih oznak kontrole kvalitete pripravili matriko uteži, kjer je vsaka sonda nekega vzorca prejela utež 0, če je katera od teh oznak kazala na slabo kvaliteto, drugače pa je bila utežena z vrednostjo 1. Če je bila utež za izbrano mikromrežno sondu enaka 0 pri več kot 90 % vzorcev, je bila slednja izločena iz nadaljnih analiz. Z drugimi besedami, dovolili smo slabo kvaliteto sonde do pri 12 vzorcih, kar je vodilo do izločitve 4,063 od 42,034 točk uporabljene mikromrežne zasnove POCI.



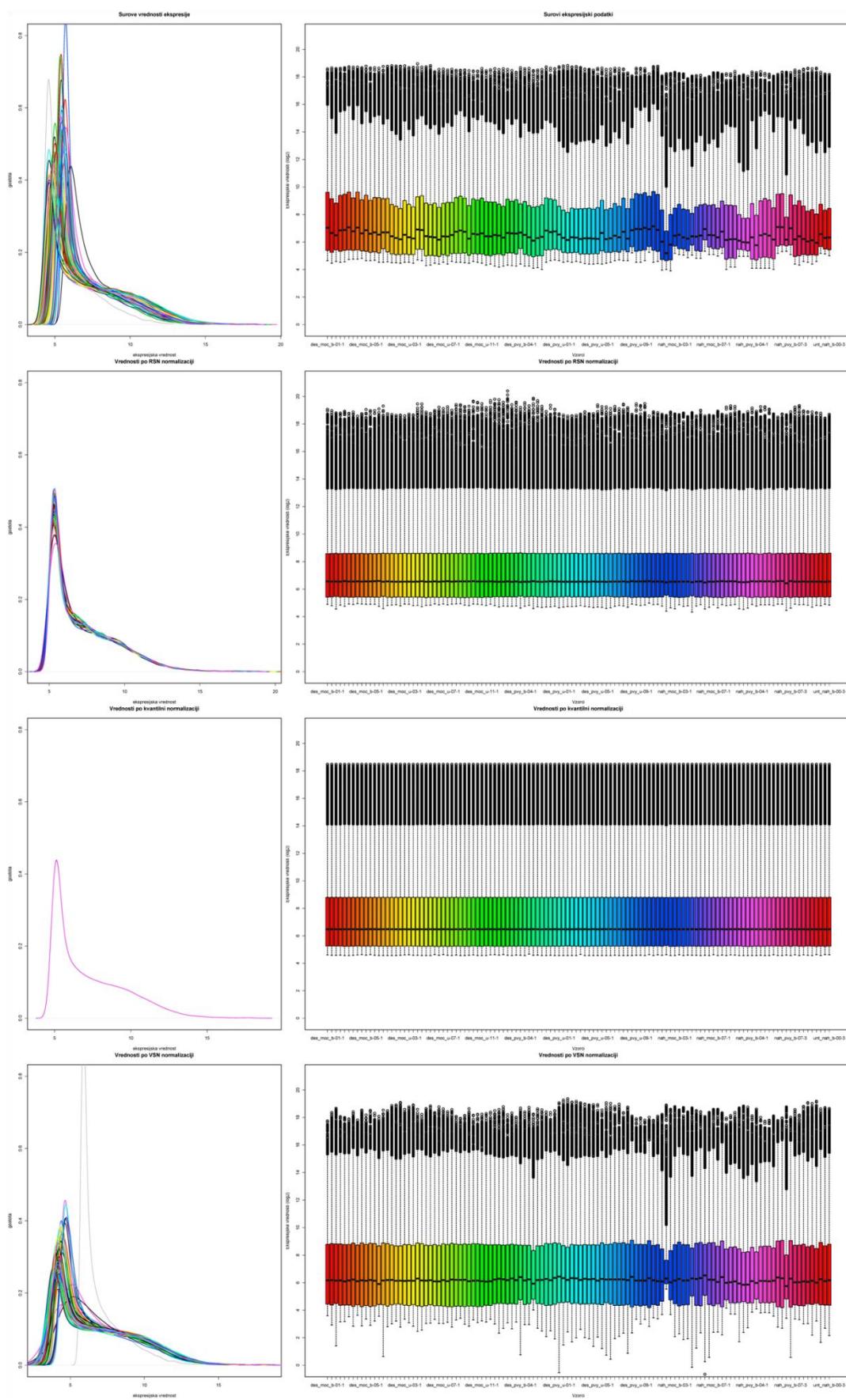
Slika 21: Koraki kvalitete kontrole pred normalizacijo.

Slika prikazuje gostotne diagrame (levo) in okvirji z ročaji (desno) vseh mikromrež, zgoraj so predstavljene intenzitete signalov, spodaj pa intenzitete ozadja.

Figure 21: Quality control steps taken prior to normalisation of microarrays.

The figure displays density plots (left) and boxplots (right) for all microarrays, on top signal intensities are represented and background intensities on the bottom.

Za zmanjšanje sistematičnih razlik intenzitet izmerjenih signalov hibridiziranih vzorcev, ki so posledica same tehnike priprave (koncentracija mRNA vzorca, učinkovitost označevanja vzorcev z barvilom, razlike v času hibridizacije), je potrebno vzorce normalizirati. Dobro je znano, da ima od različnih metod predprocesiranja normalizacija največji vpliv na končni rezultat, t.j. seznam diferencialno izraženih genov (ang. *differentially expressed genes*, DIG), zato smo na preostalih 37,971 kvalitativnih sondah mikromreže POCI uporabili tri različne metode normalizacije (kvantilno, RSN, VSN). Po normalizaciji je sledil ponoven korak kontrole kvalitete, v obliki pregleda gostotnih diagramov (Slika 22, levo) in škatel z okvirji (Slika 22, desno), kjer je bila opazna podobnost med kvantilno in RSN normalizacijo, medtem ko je VSN izstopala do takšne mere, da je nismo uporabljali v kasnejših korakih analize.



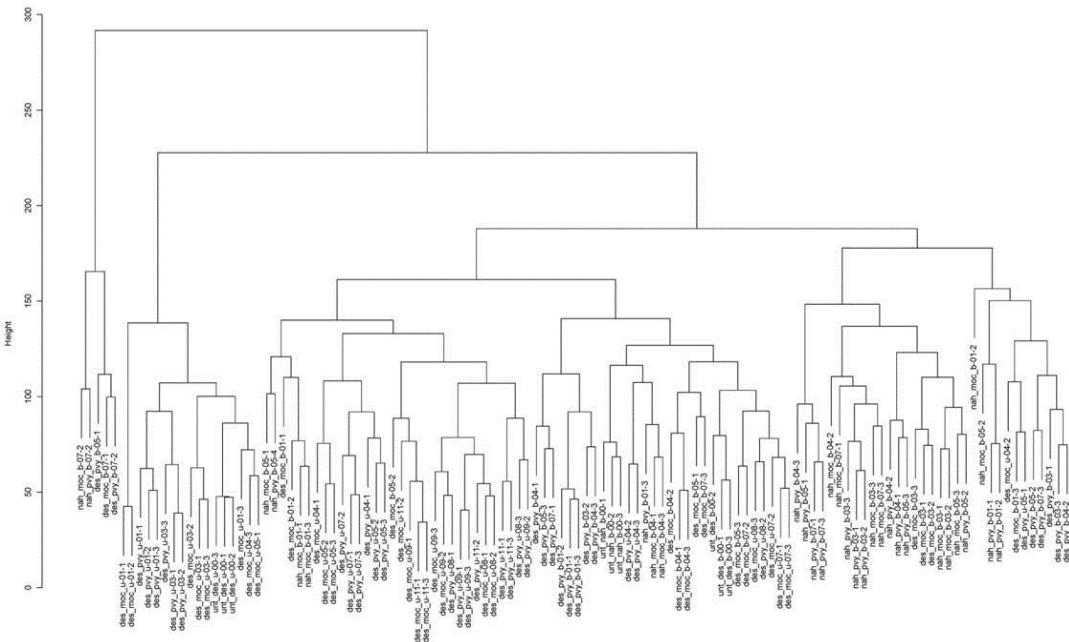
Slika 22: Primerjava normalizacij s prikazom okvirjev z ročaji ter gostotnimi diagrami.

Za surove vrednosti ekspresije (prva vrstica) in nato vse tri normalizacije (RSN, kvantilna, VSN v tem vrstnem redu) so prikazani izrisi gostotnih diagramov (levo) ter škatle z okvirji (desno) za vsakega od 118ih vzorcev. Gostotni diagrami po kvantilni normalizaciji imajo takšno podobnost, da na grafu ni možno ločiti med različnimi vzorci, medtem ko RSN gostotni diagrami še vedno nakazujejo na nekatere razlike. VSN gostotni diagrami so zelo različni, kar je razlog zakaj niso bili uporabljeni za kasnejše analize. Podobno velja za okvirje z ročaji, kjer ni razlik ob uporabi kvantilne normalizacije, obstoje manjše razlike z RSN normalizacijo ter velike razlike po VSN normalizaciji.

Figure 22: Comparison between normalisations via boxplot and density diagram displays.

For raw expression values (first line) and then all three normalisations (RSN, quantile, VSN in this order) density diagram displays (left) and boxplots (right) for each of the 118 samples are displayed. Density diagrams following the quantile normalisation are so similar, that we can't separate different samples on the diagram, while RSN density diagrams still display some differences. VSN density diagrams are very variable, which is the reason as to why they were not used in following analyses. Similar holds true for boxplots, where there are no observable differences when using the quantile normalization, smaller differences for RSN normalisation and large differences after VSN normalisation.

Zaradi podobnosti med kvantilno in normalizacijo RSN smo za dodatne poglede v strukturo podatkov uporabljali vrednosti RSN, in sicer za pregled sorodnosti med mikromrežami glede na dendrogram podobnosti evklidskih razdalj med vzorci (Slika 23) ter izris topotnih grafov neurejenih vzorcev glede na Pearsonov koeficient podobnosti (Slika 24).

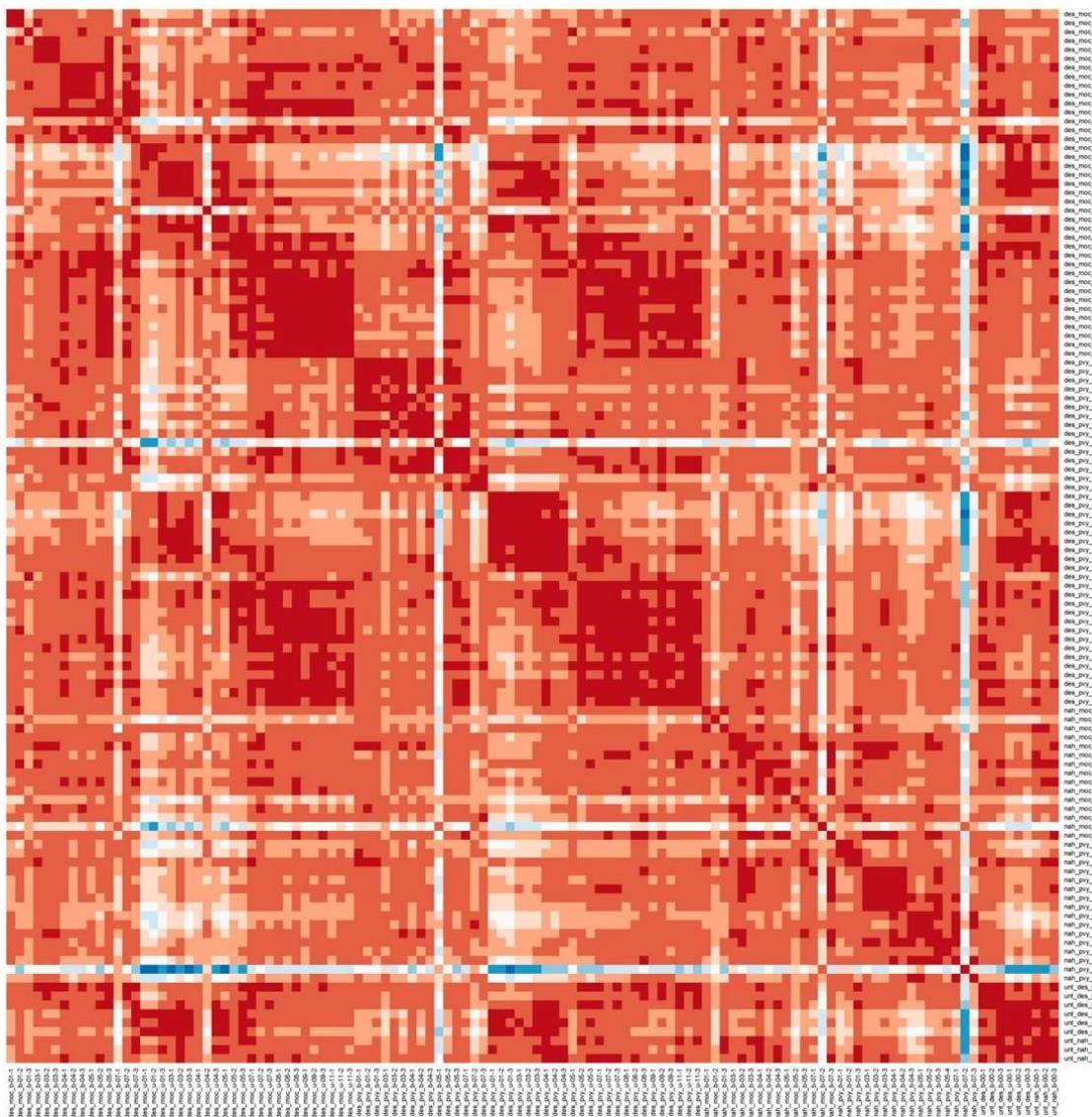


Slika 23: Rezultat razvrščanja podatkov mikromrež v obliki dendrograma.

Prikazan je dendrogram pridobljen z uporabo evklidske razdalje. Nastale skupine dendrograma so bile povečini organizirane po pričakovanjih glede na dano biološko skupino vzorcev.

Figure 23: The resulting dendrogram after using the method of hierarchical clustering.

Dendrogram obtained using the Euclidian distance is shown. Formed groups of the dendrogram were mostly organised as expected, given the biological sample background.



Slika 24: Toplotni graf podatkov mikromrež.

Toplotni grafi so bili izrisani na podlagi izračunanih Pearsonovih koeficientov med vsemi vzoreci. Med podobnim biološkim ozadjem in triplikati je opazna visoka podobnost (na diagonali topotnega grafa). Dejanske vrednosti Pearsonovega koeficiente niso bile manjše od 0,85, tudi za manj podobne vzorce (modre barve).

Figure 24: The resulting heatmap for the microarray data.

Heatmaps were drawn using Pearson coefficients as a distance measure between samples. High similarities can be observed between similar biological conditions and triplicates (as seen on the diagonal of the heatmap). Actual values of the calculated Pearson coefficients were all higher than 0,85, even for the less similar samples (blue colour).

Po normalizaciji postanejo opazne biološke razlike med vzorci brez ostalih vplivov, omogočena pa je tudi primerjava izražanja genov med različnimi mikromrežami. Za primerjavo med vzorci okuženimi z virusom PVY ter negativnimi kontrolami (voda) smo uporabili linearni model in t-test z empirično Bayesovo metodo. Zaradi velike količine

primerjav (~37,000 na primerjavo) smo uporabili še mero deleža napačnih uvrstitev FDR. Za kasnejše analize je bil gen označen kot diferencialno izražen, če je bila njegova popravljena vrednost FDR manjša od 0,05 za oba tipa normalizacij (presek med rezultati kvantilne in robustne normalizacije). S tem smo sicer zmanjšali število diferencialno izraženih genov, vendar povečali zanesljivost preostalih.

Preglednica 8: Število diferencialno izraženih genov pri obeh tipih normalizacije in njunem preseku.
Opravljenih je bilo 18 primerjav med različnimi tipi vzorcev in pripadajočih dnevi po primerjavi virusnega in kontrolnega tretmaja. Gen je bil označen kot diferencialno izražen, če je bila njegova FDR p-vrednost manjša od 0,05.

Table 8: Number of differentially expressed genes using both normalisations and their intersection.
There were a total number of 18 comparisons performed between different sample types and corresponding days after virus and mock treatments. A gene was deemed to be differentially expressed if it's FDR p-value was smaller than 0,05.

Opis primerjave	DIG (kvantilna)	DIG (RSN)	DIG preseka
Désirée inokulirani, 1 dan po okužbi	2,639	2,765	2,553
Désirée inokulirani, 3 dan	2,171	2,243	2,063
Désirée inokulirani, 4 dan	3,524	3,646	3,366
Désirée inokulirani, 5 dan	4,187	4,385	4,019
Désirée inokulirani, 7 dan	3,542	3,544	3,412
NahG-Désirée inokulirani, 1 dan po okužbi	8,040	8,201	7,731
NahG-Désirée inokulirani, 3 dan	386	367	338
NahG-Désirée inokulirani, 4 dan	4,204	4,458	4,082
NahG-Désirée inokulirani, 5 dan	1,877	2,128	1,800
NahG-Désirée inokulirani, 7 dan	2,287	2,380	2,157
Désirée ne-inokulirani, 1 dan po okužbi	4,034	4,060	3,880
Désirée ne-inokulirani, 3 dan	2,560	2,532	2,423
Désirée ne-inokulirani, 4 dan	1,729	1,752	1,652
Désirée ne-inokulirani, 5 dan	1,557	1,382	1,356
Désirée ne-inokulirani, 7 dan	901	887	841
Désirée ne-inokulirani, 8 dan	4	2	2
Désirée ne-inokulirani, 9 dan	917	950	884
Désirée ne-inokulirani, 11 dan	1,234	1,223	1,184

Končni podatki normalizacije so bili naloženi v Gene Expression Omnibus (GEO) (Edgar in sod., 2002) in so dostopni preko GEO pristopne številke GSE58593.

4.2 INTEGRACIJA ZAPOREDIJ V OKVIRU ENE RASTLINSKE VRSTE

4.2.1 Pan-genom krompirja

V prvem koraku postopka smo z uporabo algoritma CD-HIT združili oba genska modela (PGSC, ITAG) z uporabo prazne vrednosti 95 % podobnosti ter množici zaporedij unigene (POCI, StGI) pri 90 % podobnosti, rezultati so prikazani v Preglednici 9. Glede na končno število skupin CD-HIT je opazen velik delež prekrivanja med genskima modeloma, medtem ko končno število pri množicah unigene nakazuje na veliko redundantnost zaporedij, saj je končno število množic zaporedij unigene CD-HIT manjše od začetnega števila zaporedij StGI.

Preglednica 9: Zmanjšanje redundantnosti zaporedij po uporabi algoritma CD-HIT na posebej genskih modelih in posebej množicah zaporedij unigene.

Uporabljeni pragi identitete se med obema tipoma zaporedij razlikujejo – za genske modele smo uporabili strožji kriterij (95 %), saj so zaporedja genskih modelov osnovana na identičnih zaporedij molekul kromosomov. Za razliko od genskih modelov smo pri množicah unigene, osnovanih na zbirkah zaporedij EST različnih kultivarjev krompirja, uporabili manj strog kriterij (90 %).

Table 9: The decrease in sequence redundancy, after gene models and unigene sets were clustered separately, using the CD-HIT algorithm.

Unequal threshold were chosen between both sequence types – as gene models are based on identical superscaffolds, a more strict threshold was chosen (95 %). Less strict criteria was chosen for unigene sets (90 %), as they are based on a collection of EST-derived sequences from different cultivars

	Množica zaporedij	Začetno število zaporedij	Prag % identitete	Končno število skupin CD-HIT
Genska modela	PGSC	39031	95 %	44456
	ITAG	35004		
Množici zaporedij unigene	POCI	46345	90 %	51359
	StGI	62330		

V drugem koraku postopka smo uporabili algoritem BLASTx in primerjali nukleotidna zaporedja množic unigene proti proteinski bazi genskih modelov z zelo blagimi kriteriji BLASTx. Rezultat primerjave je iz tega razloga vseboval veliko število zadetkov (2,802,047), od katerih pa jih je bilo veliko lažno pozitivnih, zaradi same narave delovanja algoritma BLAST. Ker deluje na osnovi lokalnih in ne globalnih poravnav, je veliko rezultatov lahko na kratkih odsekih imelo velik odstotek identičnosti ali podobnosti med zaporedjem.

V tretjem koraku smo za odstranitev večine lažno pozitivnih zadetkov algoritma BLAST uporabili preizkusili tri algoritme za učenje odločitvenih dreves – J48, ADTree in JRip. Učna množica je vsebovala 310,179 povezav BLASTx, porazdelitev razrednega atributa je bila 30 % za 'je ortolog' in 70 % za 'ni ortolog'. Rezultat algoritma J48 je bilo obširno odločitveno drevo s 6,569 dogodkovnimi vozlišči (ang. *node*) in 3,285 končnimi vozlišči (ang. *leaf*). Odločitveno drevo algoritma ADTree je bilo precej manjše, s 99,5 % pokritostjo osnovnih podatkov, 31 končnimi vozlišči, od teh 21 napovednih vozlišč. Rezultat algoritma JRip namesto odločitvenega drevesa predstavlja 34 pravil. Vsi algoritmi so dosegli na testni množici dosegli natančnost višjo od 90 %, od katerih je imel najvišjo algoritom J48; dodatno je imel slednji tudi najboljše razmerje med lažno pozitivnimi in lažno negativnimi določitvami (Preglednica 10).

Preglednica 10: Rezultati validacije treh algoritmov učenja odločitvenih dreves.

Primerjava rezultatov validacije med tremi uporabljenimi algoritmi za učenje odločitvenih dreves – J48, ADTree in JRip. Največjo natančnost je dosegel algoritmom J48, medtem ko med natančnostjo algoritmov ADTree in JRip ni razlik. Ob primerjavi lažno pozitivnih in lažno negativnih določitev je spet najboljši J48, medtem ko imata ADTree in JRip visoko vrednost lažno pozitivnih določitev, kar je bilo nezaželeno za končni cilj uporabe.

Table 10: Validation results of the three algorithms used to learn the decision trees.

Comparison of validation results between the three used algorithms for decision tree learning – J48, ADTree and JRip. The highest accuracy is achieved by J48, while ADTree and JRip do not differ in that respect. When it comes to the false positive and false negative rates, again J48 performs best, where ADTree and JRip have a larger false positive rate, which was undesirable for the given problem.

Uporabljen algoritem	Natančnost algoritma	Vsebnost lažno pozitivnih določitev	Vsebnost lažno negativnih določitev
J48	93,0 %	3,8 %	3,2 %
ADTree	91,6 %	5,0 %	3,4 %
JRip	91,6 %	5,0 %	3,3 %

Po primerjavi rezultatov validacije vseh treh algoritmов smo zaradi najvišje natančnosti ter najmanjše vsebnosti tako lažno pozitivnih kot lažno negativnih določitev izbrali algoritmom J48. Naučeno drevo smo uporabili pri filtriranju začetne skupine rezultatov BLASTz iz skoraj 3 milijonov zadetkov na 170,905 zadetkov (94 % zmanjšanje) – obdržali smo določitve z oznako 'je ortolog'. Za te zadetke smo izračunali medsebojne uteži, kar je dodatno zmanjšalo število zaporedij na 80,962. Le-ta so nadaljevala v zadnji korak postopka, analize omrežij v programu Pajek, za združevanje v paralogne skupine.

Po uporabi razvite metodologije za združevanje zaporedne informacije iz različnih virov, smo preko primerjav zaporedij, učenja odločitvenega drevesa ter analize omrežij pridobili niz 35,609 paralognih genskih skupin krompirja, od katerih je imelo 20,809 skupin ITAG, 9,509 PGSC, 2,882 POCI ter 2,409 StGI13 reprezentativni identifikator. Skupine so predstavljene v obliki pretvorniške tabele, ki je javno dostopna kot del orodja GoMapMan. Primer ene izmed dobljenih paralognih skupin s skupno deljeno podobnostjo zaporedij je podan v Preglednici 11 ter na Sliki 25.

Preglednica 11: Primer paralogne skupine za etilensko odzivni faktor 5 (ERF5).

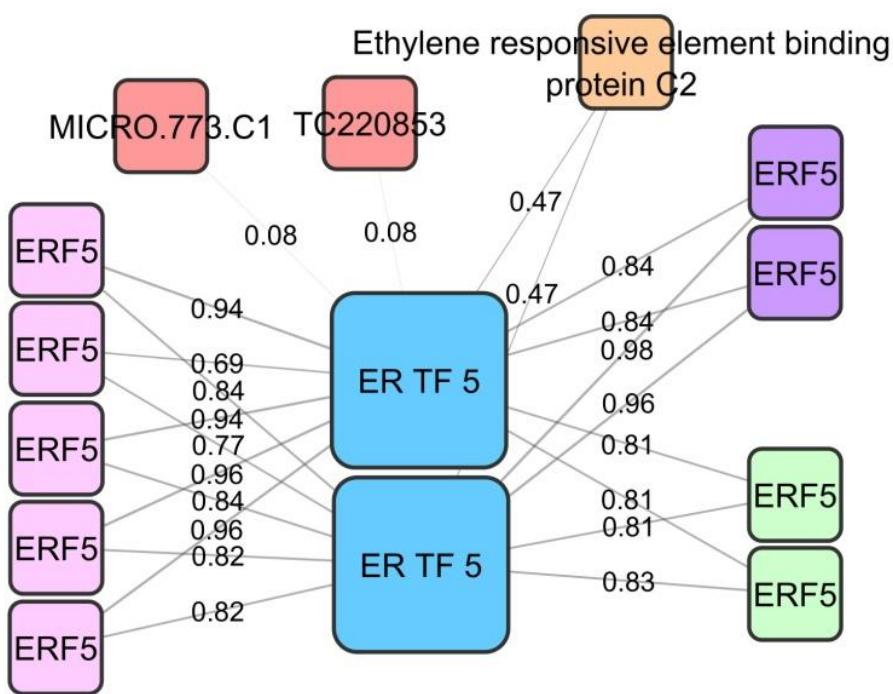
Vsaka paralogna skupina krompirja ima tako imenovan identifikator StNIB, v tem primeru je predstavljen StNIB_14511. Vsaka skupina ima reprezentativni identifikator, kot označeno v tabeli (Sotub12g018750.1.1). Identifikatorji so unikatni za celo tabelo, s čimer se izognemo potencialnim zamenjavam ob uporabi večih različnih virov hkrati, t.j. ITAG (Tomato Sequencing Consortium), PGSC (Potato Genome Sequencing Consortium), POCI (Potato Oligo Chip Initiative) ali StGI13 (Dana Farber Cancer Institute *Solanum tuberosum* Gene Index, v13). Dodatno so ohranjeni tudi opisi za identifikator iz vsakega vira.

Table 11: An example of a resulting parologue group for ethylene response factor 5 (ERF5).

Each potato parologue group receives a so-called StNIB identifier, StNIB_14511 is represented here. Each group also has a representative identifier, which has been marked in the table itself (Sotub12g018750.1.1). Identifiers are all unique for the whole table, thus avoiding potential mix-ups when using several sources, which can be ITAG (Tomato Sequencing Consortium), PGSC (Potato Genome Sequencing Consortium), POCI (Potato Oligo Chip Initiative) or StGI13 (Dana Farber Cancer Institute *Solanum tuberosum* Gene Index, v13). Additionally, the original descriptions as defined by the respective source are included.

Podskupina	Vir	Unikatni identifikator	Opis
gen1	ITAG	Sotub12g018750.1.1*	ERF5
gen1	PGSC	PGSC0003DMG400036566	ERF5
gen2	ITAG	Sotub12g018760.1.1	ERF5
gen2	PGSC	PGSC0003DMG400040046	ERF5
est1	POCI	MICRO.773.C9	ERF5
est1	StGI13	TC210717	ERF5
est2	POCI	bf_ivrootxx_0006e10.t3m.scf	ERF5
est2	POCI	MICRO.773.C3	ERF5
est2	POCI	MICRO.773.C8	ERF5
est2	StGI13	TC195612	ERF5
est2	StGI13	TC196386	ERF5
est3	POCI	MICRO.773.C2	ERF5
est3	StGI13	TC211336	ERF5
est4	POCI	MICRO.773.C1	MICRO.773.C1
est4	StGI13	TC220853	TC220853
est5	StGI13	CN463250	ET binding protein C2

* Ta gen je obenem tudi predstavnik cele skupine (ang. *representative identifier*).



Slika 25: Predstavitev paralogne skupine genov krompirja.

Vrednosti na povezavah predstavljajo vrednosti izračunanih uteži z ortogonalno regresijo za gensko skupino StNIB_14511. Centralni vozlišči predstavljata dve skupini iz genskih modelov (modro), preostanek so vsa zaporedja izpeljana iz EST-tehnologije (oranžna, rdeča, rožnata, vijolična, zelena). Uteži so povsod visoke (>0.65), z izjemo treh zaporedij: dve zaporedji nimata niti človeško berljivega opisa (rdeča) in eno zaporedje (oranžna) ima v opisu le indikacijo, da se veže z etilensko odzivnim faktorjem.

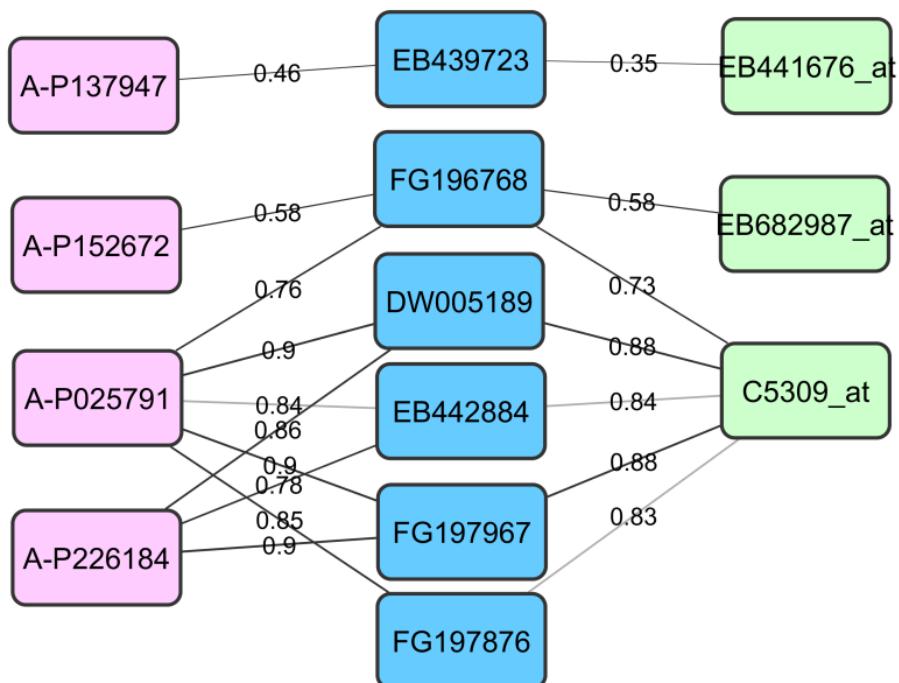
Figure 25: Representation of a potato parologue gene group.

The values of the edges represent the final calculated weights from orthogonal distance regression for the gene group StNIB_14511. The two central nodes represent two clusters from gene model predictions (blue), the remainder are all EST-derived unigene sequences (orange, red, pink, purple, green). The weights are all high (>0.65), with the exception of three sequences: for two of these there is no human readable description (red) and one (orange) is indicated as merely binding an ethylene response factor.

V okviru genskih modelov ITAG in PGSC posebej smo preverili še število paralognih skupin, ki so vsebovale več kot en gen vsake od skupin posebej (pregled razpona ponovljenih genov). Vseh 35,004 napovedanih genov skupine ITAG se je združilo v 20,809 paralognih skupin, od teh 14,481 z enim samim genom krompirja te skupine, ostalih 6,328 skupin pa je pokrilo preostalih 20,523 napovedanih genov. Od vseh 39,031 napovedanih genov skupine PGSC pa je nastalo 26,181 paralognih skupin, od katerih jih je 20,298 vsebovalo en sam gen krompirja, ostalih 5,883 skupin pa je pokrilo preostalih 18,733 genov.

4.2.2 Pan-transkriptom tobaka

Z unijo nukleotidnih zaporedij (NCBI, TIGR in ESTtobacco) prisotnih v eni ali drugi tobakovi mikromreži (Agilent, Affymetrix) in naknadno združevanje v skupke (s podobnostjo zaporedij večjo od 95 %) smo dobili 44,848 skupkov CD-HIT. Ob povezavi mikromrežnih zaporedij z genskimi skupinami NCBI UniGene (verzija 17) se je skoraj 75 % skupkov CD-HIT (33,254) povezalo s 78 % vseh obstoječih zaporedij iz skupno 24.432 skupin NCBI UniGene. Preostalih 11,594 mikromrežnih zaporedij brez zadetka v NCBI UniGene (v.17) smo definirali kot 11,594 novih transkriptov in jih združili z 24,432 skupinami NCBI UniGene. Dodatno se te skupine da razširiti na vse GenBank, DFCI in TIGR identifikatorje, kjer smo z našo integracijo pokrili 70 % vseh tobakovih znanih zaporedij iz junija 2012 (352,765 zaporedij). Primer enega od končnih skupkov CD-HIT je na Sliki 26.



Slika 26: Predstavitev paralogne skupine transkriptov tobaka.

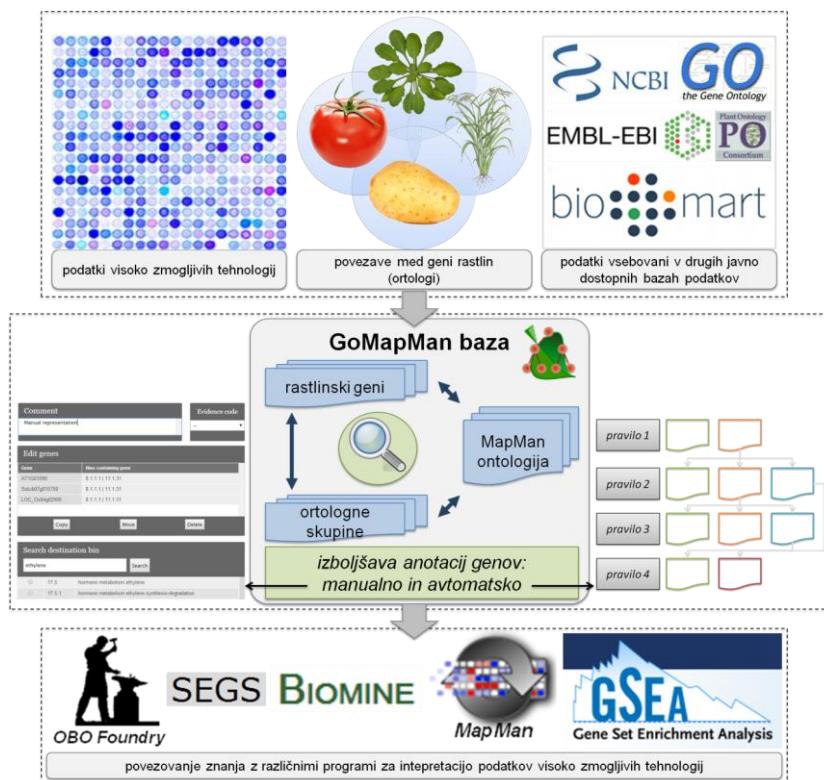
Na sliki je predstavljena skupina tobakovih transkriptov, ki vključuje zaporedja NCBI UniGene (modro), sonde mikromrež Agilent (rožnata) in Affymetrix (zelena).

Figure 26: Representation of a resulting tobacco transcript group.

The figure displays a tobacco cluster which includes the NCBI UniGene sequences (blue), the Agilent microarray probe (pink) and Affymetrix probe (green).

4.3 GOMAPMAN: POVEZOVANJE ZNANJA MED RASTLINSKIMI VRSTAMI

Za pripravo aplikacije GoMapMan (Slika 27) smo v prvem delu pripravili vse njene osnovne gradnike, t.j. ontologijo MapMan (Thimm in sod., 2004), pripravili rastlinske vrste in ortologne skupine ter povezave z javnimi bazami podatkov (Poglavlje 4.3.1). V drugem delu smo zasnovali metode popravljanja funkcionalnih opisov genov v ontologiji preko avtomatskih konsolidacij ter preko implementacije uporabniškega vmesnika za ročno popravljanje (Poglavlje 4.3.2). Obenem smo v tem koraku poskrbeli za sledljivost sprememb ter integracijo z drugimi orodji, ki omogočajo analizo podatkov visoko zmogljivih tehnologij.



Slika 27: Shematska predstavitev aplikacije GoMapMan in njenih funkcionalnosti.

GoMapMan združuje povezave med geni različnih rastlinskih vrst preko ortolognih skupin, povezave s podatki visoko zmogljivih tehnologij in povezave na razpršene javno dostopne baze podatkov. Geni so postavljeni v ontološko drevo MapMan, za kar sta razvita tako avtomatski kot način ročnega popravljanja položajev genov. Povezave med geni in ontološkimi vozlišči ontologije MapMan ter nekatere anotacije so periodično izvožene v datoteke za programe za analizo podatkov visoko zmogljivih tehnologij.

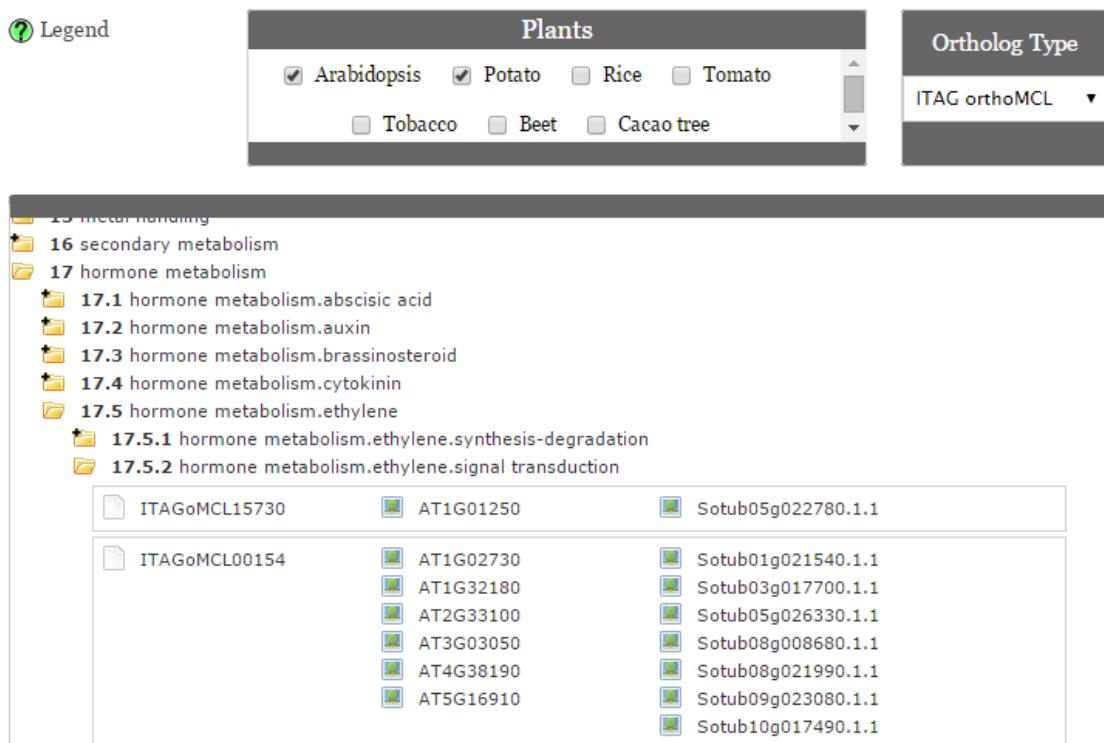
Figure 27: A schematic representation of GoMapMan application with its functionalities.

GoMapMan combines connections between different plant species genes via orthologue groups, connects them with high throughput data and dispersed publicly available databases. Genes are placed into the MapMan ontology tree, for which both automatic and manual corrections of ontology positions were developed. Connections between genes and MapMan ontology nodes and some annotations are periodically exported into files for analyses of high-throughput data sets.

4.3.1 Implementacija podatkovne baze GoMapMan

Glavni pogled, ki združuje drevo ontologije MapMan, sezname osnovnih gradnikov rastlinskih vrst (genov oz. transkriptov v primeru pan-transkriptoma tobaka) ter različne tipe ortologih skupin je na Sliki 28. Razdelke ontologije MapMan se lahko odpira dokler se ne doseže najbolj podrobne hierarhične enote ontologije, v vsakem razdelku pa so naštete ortologne skupine in geni s tem funkcionalnim opisom. Uporabnik lahko izbere rastline in ortologni tip, v odvisnosti od željene širine ortologih skupin. Zaradi boljše preglednosti se lahko naenkrat opazuje samo en ortologni tip naenkrat.

Ontology tree



Slika 28: Pogled ontološkega drevesa aplikacije GoMapMan.

Hierarhično organizirani funkcionalni opisi ontologije MapMan so prikazani z ikonami odprte ali zaprte mape, kjer ima slednja plus simbol (indikator da vsebuje še bolj specifični funkcionalni opisi). Vsaka funkcionalna kategorija nato vsebuje naštete identifikatorje izbranega ortolognega tipa in genov izbranih rastlin.

Figure 28: Ontology Tree View of GoMapMan application.

Hierarchically organized MapMan BINs are displayed as opened or closed folders, the latter having a plus symbol (indicating that they contain further subBINs). Each BIN then lists identifiers of the selected orthologue type and genes of selected plants.

4.3.1.1 Popravljanje ontologije MapMan

Ontologija MapMan je bila zasnovana tako fleksibilno, da se lahko ob pridobivanju novega znanja o bioloških poteh in procesih ontološko drevo širi naprej preko dodajanja novih otroških vej ontologije (dokler le-te že niso definirane). To je v primeru priprave skupne ontologije MapMan za vnos v aplikacijo GoMapMan pomenilo potrebo po usklajevanju vseh terminov različnih verzij datotek. Zadnja verzija ontologije (april 2015) vsebuje 1895 vozlišč, in je bila od prve implementacije v aplikaciji GoMapMan (marec 2012) do sedaj izboljšana dvakrat. Prvič je bilo z dodatkom bolj nadrobnega opisa funkcij genov, ki sodelujejo pri utišanju RNA (razdelki skupine 27.5, RNA.silencing), kar je bil rezultat diplomskega dela Marka Cheriscole. Dodatno smo v okviru dela izboljšali tudi strukturo razdelkov ontologije, ki opisujejo receptorje in PR-proteine prisotne v procesu biotskega stresa. Starejše verzije ontologije so imele definirane le skupini 20.1.2 (stress.biotic.receptors) ter 20.1.7 (stress.biotic.PR-proteins), mi pa smo dodali še podskupine za oba razdelka, kot je prikazano v Preglednici 12.

Popravki ontologije so možni tudi v obratno smer, z odstranjevanjem določenih razdelkov, kar se je zgodilo v primeru razdelkov 36 (C4) in 36.1 (C4.photosynthesis) in kasneje izkazalo, da je artefakt ene od datotek specifične za koruzo (*Zea mays*). Razlogi za implementacijo takega razdelka so nam bili neznani in ker aplikacija trenutno še ne vsebuje rastlin s C4 tipom fotosinteze, sta bila razdelka odstranjena. V primeru prihodnje implementacije npr. koruze v ontologijo pa bodo potrebne spremembe na ontologiji v razdelku 1 (s funkcionalnim opisom fotosinteze), kar bo moralo potekati skladno z eksperti področja.

Preglednica 12: Izboljšava strukture ontologije MapMan za biotski stres.

Skupina genov za odpornost je bila dalje razdeljena tako, da vključuje bolj natančno definicijo dveh možnih oblik, toll interleukin 1 receptor (20.1.2.1) in coiled-coil strukturirane (20.1.2.2). Novi razdelki za proteine povezane s patogenezo so bili dodani glede na eksperumno znanje (Van Loon in Van Strien, 1999). Dodatno poenostavitev predstavlja združitev definiranih različnih hitinaznih PR-proteinskih tipov v en sam funkcionalni opis (20.1.7.3, hitinaze in proteini, ki vežejo hitin), ampak smo pustili prostor za nadgradnjo te strukture v prihodnosti (v tabeli niso definirani proteini PR4, ki vežejo hitin; PR8 hitinaze tipa III ter PR11, hitinaze tipa I).

Table 12: The improved structure of MapMan ontology for biotic stress.

Resistance genes group was further split to include a more fine grained definition of the two types, the toll interleukin 1 receptor (20.1.2.1) and the coiled-coil structured ones (20.1.2.2). The new pathogenesis-related (PR) protein BINs were added according to a merge of information from expert knowledge (Van Loon in Van Strien, 1999). An additional simplification is represented in the merger of the defined different chitinase PR-protein types into a single BIN (20.1.7.3, chitinases and chitin binding proteins), but placeholders have been preserved for a future improvement (note that there are no BINs for PR4 chitin binding proteins, PR8 chitinases III or PR11 chitinase I).

Koda razdelka	Funkcionalni opis kategorije
20.1.2.1	stress.biotic.receptors.CC-NB-LRR
20.1.2.2	stress.biotic.receptors.TIR-NB-LRR
20.1.7.1	stress.biotic.PR-proteins.PR1 (antifungal)
20.1.7.2	stress.biotic.PR-proteins.PR2 (endo-beta-1,3-glucanases)
20.1.7.3	stress.biotic.PR-proteins.PR3 (chitinases and chitin binding proteins)
20.1.7.5	stress.biotic.PR-proteins.PR5 (thaumatin-like proteins)
20.1.7.6*	stress.biotic.PR-proteins.PR6 (proteinase inhibitors)
20.1.7.7	stress.biotic.PR-proteins.PR7 (endoproteases)
20.1.7.9	stress.biotic.PR-proteins.PR9 (peroxidases)
20.1.7.10	stress.biotic.PR-proteins.PR10 (ribonucleases)
20.1.7.12*	stress.biotic.PR-proteins.PR12 (plant defensins)
20.1.7.13	stress.biotic.PR-proteins.PR13 (thionins)
20.1.7.14	stress.biotic.PR-proteins.PR14 (lipid transfer proteins)
20.1.7.15	stress.biotic.PR-proteins.PR15 (oxalate oxidases)
20.1.7.16	stress.biotic.PR-proteins.PR16 (oxalate oxidase-like proteins)

* Ta kategorija je bila prisotna že pred našimi popravki ontologije.

V splošnem so popravki ontologije zelo strogo nadzorovani in dovoljeni samo za skrbnike podatkovne baze z ročnim vnosom v za to določeno tabelo. Skripte za vnos v velikem merilu tako tudi niso razvite, saj iz vidika sprememb ontologije MapMan ne pričakujemo pogostih sprememb.

4.3.1.2 Vnos rastlinskih vrst

Aplikacija GoMapMan trenutno vsebuje genske modele petih rastlin, pan-genom krompirja (iz Poglavlja 4.2.1) ter pan-transkriptom tobaka (iz Poglavlja 4.2.2). Genski model rastline navadnega repnjakovca je zastarel (TAIR9), vendar so bile razlike med leto in najnovejšo TAIR10 verzijo minimalne, 126 novih genov, kar predstavlja 0.3 % celotnega proteoma rastline. Podobno kot velja za krompir, tudi za riž obstajata dva genska modela napovedana na enakih psevdomolekulah (kromosomih) s spremljajočo pretvorniško tabelo. V bazi GoMapMan je vnešen genski model riževega genomskega projekta MSU-v7. Za paradižnik, kakav in sladkorno peso so vnešeni najnovejši genomski modeli in sicer SL2.40/ITAG2.3 za paradižnik, Phytozome9.1 za kakav ter RefBeet1.1 za sladkorno peso. Pan-genom krompirja ter pan-transkriptoma pa sta vnešena kot smo jih pripravili v Poglavlju 3.2.

Informacija vsakega gena je združena v posebnem pogledu o genski informaciji (Slika 29), ki vsebuje gensko specifično informacijo (rastlinsko vrsto, opis, kratko ime, sinonime ter opis lokacije v genomu), če je bila ta podana v osnovnih podatkih. Sledijo našteti razdelki MapMan, ki izbrani gen funkcionalno opisujejo ter ortologne skupine, v katerih se ta gen nahaja. Zadnji del vsebuje naštete transkripte ali identifikatorje mikromrežnih sond, v primeru da le-te obstoje.

Gene details

Identifier	Sotub05g022780.1.1
Plant	potato
Description	Ethylene-responsive transcription factor 9
Short Name	-
Synonym	-
Genomic Context	-
Source	stNIB-v1 (PGSC_DM_v3.4 and iTAG Potato Gene Model (to PGSC v2.1.11 Pseudomolecules).)
Last Modified	21 Dec 2013

Ontology annotations

Click on bin to open it in ontology tree.

Bin Code	Bin Name	Evidence Code
17.5.2	hormone metabolism.ethylene.signal transduction	ISS
27.1	RNA.processing	ISS
27.3.3	RNA.regulation of transcription.AP2/EREBP, APETALA2/ethylene-responsive element binding protein family	ISS

Orthologues

PGSCoMCL01203 PGSCoMCL11770 ITAGoMCL15730 wc00030

Transcripts or microarray features for the gene

Microarray feature	Description	Source
ACDA03544H04.T3m.scf	transcription factor JERF1 [Lycopersicon esculentum]	stu_Agilent_4x44k
MICRO.6769.C3	transcription factor JERF1 [Lycopersicon esculentum]	stu_Agilent_4x44k

Slika 29: Pogled informacije o posameznem genu v aplikaciji GoMapMan.

Vsek izbran gen/transkript se povezuje s starševsko podatkovno bazo rastline (npr. TAIR za navadni repnjakovec), sledi zbrana informacija iz iste starševske baze, če je na voljo (opisi genov, kratka imena, sinonimi, genomski kontekst). Naslednji trije deli opisa nato naštejejo vsi ontološki funkcionalni opisi gena, v katere ortologne skupine pripada ter končno, če se nanj navezujejo transkripti ali identifikatorji mikromrežnih sond.

Figure 29: Gene Details View of a single gene in GoMapMan application.

Each selected gene/transcript links out to the parental database of that plant (e.g. TAIR for arabidopsis), followed by other information retrieved from the same parental database (gene descriptions, short names, synonyms and genomic context), if available. The next three sections then list all ontology annotations for that gene, to which orthologue clusters the gene belongs to and lastly, if any transcripts or microarray features are connected to it.

4.3.1.3 Vnos ortolognih skupin

Aplikacija GoMapMan vsebuje ortologne skupine iz javnih virov in je bila zasnovana tako, da omogoča več različnih ortolognih razvrstitev, saj se metode določanja ortologih skupin razlikujejo glede na njihovo občutljivost in specifičnost (Altenhoff in Dessimoz, 2009). Tako se določene skupine razlikujejo zaradi značilnosti samih podatkov in taksonomskega konteksta (Chan in sod., 2013; Powell in sod., 2014). S tem razlogom smo dovolili vključitev ortolognih skupin s širšim ali ožjim taksonomskim razponom, katerih rezultat so skupine nižje ali višje specifičnosti (Slika 30).

Orthologue details

Identifier	ITAGoMCL15730
Description	-
Type	ITAG_orthoMCL
Last Modified	18 Sep 2014

Ontology annotations

[Click on bin to open it in ontology tree.](#)

Bin Code	Bin Name	Evidence Code
17.5.2	hormone metabolism.ethylene.signal transduction	ISS
27.1	RNA_processing	ISS
27.3.3	RNA.regulation of transcription.AP2/EREBP_APETALA2/ethylene-responsive element binding protein family	ISS

Orthologue genes

Arabidopsis
AT1G01250: AP2 domain-containing transcription factor; putative
Potato
Sotub05g022780.1.1: Ethylene-responsive transcription factor 9

Slika 30: Pogled informacije o ortologni skupini v aplikaciji GoMapMan.

Podobno kot pri genski strani je tudi tu najprej podana informacija za ortologno skupino (identifikator, ortologni tip ter opis) in funkcionalni opisi. Nato sledijo po rastlinah ločeni nasteti geni rastlin in njihovi opisi, ki pripadajo izbrani ortologni skupini.

Figure 30: Orthologue Details View of GoMapMan application.

Similarly as in the Gene Details View information for the orthologue group is listed (identifier, orthologue type and description) and MapMan functional descriptions. At the bottom the genes belonging to orthologue grouping with descriptions are listed, separated by plant species.

Od vključenih najbolj strog ortologni tip predstavlja ITAG_RSD, kjer so ortologne skupine najmanjše glede na povprečno število genov v enaki ortologni skupini. PGSC_oMCL predstavlja vmesni nivo natančnosti, saj je bil pripravljen s proteomi petih bližnje sorodnih rastlinskih vrst, PGSC_oMCL pa je vseboval 12 rastlin vseh področij filogenetskega drevesa rastlin. Zaradi teh razlik med različnimi ortolognimi tipi, smo se odločili vse povezave združiti z uporabo šibkih komponent v programu Pajek (Batagelj in Mrvar, 1998). Dodatna motivacija so bila visoka razmerja genov, ki so bili vključeni vsaj v eno ortologno skupino (Preglednica 13).

Preglednica 13: Število proteinsko kodirajočih genov za vsako rastlino in njihova vključenost v katerokoli od ortolognih skupin.

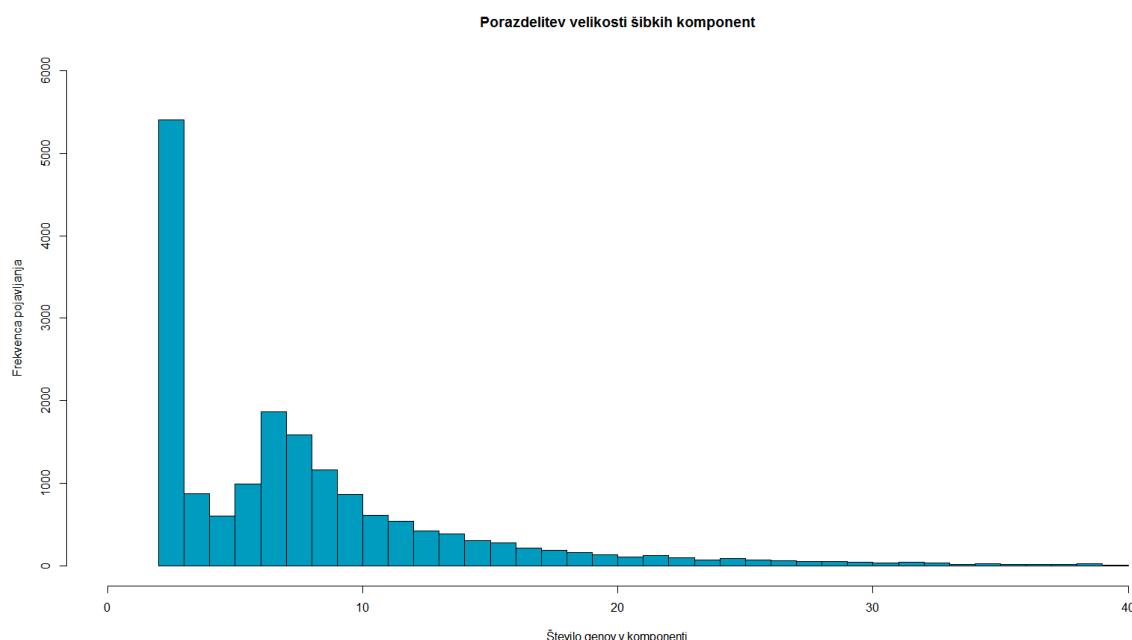
Preglednica vsebuje samo štiri od trenutno sedmih rastlin vključenih v aplikaciji GoMapMan, saj preostale tri rastline (tobak, sladkorna pesa ter kakav) niso bile vključene v katerega od trenutnih ortolognih skupkov. Pri vseh rastlinah je razmerje vključenih proteinsko kodirajočih genov večje od 60 %, za navadni repnjakovec in riž tudi preko 70 %. Krompir in paradižnik imata nižje vrednosti, saj sta evolucijsko bolj oddaljena, pa tudi nista vsebovana v vseh ortolognih skupkih.

Table 13: Number of protein-coding genes for each species and % of genes included in any of the orthologue groups.

The table contains only four out of all seven plants currently included in GoMapman, as the remaining three plant species (tobacco, sugarbeet and cacao tree) were not included in any of the implemented orthologue clusterings. All the plants contain over 60 % of all protein coding genes in at least one orthologue clustering, over 70 % for Arabidopsis and rice. Potato and tomato have slightly lower percentages of inclusion, as they are evolutionarily more distant and additionally are not included in all orthologue clusterings.

Rastlinska vrsta	Število proteinsko kodirajočih genov	Število proteinsko kodirajočih genov prisotnih vsaj v eni ortologni skupini
navadni repnjakovec	27416	23446 (86 %)
krompir	34732	22186 (64 %)
paradižnik	35609	23882 (67 %)
riž	39045	29183 (74 %)

Na podlagi filogenetske informacije vsebovane v različnih ortolognih nizih (PGSC_oMCL, ITAG_oMCL, ITAG_RSD) smo z uporabo analize omrežij na podlagi skupine 174,035 genov štirih rastlin vključenih v kateregakoli od nizov (navadni repnjakovec, krompir, paradižnik, riž) pridobili 73,701 šibkih komponent. Od tega je ena šibka komponenta zaradi eksperimentalnih (pridobitev zaporedij) ali pa računskih težav (bioinformatska orodja) med seboj povezovala preko 9,000 različnih genov in je nismo vključili v konsolidacijski postopek. Dodatno je 57,274 šibkih komponent vsebovalo le en gen in prav tako niso nadaljevali v postopek konsolidacije. Preostalih 16,426 šibkih komponent je pokrilo 107,683 genov, na katerih smo izvedli postopek konsolidacije, izboljšave položajev v ontologiji glede na informacijo cele ortologne skupine. Velikosti šibkih komponent, ki so vsebovale vsaj en gen so bile pri večini pod 10 genov, kot je razvidno iz histograma na Sliki 31.



Slika 31: Porazdelitev velikosti šibkih komponent pridobljenih iz treh ortolognih nizov.

Šibke komponente smo pripravili iz treh ortolognih nizov, in sicer PGSC_oMCL, ITAG_oMCL ter ITAG_RSD. Histogram prikazuje le komponente velikosti med 2 in 40 genov, saj je bila večina v tem območju. Od preostanka je 55,651 šibkih komponent vsebovalo le en gen in 381 več kot 40 genov (z največjo komponento velikosti 3132 genov).

Figure 31: Size distribution of weak components derived from three orthologous groupings.
The weak components were prepared using three orthologue types, PGSC_oMCL, ITAG_oMCL and ITAG_RSD. The histogram only includes size ranges between 2 and 40 genes per weak component, as the majority of the weak components are placed here. As for the remainder, 55,651 weak components only contain one gene and 381 contain more than 40 genes (the largest component containing 3132 genes).

4.3.1.4 Povezovanje z znanjem v drugih javnih bazah podatkov

Pomemben del vsebine baze je znanje vsebovano v različnih javno dostopnih razpršenih bioloških bazah podatkov, ki obsegajo opise funkcij genov, znanje o proteinskih strukturah in medsebojnih interakcijah, vključenost v procese ter povezavo z ostalimi ontologijami. S tem je uporabnikom aplikacije GoMapMan olajšan dostop do zunanjih virov, saj so našteti (če obstoje) v informacijski strani za vsak rastlinski gen ter ortologno skupino (Slika 32). Vsaka anotacija gena je oblikovana kot internetna povezava, kar olajša poizvedbo o nekem genu. Pomembno je omeniti, da se anotacij genov ali ortolognih skupin iz javnih virov ne validira interna ter da se končno odločitev o zanesljivosti anotacijskega vira ter anotacije same prepušča končnemu uporabniku.

Annotations

GO	GO:0001071 nucleic acid binding transcription factor activity GO:0003674 molecular_function GO:0003677 DNA binding GO:0003700 sequence-specific DNA binding transcription factor activity GO:0005575 cellular_component GO:0005622 intracellular GO:0005623 cell GO:0005634 nucleus GO:0006351 transcription, DNA-templated GO:0006355 regulation of transcription, DNA-templated GO:0008150 biological_process GO:0009058 biosynthetic process GO:0034641 cellular nitrogen compound metabolic process GO:0043226 organelle
PFAM	PF00847 AP2 domain
SMART	SM00380 DNA-binding domain in plant proteins such as APETALA2 and EREBPs
SPRINT	PR00367
Gene3D	G3DSA:3.30.730.10
PANTHER	PTHR31190 , PTHR31190-SF0 , PTHR31985
InterPro	IPR001471 AP2/ERF domain [domain] IPR016177 DNA-binding domain [domain]
PROSITE	PS51032 AP2/ERF domain profile.
SUPERFAMILY	SSF54171

Slika 32: Prikaz povezav na zunanje baze podatkov v aplikaciji GoMapMan.

Vsaka stran, ki vsebuje podrobnejše informacije o nekem genu je sestavljena iz večih delov, eden izmed njih je tudi del t.i. anotacij, ki so dejansko povezave na zunanje baze podatkov. Za nekatere so poleg samih internetnih povezav dodani tudi še krajši opisi, ki olajšajo pogled in zmanjšajo količino potrebnih naknadnih preusmeritev v brskalniku.

Figure 32: A display of connections to external databases as in GoMapMan application.

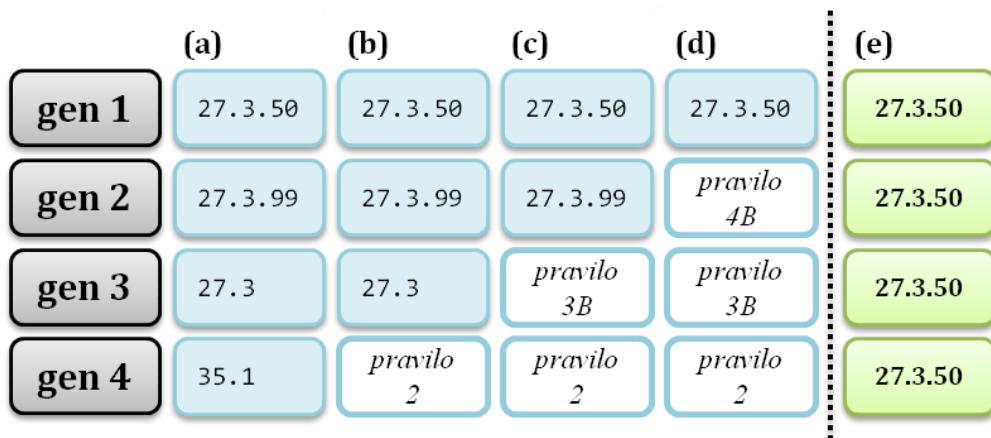
Each page dedicated to contain gene information (Gene Details View) is composed of several parts, one of them is the section of Annotations, which are acrually hyperlinks to external data sources. For some of these sources, short descriptions are given also, to ease the user's access and reduce the need for additional reroutes in the research of a particular gene.

4.3.2 Funkcionalnosti aplikacije GoMapMan

Vnos podatkov v podatkovno bazo GoMapMan je odvisen od tipa sprememb. Manjše spremembe, ki spremenjajo manjše tabele entitet (seznam rastlin, genskih modelov, ortolognih tipov, anotacijskih tipov, mikromrežnih tipov) se lahko spreminja le v ukaznem okolju. Večje spremembe (dodajanje, popravljanje, ipd.) v podatkih baze pa so implementirane v obliki modularnih paketnih skript. Tako kot so večje tabele ločene med vsebinske in povezujoče, so tako ločene tudi skripte za njihovo spreminjanje. Vse skripte imajo vnaprej določeno strukturo, kar olajša vodenje sprememb in pohitri vnos novih podatkov v bazo. Obenem jasna delitev nalog, ki jo vsaka skripta opravlja, ter njihova inherentna modularnost dovoljujeta razvoj bolj specializiranih funkcij v aplikaciji. Biološki ekspert izpolni vhodno datoteko s popravki gensko ontoloških anotacij, nato pa administrator sistema te popravke implementira. Možno je pa tudi ročno popravljanje anotacij genov neposredno v razvitem uporabniškem vmesniku.

4.3.2.1 Avtomatski popravki preko konsolidacije genskih funkcionalnih anotacij

Vključitev ortologne informacije omogoči hkraten vpogled v gene večih rastlinskih vrst (s čimer rešuje težavo asinhronega upravljanja s funkcionalnimi anotacijami) in omogoča uporabo avtomatskih konsolidacijskih skript za ontološke anotacije genov. Informacijo v ortolognih skupinah aplikacija GoMapMan uporablja kot glavni vir prevoda funkcionalne informacije med različnimi vrstami, s tako imenovano konsolidacijo pozicij genov med vrstami. Postopek konsolidacije je shematsko predstavljen na Sliki 33 na izmišljenem primeru ortologne skupine s štirimi geni in njihovimi začetnimi položaji ter končnim položajem skupine po zaključenem postopku. Primer na Sliki 34 pa prikazuje primer ene od ortolognih skupin med rastlinami krompirja, paradižnika, riža in navadnega repnjakovca.

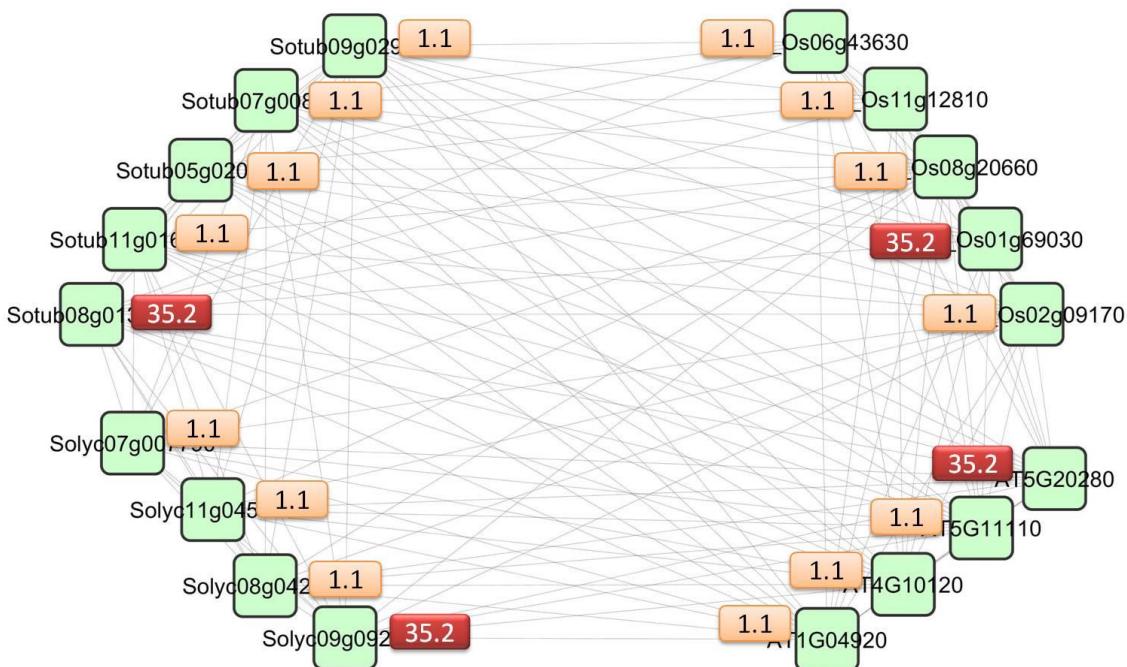


Slika 33: Ponazoritev postopka konsolidacije primera ortologne skupine.

Ortologno skupino predstavljajo širje geni (gen 1/2/3/4) označeni sivo. (a) Pred začetkom konsolidacije so vsi v različnih razdelkih, kjer se v naslednjem koraku (b) s pravilom 2 najprej izloči funkcionalna kategorija 35.1 (*brez ontološke anotacije*). (c) Zaradi hierarhičnega pravila 3 se potem izloči še kategorijo 27.3 (*regulacija transkripcije RNA*), saj sta v skupini prisotna dva bolj detajlni otroški kategoriji. (d) Ker je kategorija 27.3.50 bolj specifična kot manjvredna oblika 27.3.99 (*neklasificirana regulacija transkripcije RNA*), se odstrani še slednjo. (e) Končni rezultat konsolidacije predstavlja funkcionalna kategorija 27.3.50 (*splošna regulacija transkripcije RNA*), katerega se končno pripiše še vsem ostalim genom ortologne skupine.

Figure 33: A representation of the consolidation procedure for a fictitious orthologous group.

The orthologue group is composed of four genes marked grey. (a) Prior to consolidation start all the genes are assigned with different BINs, which are first reduced in the next step (b) with rule 2, where BIN 35.1.3 (*not assigned.no ontology assigned*) is removed. (c) Because of the hierarchical rule 3 BIN 27.3 (*RNA.regulation of transcription*) is further excluded, as the group also contains two more fine grained child BINs. (d) As BIN 27.3.50 is more specific than the lesser valued BIN 27.3.99 (*unclassified*), the latter is also removed. (e) The final consolidation result is represented by BIN 27.3.50 (*RNA.regulation of transcription.general transcription*), which is then assigned to all the other genes of the orthologous group also.



Slika 34: Primer genov v ortologni skupini in njihovih funkcionalnih anotacij.
Ortologna skupina v tem primeru vsebuje kombinacijo genov rastlin navadnega repnjakovca, paradižnika, krompirja in riža. Večina genov je anotirana z razdelkom MapMan, ki nakazuje funkcijo v svetlobno reakcijskem delu fotosinteze (funkcionalna kategorija 1.1). Štiri geni pa so označeni s statusom neznane funkcije (funkcionalna kategorija 35.2). Ideja za popravljanjem celotne ortologne skupine temelji na dejstvu, da je potrebno te štiri gene iz funkcionalne kategorije 35.2 premakniti v kategorijo 1.1.

Figure 34: An example of orthologous genes and their functional annotations.
The gene group for this example is a combination of *A. thaliana*, tomato, potato and rice genes. The majority of them are annotated with the MapMan BIN denoting a function in lightreaction part of photosynthesis (BIN 1.1). Four genes however are denoted as not having any function assigned (BIN 35.2). The idea behind curation of an orthologue group is that these four genes should all be correctly functionally annotated to BIN 1.1.

Primerjava stanja osnovnih izvornih datotek s stanjem po prvi konsolidaciji je podana v Preglednici 14 za izbrane izvorne in izvozne datoteke. Primerjava je bila narejena na tistih datotekah, ki so prisotne na obeh straneh (originalna internetna stran MapMan ter naše implementacije rastlinskih vrst) in samo na tistih genih/identifikatorjih, ki so bili skupni obema datotekama (presek). S postopkom konsolidacije smo izboljšali funkcionalne opise med 24 % in 70 % genom, v odvisnosti od rastlinske vrste. Največja izboljšava je opažena pri rastlini krompirja, kjer smo okoli 35 % enotam (genom oz. mikromrežnim sondam) pripisali funkcijo (prej so bili opisani s funkcionalno kategorijo 35) in okoli 30 % enotam pripisali bolj natančen opis funkcije (premik globlje v hierarhičnem drevesu ontologije). Na primeru tobaka smo izboljšali funkcionalni opis 55 % unigenov našega pripravljenega pan-transkriptoma, kjer je šlo kot pri krompirju v polovici primerov za

pripis novega funkcionalnega opisa in drugi polovici za pripis bolj natančnega opisa. Sledijo funkcionalni opisi navadnega repnjakovca in paradižnika z izboljšavami v približno 40 % genov, kjer pa se je razmerje med pripisom nove funkcije oz. izboljšanjem opisa funkcije premaknilo v smeri slednje. Z drugimi besedami, večim genom smo pripisali bolj natančen opis funkcije, kot pa pripisali novo. Najmanj sprememb funkcionalnih anotacij je v primeru riževega genoma (24 %), kar izvira iz samih uporabljenih ortolognih tipov in s tem slabe pokritosti riževih ortolognih genov – le en ortogni tip od treh je pri pripravi skupin vključil gene riža.

Preglednica 14: Izboljšava funkcionalnih opisov genov različnih rastlinskih vrst s pomočjo konsolidacije znanja v aplikaciji GoMapMan.

Prikazana je primerjava osnovnih izvornih datotek s stanjem po prvi konsolidaciji preko primerjav presekov genov med pred- ter po-konsolidacijskim stanjem položajev v ontologiji. Geni s spremenjenimi položaji v ontologiji se delijo na dva dela: geni, katerih anotacija je bila izboljšana iz razdelka 35 (neznana funkcija) ter gene katerih funkcija je bila bolj natančno določena (so bili hierarhično postavljeni globje v drevesu).

Table 14: Gene functional annotation improvement via knowledge consolidation for various species in GoMapMan.

Displayed is the comparison of the source export files with the status after the first consolidation through cross-section comparisons between pre- and post-consolidation gene position states. The genes with changed ontological positions separate into two sections: those whose annotation was improved from BIN 35 (unknown function) and genes whose function was specified more exactly (placed hierarchically deeper in the tree).

Rastlina	Vir	Presek	Razmerje sprememb	Nove anotacije (premik iz 35)	Izboljšava (hier. nadgradnja)
navadni repnjakovec	TAIR9	33,518	41 %	4,691 (14 %)	9,062 (27 %)
tobak	Agi4x44	34,734	55 %	10,235 (29 %)	8,818 (25 %)
riž	MSU_v7	55,801	24 %	4,908 (9 %)	8,601 (15 %)
paradižnik	ITAG2.3	34,727	44 %	4,665 (13 %)	10,711 (31 %)
krompir	Agi4x44	42,034	65 %	15,299 (36 %)	12,019 (29 %)
krompir	PGSC	39,031	69 %	14,876 (38 %)	12,056 (31 %)

4.3.2.2 Ročno popravljanje funkcionalnih opisov genov v ontologiji

V aplikaciji GoMapMan je implementirana sledljivost vnašanja sprememb, ki omogoča boljši nadzor verodostojnosti sprememb in jo lahko vidijo registrirani uporabniki (Slika 35).

Action	Starting Bin	End Bin	Description	Timestamp	Evidence Code	User
Insert	-	35.20	Adding Arabidopsis into 35.20.	2012-09-03	ND	auto
Insert	-	20.2.99	MapMan TAIR9 export.	2012-09-03	ISS	auto
Delete	35.20	-	Deleting from 35.20.	2012-09-03	ND	auto
Insert	-	20.1	Consolidation 13-12-21: severals (manual check)	2013-12-21	ISO	auto
Delete	20.1	-	rank4_wc00037_anno_stress	2014-09-16	IC	zivar
Move	20.2.99	20.1.7.1	PR1 based on protein domain classifications.	2014-09-16	IC	zivar

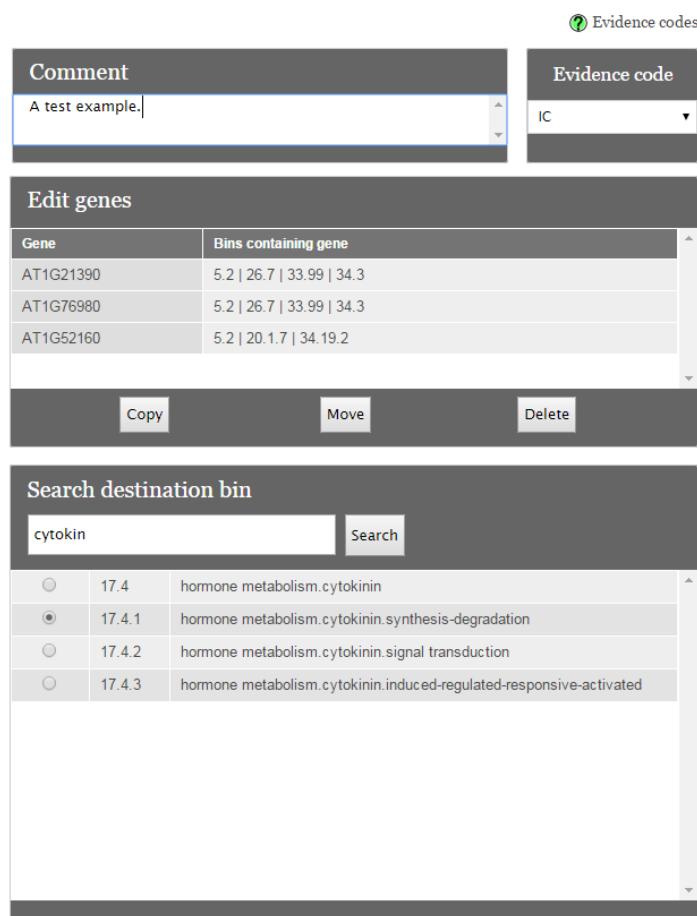
Slika 35: Informacija o zgodovini gena v ontologiji MapMan, kot jo vidi prijavljen uporabnik.

Prikazana informacija na sliki vsebuje naprej podatek o tipu modifikacije genskega položaja v ontologiji MapMan (možnosti: vnos, izbris, kopiranje, premik), o razdelkih MapMan in obrazložitvi modifikacije ter datum spremembe z evidenčno kodo, ki pove več o zanesljivosti ontološkega položaja. V zadnjem stolpcu je prisotna še informacija o uporabniku, ki je zadolžen za spremembo – 'auto' pomeni spremembe večjega obsega narejene z paketnimi skriptami, sicer pa se izpiše uporabniško ime.

Figure 35: Gene history information in the MapMan ontology, as is shown to a logged in user.

The figure displays information, first about the type of the gene position modification (options: insertion, deletion, copy or move), followed by MapMan BINs it was placed into and the reason of the modification, the date of the modification and evidence code, which informs the user of information reliability. The last column displays the user who performed the change, where 'auto' denotes batch scripts, otherwise the username is displayed.

Iz tehničnega vidika popravkov je preko grafičnega uporabniškega vmesnika v aplikaciji GoMapMan implementirana preprosta interaktivna rešitev (Slika 36). Možen je izbor večih genov hkrati, ki se jih nato glede na biološko znanje ali kopira v drugo funkcionalno kategorijo, premika med kategorijami oz. izbriše iz kategorije. Vsak tip modifikacije je pred samo izvedbo v ozadju tehnično preverjen in izведен, v primeru da ni težav (npr. poskus izbrisa gena iz razdelka, če je ta funkcionalna kategorija edina anotacija tega gena).



Slika 36: Vmesnik za vnašanje ročnih popravkov aplikacije GoMapMan.

Za vse izbrane gene mora uporabnik vnesti komentar, s katerim obrazloži spremembo, ter izbrati dokazno oznako iz seznama (vrh slike). Dodatno mora poiskati in izbrati končno funkcionalno kategorijo. Končno izbere še tip spremembe – kopija funkcionalnega opisa v kategorijo, premik v drugo kategorijo oz. izbris iz trenutno izbrane kategorije.

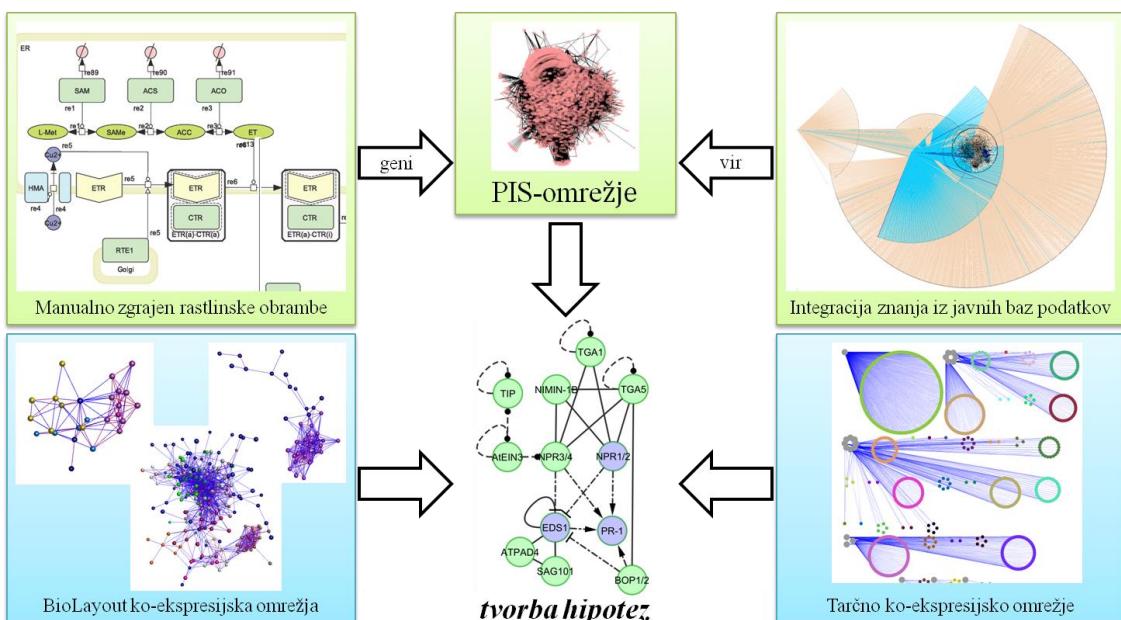
Figure 36: GoMapMan manual curation interface.

For all selected genes the user must insert a comment, describing the reasoning behind the change and select the evidence code from a drop-down list (figure top). Additionally, the destination BIN has to be searched for and selected. Lastly, the action is selected – copy/move from BIN to BIN or delete from the selected BIN.

V okviru skupine smo uredili izbrane skupine treh družin transkripcijskih faktorjev, ki vsebujejo domeno jasmonat-ZIM, domeno NAC ali domeno WRKY. Dodatno so bili popravljeni geni pripadajoč skupinama receptorjev ali PR-proteinov v biotskem stresu, v povezavi s popravkom ontologije (Poglavlje 4.3.1.1). V tem primeru se je pogosto zgodilo, da so bili receptorski proteini prisotni v razdelku, ki pa dejansko opisuje funkcijo PR-proteinov. Glede na opise genov smo velik delež lahko opisali s pravilnim funkcionalnim opisom, obenem pa smo tudi samo strukturo PR-proteinov bolj podrobno opisali in jih prenestili v prave podrazdelke.

4.4 INTEGRACIJA HETEROGENIH TIPOV PODATKOV

Predhodno zgrajen model rastlinske obrambe (Miljkovic in sod., 2012) smo v prvem delu (poglavlje 4.4.1) nadgradili in prenesli v orodje CellDesigner, iz literature in javnih baz podatkov zgradili veliko omrežje znanja o povezavah med geni navadnega repnjakovca ter na podlagi komponent modela obrambe izluščili podomrežje rastlinske imunske signalizacije (AtPIS) in ga prevedli na gene krompirja (StPIS). V drugem delu smo pripravili ko-ekspresijska omrežja (poglavlje 4.4.2), ki smo jih nato združili z znanjem v modelu StPIS (poglavlje 4.4.3). V zadnjem poglavju (4.4.4) pa sledijo rezultati analize pridobljenih omrežij z namenom generiranja novih hipotez o delovanju obrambnega odgovora rastlin. Kot shematski pregled pristopov služi Slika 37.



Slika 37: Koraki nadgradnje strukturnega modela rastlinske imunske signalizacije in njegove analize.

Razširjen in preveden model rastlinske obrambe je bil združen z zgrajenim velikim omrežjem znanja navadnega repnjakovca, iz česar smo izluščili omrežje rastlinske imunske signalizacije (PIS). Iz druge strani smo se poslužili dveh pristopov gradnje ko-ekspresijskih omrežij (modri ozadji). Različne kombinacije končnih omrežij so bile natančneje analizirane z različnimi pristopimi, s končnim ciljem tvorbe novih hipotez.

Figure 37: Steps taken to upgrade the structural plant immune signalling model and its analyses.
 Expanded and translated plant defense model was combined with the built large knowledge network for *A. thaliana* and used to extract the plant immune signalling (PIS) network. On the other hand, two approaches towards building co-expression networks were used (blue background). Different combinations of resulting networks were analysed in more detail using different approaches, with the final goal of forming new hypotheses.

4.4.1 Nadgradnja strukturnega modela rastlinske obrambe

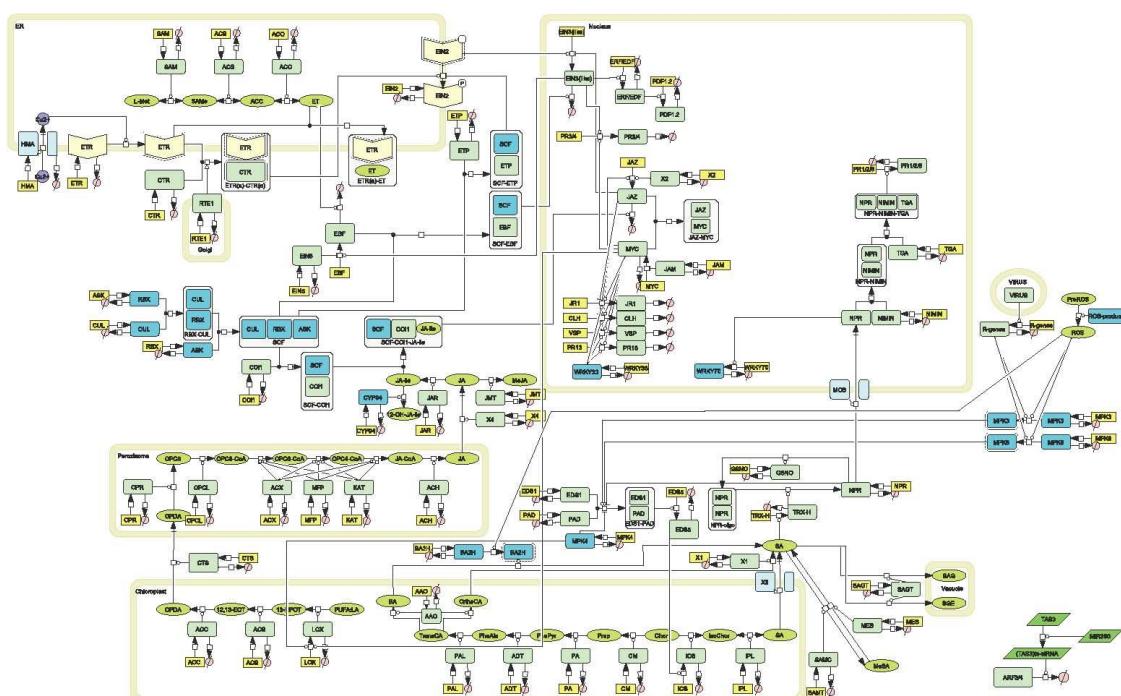
4.4.1.1 Priprava modela imunske signalizacije navadnega repnjakovca

Osnovni model (Miljkovic et al, 2012) je bil zgrajen v orodju CellIllustrator (Nagasaki in sod., 2010), katerega izvozna datoteka je pripravljena v formatu CSML (ang. *Cell System Markup Language*). Ker se danes kot standard bolj pogosto uporablja SBML (ang. *Systems Biology Markup Language*) in je avtomatska pretvorba iz med obema v orodju CellIllustrator nezanesljiva (vodi do precejšnih napak), smo model najprej prenesli v orodje CellDesigner (Funahashi in sod., 2003). Orodje CellDesigner podpira uporabo orodja COPASI (Hoops in sod., 2006) za dinamično modeliranje, medtem ko so rešitve za dinamično modeliranje v programu CellIllustrator slabo podprte.

Ob prevajanju modela iz formalizma CSML v formalizem SBML (Slika 38) smo natančneje definirali tip reakcije in s tem nadgradili ontologijo reakcij. Dodatno smo tudi razširili taksonomijo komponent modela. Vse komponente in reakcije so bile tudi opisane iz vidika pripadnosti različnim biološkim potem, npr. sintezi, signalizaciji ali skladiščenju rastlinskega hormona, dodatno definirane pa so t.i. komponente ali reakcije medsebojnega pogovora (splošne komponente rastline, npr. kinaze).

Osnovni model (2012) je vključeval 73 bioloških reakcij na nivoju družine, kar se razširi v 387 reakcij na nivoju genov. Z novim znanjem iz literature smo dodali 32 novih reakcij in ga tako povečali na 642 reakcij na nivoju genov (Priloga A). Od teh 105 definiranih reakcij, jih je 21 specifičnih za etilenske biološke poti, 36 za poti jasmonske kisline, 28 salicilne kisline, preostalih 20 pa sodeluje v medsebojnem pogovoru med rastlinskimi hormoni. Reakcije so razdeljene tudi glede na njihovo ontologijo, kjer je pri 13 reakcijah prišlo do vezave dveh komponent, pri 69 do aktivacije preko enega izmed definiranih mehanizmov, pri 16 do inhibicije. Preostalih 7 zaznamuje reakcije med virusnimi proteini z rastlino.

Na nivoju družine je osnovni model (2012) gradilo 79 komponent, kar se razširi v 185 komponent na nivoju posameznih genov (npr. družino S-adenozilmetionin sintetaz SAM v navadnem repnjakovcu sestavlja skupina štirih genov – SAM1, SAM2, MAT3 in MAT4). Trenutno najnovejša verzija modela vsebuje 101 komponent na nivoju družine (298 v razširjenem modelu), od katerih je 68 proteinov rastline navadnega repnjakovca, 31 metabolitov ter 2 pripadnika malih RNA (Priloga B). Od 95 definiranih družin komponent jih 20 sodeluje v etilen specifičnih reakcijah, 34 v reakcijah jasmonske kisline, 30 salicilne, preostalih 11 komponent pa reakcijah medsebojnega pogovora.



Slika 38: Nadgrajen model rastlinske imunske signalizacije.

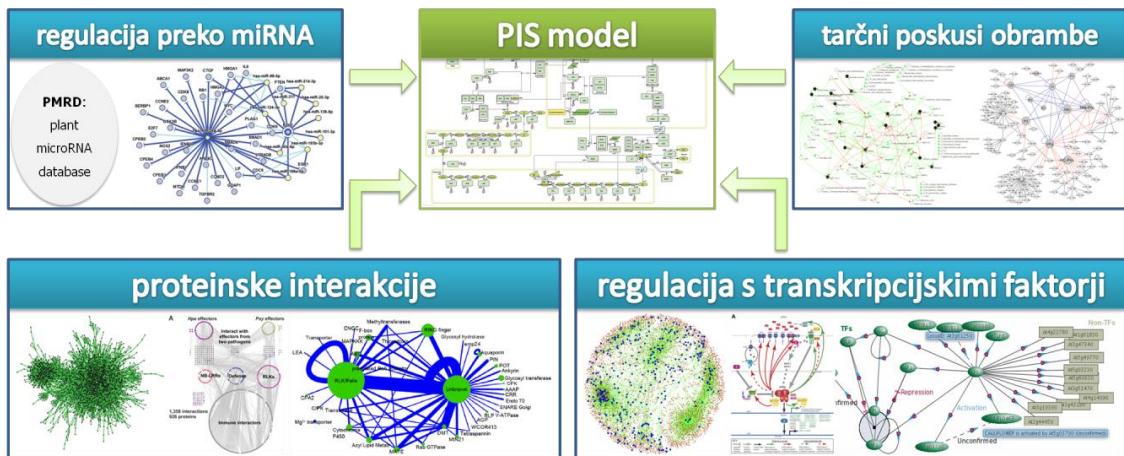
Orodje CellDesigner omogoča delitev v različne celične razdelke, ki so na sliki prikazani kot svetlo rumeni okvirji in predstavljajo jedro, kloroplast, peroksisom in endoplazmatski retikulum (ER). Vsaka proteinska točka (zelene barve za hormonsko specifične in modro za bolj splošne) ima korak sinteze (rumeni pravokotnik) in degradacije (krog svetlo rdeče barve). Nekatere bolj specifične biološke vloge so tudi vključene, bolj natančno transporterji (svetlo modre barve), membranski receptorji (svetlo rumena) in metaboliti (zelene ovalne oblike).

Figure 38: Upgraded plant immune signalling model.

CellDesigner allows separation into different compartments, which are shown as bright yellow brackets and represent the nucleus, chloroplast, peroxisome and endoplasmatic reticulum (ER). Each protein node (green for hormone specific nodes and blue for the general nodes) has a synthesis (yellow rectangle) and degradation (bright red circle) reaction. Some more specific biological roles are also included, namely transporters (light blue), membrane receptors (light yellow) and metabolites (green oval shapes).

4.4.1.2 Integracija javno dostopnega znanja

Ker smo želeli združiti čim več obstoječega znanja o delovanju rastline navadnega repnjakovca, smo za navadni repnjakovec v znanstvenih objavah poiskali vse eksperimente metodologij visokih zmogljivosti, ki so se lahko povezali z našim signalizacijskim omrežjem. Shema združevanja podatkov je predstavljena na Sliki 39.



Slika 39: Uporabljeni podatki za izgradnjo velikega omrežja znanja navadnega repnjakovca.
Podatki se razdelijo na pet delov: proteinske interakcije, omrežja transkripcijskih faktorjev, gensko regulacijo preko malih RNA, objave specifične za rastlinsko obrambo ter končen (ročno zgrajen in izboljšan) model rastlinske obrambe PIS (zelene barve).

Figure 39: The data used as building blocks for the large knowledge network for *A. thaliana*.
Data contains five sections, the protein-protein interactions, gene regulation (transcription factor networks and microRNA regulation), plant defence specific publications and lastly, the manual NIB built and improved plant defense model (green).

Veliko omrežje znanja navadnega repnjakovca je vsebovalo podatke ne-tarčnih proteinsko-proteinskih interakcij, genske regulacije, vključujuč regulacijo na nivoju transkripcijske regulacije in malih RNA, ter zbrane podatke iz modela rastlinske imunske signalizacije (Preglednica 15). Ker so nekatere od povezav pokrite v več kot samo enem podatkovnem tipu in so bile v okviru večjih skupin tipov podatkov (proteinskih interakcij, transkripcijskih faktorjev, malih RNA) odstranjene, smo namesto pričakovane velikosti omrežja s 63,593 povezavami dobili končno usmerjeno omrežje velikosti 53,880 medsebojnih povezav med geni – od tega 26,428 neusmerjenih, ki predstavljajo vezavo, ter 27,452 usmerjenih (regulacija preko transkripcijskih faktorjev ali malih RNA ter reakcije aktivacije, inhibicije in sinteze iz modela).

Preglednica 15: Pregled velikost manjših podomrežij, pridobljenih iz posameznih podatkovnih virov. PPI pomeni proteinske interakcije, TF transkripcijsko regulacijo in miRNA regulacijo preko malih RNA. V prvih dveh stolpcih je opisan vir podatka in njegova kvaliteta (izvor iz tarčnih poskusov, visoko zmogljivih študij, *in silico* napovedanih podatkov ali modela PIS). Stopnja odseva kvalitete podatkov, kjer 1 pomeni visoko kvalitetne podatke iz tarčnih poskusov iz literature; 2 rezultate visoko zmogljivih tehnologij; 3 *in silico* validirane napovedane povezave; 4 *in silico* nevalidirane napovedane povezave; 5 pa sintezne reakcije iz modela PIS. Sledita število unikatnih povezav v samem nizu in število ohranjenih povezav v velikem omrežju znanja navadnega repnjakovca (nekatere povezave so bile redundantne v okviru celotnega večjega podatkovnega tipa, npr. med vsemi skupinami regulacije s transkripcijskimi faktorji).

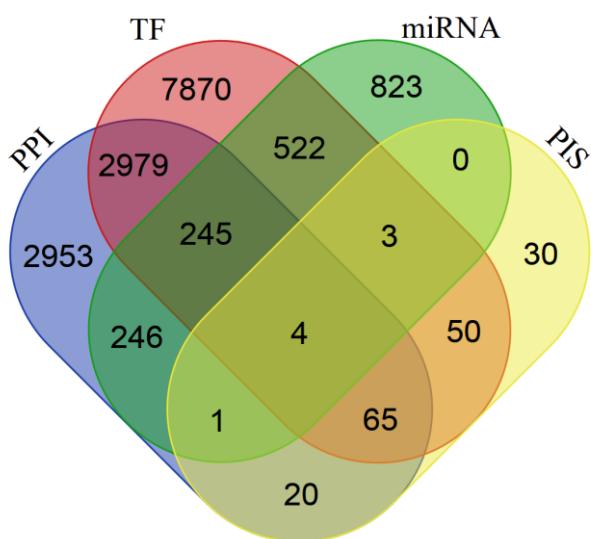
Table 15: Overview of small subnetwork sizes from different data sources.

PPI denotes protein-protein interactions, TF transcriptional regulation and miRNA micro-RNA regulation. The first two columns describe the data source and its quality (experimental data of targeted experiments, high-throughput studies data, *in silico* putative data or PIS model data). The reliability level reflects the data quality, where 1 denotes data of high quality from target experiments; 2 high-throughput technology data; 3 *in silico* validated predicted connections; 4 *in silico* unvalidated predicted connections; and 5 the synthesis connections from the PIS model. The number of unique connections within that set and the number of kept connections in the large knowledge network for *A. thaliana* follow (some connections were redundant within the large data type, e.g. between all transcription factor groups).

Tip podatka	Vir	Stopnja zaupanja	Število vseh povezav	Število ohranjenih povezav
PPI tarčni	atPIN	1	6637	5768
PPI tarčni	STRING	1	2709	2696
PPI visoko zmogljivi	Elena in Rodrigo	2	206	206
PPI visoko zmogljivi	athConsortium	2	11351	4737
PPI visoko zmogljivi	Jones	2	12102	12090
PPI visoko zmogljivi	Mukhtar	2	2795	858
PPI napoved	DeBodt	3	18674*	-
PPI napoved	atPIN	3	89766*	-
TF tarčni	atRegNet	1	823	801
TF visoko zmogljivi	atRegNet	2	10691	10684
TF visoko zmogljivi	ChIP-seq ET	2	1314	1305
TF napoved	Srivastava e2f	3	178	178
TF napoved	Srivastava ostali	4	12155	12040
miRNA tarčni	miRTarBase	1	101	94
miRNA visoko zmogljivi	PMRD	2	1889	1823
model PIS	aktivacija	1	204	190
model PIS	vezava	1	97	73
model PIS	inhibicija	1	132	128
model PIS	sinteza	5	209	209
			63593	53880

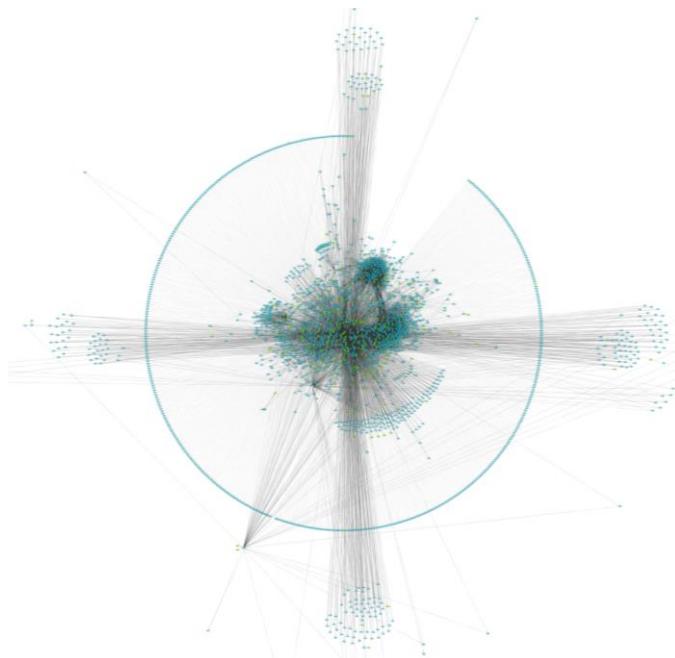
* Povezave niso bile vključene v končno veliko omrežje znanja navadnega repnjakovca.

Vsebovane reakcije so skupno zajele 15,811 unikatnih komponent, t.j. genov, malih RNA ter 11 virusnih komponent (Slika 40). Omrežje ni popolno, saj se število vseh genov rastline navadnega repnjakovca giblje v okolici 33 tisoč (od tega nekodirajočih genov okoli 5 tisoč). Manjkajoče komponente so v večini posledica eksperimentalnih metod pridobivanja teh povezav, ki še niso sposobne pokriti celotnega spektra.



Slika 40: Vennov diagram pokritja genov vsebovanih v velikem omrežju navadnega repnjakovca.
PPI pomeni proteinske interakcije, TF transkripcijsko regulacijo, miRNA regulacijo preko malih RNA, PIS pa komponente vsebovane v modelu PIS.

Figure 40: Venn diagram of gene coverage included in the large knowledge network for *A. thaliana*.
PPI denotes protein-protein interactions, TF transcriptional regulation, miRNA micro-RNA regulation and PIS the components of the PIS model.



Slika 41: Vizualizacija zgrajenega velikega omrežja znanja navadnega repnjakovca.
Predstavitev velikega omrežja znanja navadnega repnjakovca vsebujoč 53,880 skupnih povezav, rezultat uporabe organske ureditve točk (Tuikkala in sod., 2012), implementirane v orodju Cytoscape (Shannon in sod., 2003).

Figure 41: Visualisation of the large knowledge network for *A. thaliana*.
A representation of the full large knowledge network for *A. thaliana* containing 63,593 total connections, the result of using the organic layout (Tuikkala in sod., 2012) implemented in Cytoscape (Shannon in sod., 2003).

4.4.1.3 Priprava omrežij rastlinske imunske signalizacije

V prvem koraku prevajanja velikega omrežja znanja iz navadnega repnjakovca na krompir smo omrežje najprej skrčili na ortologne skupine, nato pa razširili na pripadajoče gene krompirja. Za prevajanje smo preizkusili šest različnih ortolognih tipov (Preglednica 16), ker ima vsak tip lahko različno pokritje prevajanja.

Preglednica 16: Pokritje genov navadnega repnjakovca in/ali krompirja glede na ortologni tip.

Drugi stolpec prikazuje število ortolognih skupin ortolognega tipa, ki vsebujejo vsaj en gen ali navadnega repnjakovca ali krompirja (in ne polno število vseh ortolognih skupin ortolognega tipa), sledеči stolpci pa prikazujejo število ortolognih skupin, ki vsebujejo gene obeh rastlin, gene le navadnega repnjakovca ali le krompirja.

Table 16: Properties of the six orthologue types in relation to *A. thaliana* and/or potato genes.

The second column shows the number of orthologue groups of the orthologue type that contain at least one *A. thaliana* or potato gene (this is not the full cluster count of the type), and the following columns list the number of orthologue groups containing genes of both plant species, only *A. thaliana* or only potato.

Ortologni tip	Skupno število ortolognih skupin	Skupine z obema rastlinama	Skupine z geni navadnega repnjakovca	Skupine z geni krompirja
ITAG_orthoMCL	20,440	11,475	1,615	7,350
ITAG_RSD	22,081	11,866	0	10,215
PGSC_orthoMCL	17,167	11,224	2,786	3,157
Šibke komponente	34,303	10,256	10,767	13,280
PLAZA-v3 (družine)	11,417	6,803	2,607	2,007
PLAZA-v3 (pod-družine)	38,783	10,201	11,907	16,675

Zaradi teh razlik med različnimi algoritmi in uporabljenimi rastlinskimi vrstami za pripravo ortolognega tipa, so tudi prevedena omrežja krompirja lahko precej različna (Preglednica 17). Zaradi evolucijskih razlik v razvoju družin rastlinskih genov preslikava med geni navadnega repnjakovca in krompirja ni ena na ena – en gen navadnega repnjakovca lahko niti nima ortologa v krompirju, družina pa se lahko tudi ali skrči ali razširi med obema vrstama. Razlike med razvojem genskih družin smo v prvem koraku pri prevajanju zanemarili in smo prevedena omrežja pustili na nivoju ortologne skupine, omrežja pa smo razširili oz. skrčili glede na dejanske gene krompirja v ortolognih skupinah šele kasneje (Preglednica 18).

Preglednica 17: Število prevedenih genov v velikem omrežju znanja iz navadnega repnjakovca na krompir je odvisno od izbranega ortolognega tipa.

Drugi stolpec prikazuje število vseh genov v ortolognih skupinah, ki vsebujejo vsaj en gen navadnega repnjakovca ali krompirja, sledeči stolpcji pa prikazujejo razporeditev teh genov – za vsako rastlino, ali so geni v ortologni skupini z drugo rastlino (stolpca 3 in 5) ali ne (stolpca 4 in 6).

Table 17: Amount of translated genes from the large knowledge network for *A. thaliana* to potato is dependent on the selected orthologue type.

The second column shows the number of all genes in the orthologue groups that contain at least one *A. thaliana* or potato gene, and the following columns display the distribution of these genes – for each plant, whether they are in an orthologue group with the other plant species (columns 3 and 5) or not (columns 4 and 6).

Ortologni tip	Skupna vsota genov	Geni		Geni	
		navadnega repnjakovca s prevodom	navadnega repnjakovca brez prevoda	krompirja s prevodom	krompirja brez prevoda
ITAG_orthoMCL	49,354	17,401	4,957	16,122	10,874
ITAG_RSD	34,314	11,939	0	12,017	10,358
PGSC_orthoMCL	71,438	17,304	5,929	28,676	19,529
Šibke komponente	76,443	19,695	13,021	26,932	16,795
PLAZA-v3 (družine)	62,501	23,280	4,136	31,331	3,754
PLAZA-v3 (pod-družine)	62,501	14,356	13,060	15,040	20,045

Preglednica 18: Velikosti omrežij znanja po abstrakciji na ortologne skupine in nato razširitvi na posamezne gene.

Od šestih virov prevoda na krompir so širje že vključeni v bazo GoMapMan, dodatno pa smo testirali še ortologe na nivoju družin in pod-družin skupine PLAZA. Tretji stolpec nakazuje velikost omrežij glede na število povezav za abstraktne ortologne skupine, četrти stolpec pa velikost omrežij, če vsako od ortolognih skupin razširimo na vse gene krompirja, ki jih vsebuje.

Table 18: Knowledge network sizes following the abstraction to orthologue groups and then the expansion to separate genes.

From the six potato translation possibilities, four are already included in GoMapMan, additionally added have been orthologues on the level of families and subfamilies of the PLAZA group. The third column shows the network sizes according to the number of connections for abstracted orthologue groups, and the fourth column shows the network sizes, when each of the orthologue groups is expanded to all contained potato genes.

Rastlina	Vir	Število povezav (nivo ortologa)	Število povezav (nivo gena)
navadni repnjakovec	Podatki	/	53,880
krompir	ITAG_orthoMCL	47,386	75,771
krompir	ITAG_RSD	53,800	51,572
krompir	PGSC_orthoMCL	47,055	92,543
krompir	Šibke komponente	44,442	393,121
krompir	PLAZA-v3 (družine)	31,962	4,641,954
krompir	PLAZA-v3 (pod-družine)	49,393	81,934

Od vseh preizkušenih ortolognih tipov smo se za prevajanje iz omrežja genov navadnega repnjakovca na krompir odločili za ortolome skupine PLAZA pod-družin. Pred tem smo veliko omrežje znanja označili na osnovi 181 genov iz nadgrajenega modela PIS (iz Poglavlja 4.4.1.1). Podomrežje rastlinske imunske signalizacije (AtPIS) nato dobimo tako, da se odločimo za mejo, do katere povezave obdržimo. Na tej podlagi dobimo lahko podomrežja različnih velikosti, npr. AtPIS-g0 (obdržane le ničte-ničte povezave) ali AtPIS-g1 (ničte-ničte, ničte-prve, prvo-ničte in prvo-prve povezave). V našem primeru smo se odločili za omrežje AtPIS-g0, katerega smo nato z uporabo PLAZA ortolognih skupin poddružin prevedli v omrežje StPIS-g0, kar je pomenilo velikosti 11,534 povezav za omrežje navadnega repnjakovca ter 11,764 za omrežje krompirja (Preglednica 19).

Preglednica 19: Pregled povezav rangiranih omrežij PIS glede na izvorni seznam iz modela rastlinske imunske signalizacije.

Povezave so razdelje glede na generacijske tipe povezav – stetja ne upoštevajo usmerjenosti povezave, tako da je 0th – 1st povezava lahko tudi 1st – 0th povezava v dejanskem omrežju.

Table 19: Overview of connections in ranked networks according to the seed list of the plant immune signalling model.

Connections are split into generation types – the counts disregard directionality of the connection, so in fact the 0th – 1st can also be a 1st – 0th connection in the actual network.

Tip povezave	Število povezav v omrežju navadnega repnjakovca	Število povezav v omrežju krompirja (PLAZA)
0th – 0th	845	361
0th – 1st	3,082	2,936
1st – 1st	7,607	8,467
1st – 2nd	25,204	23,283
2nd – 2nd	11,534	10,567
2nd – 3rd	4,735	3,344
3rd – 3rd	419	204
3rd – 4th	332	147
4th – 4th	19	10
4th – 5th	10	4
5th – 5th	1	0
Brez oznake	92	70

Od vseh 15,811 vozlišč velikega omrežja znanja navadnega repnjakovca jih je bilo od 181 izvornih 176 označenih kot gen ničte generacije – petih genov iz seznama ni bilo prisotnih v velikem omrežju znanja. Z iskanjem sosedov je nato 2,199 genov dobilo oznako prve, 10,192 druge, 2,825 tretje, 299 četrte in 10 pete generacije; preostalih 110 pa brez oznake generacije (podgrafi, ki niso bili del velike povezane komponente LLC).

Izbrane gene iz modela PIS (120) smo prevedli na 229 genov krompirja, od katerih smo lahko v velikem omrežju znanja krompirja (prevedenem preko PLAZA-v3 poddržin) 109 genov označili kot gene ničte generacije (zaradi nepopolnosti omrežja; Slika 40). Z iskanjem sosedov po enakem postopku kot pri navadnem repnjakovcu je tako 2,109 genov krompirja dobilo oznako prve, 8,709 druge, 2,182 tretje, 130 četrte in 4 pete generacije; preostalih 84 je bilo brez oznake generacije.

4.4.2 Priprava ko-eksprejskih omrežij

Tako AtPIS kot StPIS omrežju smo, kjer je bilo možno pripisali eksperimentalne podatke, za kar je morala ortologna skupina vsebovati vsaj en gen tako navadnega repnjakovca kot krompirja, ki se je prevedel vsaj v eno mikromrežno sondu POCI, obdržano po kvaliteti kontrole mikromrež. Za pripravo eksperimentalnih podatkov za uporabo v programu BioLayout smo se odločili za uporabo ortolognih skupin PLAZA-v3 poddržin dvokaličnic, saj je bilo tu pokritje prevajanja največje (Preglednica 20).

Preglednica 20: Število ortolognih skupin v rangiranih podomrežjih PIS s pripisanimi mikromrežnimi sondami.

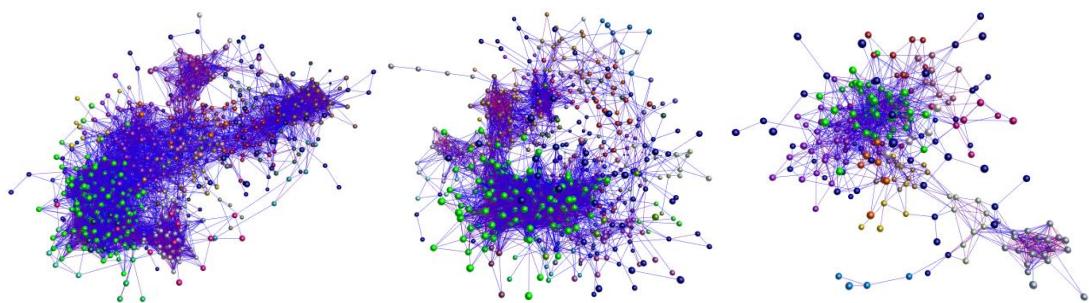
Vsote ortolognih skupin so enake skupnemu številu ortolognih skupin, ki vsebujejo vsaj en gen ali navadnega repnjakovca ali krompirja iz Preglednica 16. Razmerja in odstotki predstavljajo število teh genov ali ortolognih skupin, katerim smo lahko pripisali eksperimentalno mikromrežno sondu POCI.

Table 20: Number of genes with assigned microarray POCI probes.

The gene sums of the second column correspond to total *A. thaliana* genes from Table 17 and orthologue group sums of the third column correspond to total orthologue groups containing at least one *A. thaliana* or potato gene from Table 16. The ratios and percentages represent the number of these genes or orthologue groups that could be assigned with an experimental microarray POCI probe.

Ortologni tip	Ortologne skupine s pripisanimi sondami POCI
ITAG_orthoMCL	16,653 / 20,440 (81.5 %)
ITAG_RSD	19,725 / 22,081 (89.3 %)
PGSC_orthoMCL	12,941 / 17,167 (75.4 %)
Sibke komponente	12,163 / 34,303 (35.5 %)
PLAZA-v3 (družine)	7,443 / 11,417 (65.2 %)
PLAZA-v3 (pod-družine)	22,595 / 38,783 (58.3 %)

Po prevodu podomrežja StPIS-g1 velikosti 11,764 povezav in obsegajoč 2,218 genov krompirja na mikromrežne sonde POCI smo obržali le tiste povezave, kjer sta imeli obe točki pripisano sondo. Po odstranitvi vseh povezav, kjer katerikoli oz. oba interaktorja nista imela pripisane vsaj ene mikromrežne sonde, nam je ostalo 9,247 povezav pokrivajoč 1,989 genov, ki so se prepisali v 1,510 različnih mikromrežnih sond POCI. Tem mikromrežnim sondam smo pripisali eksperimentalne podatke časovne serije iz Poglavlja 4.1 ter jih uporabili v programu BioLayout pri treh različnih vrednostih Pearsonovega korelacijskega koeficiente (0,85/0,90/0,95), rezultati so prikazani na Sliki 42.



Slika 42: Tri omrežja programa BioLayout izhajajoč iz seznama sond omrežja StPIS-g1.

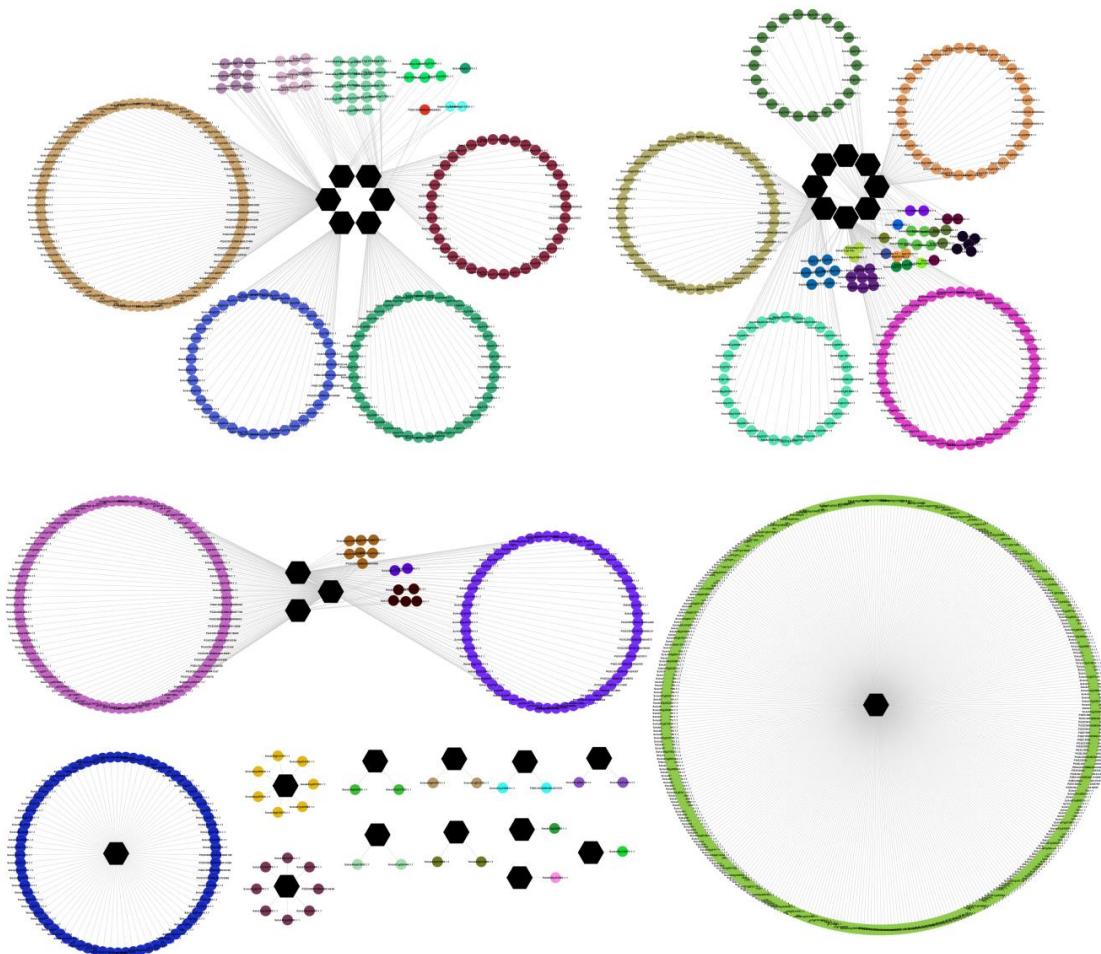
Od leve proti desni so predstavljeni rezultati BioLayout omrežij v odvisnosti od večanja prazne vrednosti Pearsonovega korelacijskega koeficiente (0,85/0,90/0,95). Različne barve točk v omrežjih predstavljajo skupine, ki so rezultat algoritma MCL, ki je vključen v program BioLayout, a jih mi nismo uporabili pri združevanju z ostalimi omrežji.

Figure 42: Three resulting BioLayout networks originating from the node list of the StPIS-g1 network.

From left to right the networks are a result of increasing Pearson correlation threshold (0,85/0,90/0,95) during the creation of the network. Different colours of the nodes represent clustering as was done by the MCL algorithm implemented in BioLayout, but were not used whilst combining with other networks.

Rezultat tarčne ko-ekspresijske analize z uporabo vab etilenske poti na celotnem nizu ekspresijskih podatkov (vsi genotipi) je predstavljalo šest različnih omrežij (dva seznama vab etilenske signalizacije in tri skupine eksperimentalnih podatkov). Za kasnejše analize smo se osredotočali na omrežje pridobljeno s kratkim seznamom etilena (samo zadnji korak sinteze etilena) na eksperimentalnih podatkih Désirée/NahG-Désirée. Končno omrežje velikosti 1,256 povezav je iz vidika komponent sestavljalo 34 vab ter dodatnih 1,143 genov krompirja s podobnimi vzorci izražanja eni ali večim vabam hkrati (Slika 43). V omrežju je bilo prisotnih 16 podgrafov, 13 takih, ki so se povezali le z eno izmed vab naenkrat, preostale tri pa so med seboj povezovale gene etilenskega odzivnega

faktorja (ERF) s (a) proteinski kinazami oz. (b) proteini SCF kompleksa, ki je odgovoren za katalizo razgradnje proteinov preko ubikvitinacije, (c) zadnji graf pa je povezoval proteinske kinaze z ATPazami odgovornimi za prenos ionov.



Slika 43: Končno tarčno ko-ekspresijsko omrežje izluščeno preko uporabe etilenskih genov kot vabe na eksperimentalnih podatkih mikromrež.

Vse črna vozlišča predstavljajo enega izmed 34 podanih genov kot vab, preostali geni pa so obarvani glede na to s katerim od tarčnih genov (enim ali več) si delijo podobnost genske ekspresije.

Figure 43: The resulting targeted co-expression network elucidated via usage of ethylene genes as baits on experimental microarray data.

All the black nodes represent one of the 34 given bait genes, while the remaining genes are coloured depending on which bait gene(s) they correlate to according to gene expression.

4.4.3 Združitev transkriptomskih podatkov z omrežjem PIS

Rezultat združevanja omrežja StPIS-g0 z rezultati ko-ekspresijske analize so bila tri končna omrežja, t.i. BTP (kombinacija BioLayout, tarčnega ter omrežja StPIS) pri treh vrednostih korelacije (BTP-0,85/0,90/0,95). Za vsakega smo pregledali njihovo medsebojno prekrivanje, torej koliko povezav se prekriva med vsemi tremi združenimi omrežji pri vsakem pragu Pearsonovega korelacijskega koeficienta (Preglednica 21).

Preglednica 21: Velikosti omrežij po združevanju velikega omrežja znanja z rezultati ko-ekspresijske analize.

Za unije treh omrežij smo izračunali razmerje pokritja za vsako povezavo – 3/3 pomeni prisotnost povezave v vseh treh omrežjih (BioLayout pri prazni vrednosti Pearsonovega korelacijskega koeficienta, tarčnem in StPIS-g0). Z večanjem pragu Pearsonovega korelacijskega koeficienta pri analizi BioLayout se velikosti omrežij temu primerno zmanjšajo, enako velja tudi za velikost presekov med omrežji.

Table 21: Network sizes after combining the large knowledge network with co-expression results.
We calculated the coverage of each connection – 3/3 means the occurrence of the same connection in all three networks (BioLayout at a particular Pearson correlation threshold, target network and StPIS-g0 network). Increasing the Pearson correlation coefficient threshold accordingly reduces the network sizes, the same holds true for intersections between the three originating networks.

Omrežje	Število vseh povezav	Pojavljanje povezav		
		3/3	2/3	1/3
BTP-0,85	133491	78	973	132440
BTP-0,90	105741	40	494	105207
BTP-0,95	87560	6	117	87437

Za primerjavo uspešnosti napovedovanja novih povezav smo kot »zlati standard« pripravili povezave iz modela PIS. Le-ta vključuje 105 reakcij na nivoju genske družine, od katerih smo v primerjavi zanemarili 51 reakcij. Od teh jih je bilo 29 posrednih oz. nepopolnih (reakcije za katere nimamo obeh genskih identifikatorjev v navadnem repnjakovcu ali vemo, da nam manjka vsaj en vmesni korak) ter 22 reakcij katalize, translokacije ali vezave med metabolitom in proteinom (saj naša omrežja niso vsebovala metabolitov). Ostalo nam je 11 reakcij proteinske (de)aktivacije, 3 reakcije degradacije, 29 reakcij translacije oz. transkripcije ter 11 reakcij vezave med proteini.

Gene družin preostalih 54 reakcij smo nato prevedli iz navadnega repnjakovca na PLAZA-v3 ortologne poddružine dvokaličnic in jih na tem nivoju tudi ohranili (izogib težavam različne evolucije genskih družin v obeh rastlinah). Obdržali smo le tiste povezave, kjer se obe komponenti lahko prevedeta na vsaj eno ortologno skupino, ki vsebuje tako gene navadnega repnjakovca kot krompirja. Končno število povezav v tako pripravljenem »zlatem standardu« je bilo vsebovalo 30 reakcij (Preglednica 22): 6 reakcij proteinske (de)aktivacije, 2 reakciji razkroja, 13 reakcij translacije oz. transkripcije ter 9 reakcij vezave med proteini.

Na ortologne skupine PLAZA-v3 poddružin dvokaličnic smo prevedli tudi veliko omrežje znanja navadnega repnjakovca, ko-ekspresijsko omrežje programa BioLayout ter tarčno ko-ekspresijsko omrežje pridobljeno s kratkim seznamom etilena. V velikem omrežju znanja smo obdržali proteinske interakcije in transkripcijske regulacije, ne pa tudi regulacije preko malih RNA (naših mikromrežni podatki niso zajeli te informacije). Tudi v prevedenih omrežjih smo obdržali le tiste povezave, kjer se obe komponenti lahko prevedeta na vsaj eno ortologno skupino, ki vsebuje tako gene navadnega repnjakovca kot krompirja. Tako zmanjšana omrežja smo primerjali s pripravljenim »zlatim standardom« 30 reakcij iz modela PIS (Preglednica 22).

Preglednica 22: Pokritje povezav preko različnih virov podatkov glede na pripravljen »zlati standard«.

Preostale povezave iz modela PIS (30) smo primerjali s povezavami v štirih zmanjšanih omrežjih (kjer sta obe komponenti predstavljali ortologne skupine z geni tako navadnega repnjakovca kot krompirja). Številke v glavi tabele predstavljajo število povezav v viru podatkov.

Table 22: Coverage of connections across different data sources as compared with golden standard.
 Remaining connections from the PIS model (30) were compared with four reduced networks (where both interactors represent orthologue groups which contain both *A. thaliana* and potato genes). Numbers in the header denote number of connections for each data type.

Tip reakcije	Opis reakcije	Omrežje PPI [10,344 p.]	Omrežje TF [9,565 p.]	Ko-ekspresija (BioLayout) [14,125 p.]	Ko-ekspresija (tarčna) [1,757 p.]
proteinska (de)akt.	HMA > ETR	-	-	-	-
proteinska (de)akt.	MPK3 > EDP	-	-	-	-
proteinska (de)akt.	MPK4 > EDP	-	-	-	-
proteinska (de)akt.	MPK6 > EDP	-	-	-	-
proteinska (de)akt.	NPR > EDP	-	-	+	-
proteinska (de)akt.	RTE1 > CTR	-	-	-	-
razkroj	COI1 > JAZ	+	-	-	-
razkroj	EBF > EIN3	+	+	+	+
transl./transkripcija	EDPA > EDS	-	-	-	-
transl./transkripcija	EIN3 > EBF	+	+	+	+
transl./transkripcija	EIN3 > ERF	-	+	-	-
transl./transkripcija	EIN3 > PR3/4	-	-	-	-
transl./transkripcija	EIN5 > EBF	-	-	+	+
transl./transkripcija	JAM > MYC	-	+	+	-
transl./transkripcija	JAZ > EIN3	-	+	-	-
transl./transkripcija	MPK4 > LOX	-	-	-	-
transl./transkripcija	MYC > JAZ	+	-	-	-
transl./transkripcija	MYC > LOX	-	+	-	-
transl./transkripcija	MYC > PR34	-	-	-	-
transl./transkripcija	NPR > WRKY	-	-	-	-
transl./transkripcija	TGA > PR125	-	+	-	-
vezava (proteinska)	ASK > COI1	+	-	-	-
vezava (proteinska)	ASK > EBF	+	-	-	-
vezava (proteinska)	CUL > ASK	+	-	-	-
vezava (proteinska)	EDS1 > PAD	+	-	-	-
vezava (proteinska)	ETR > CTR	+	-	-	-
vezava (proteinska)	JAZ > MYC	+	-	-	-
vezava (proteinska)	NPR > NPR	-	-	-	-
vezava (proteinska)	NPR > TGA	+	-	+	-
vezava (proteinska)	RBX > CUL	+	-	-	-

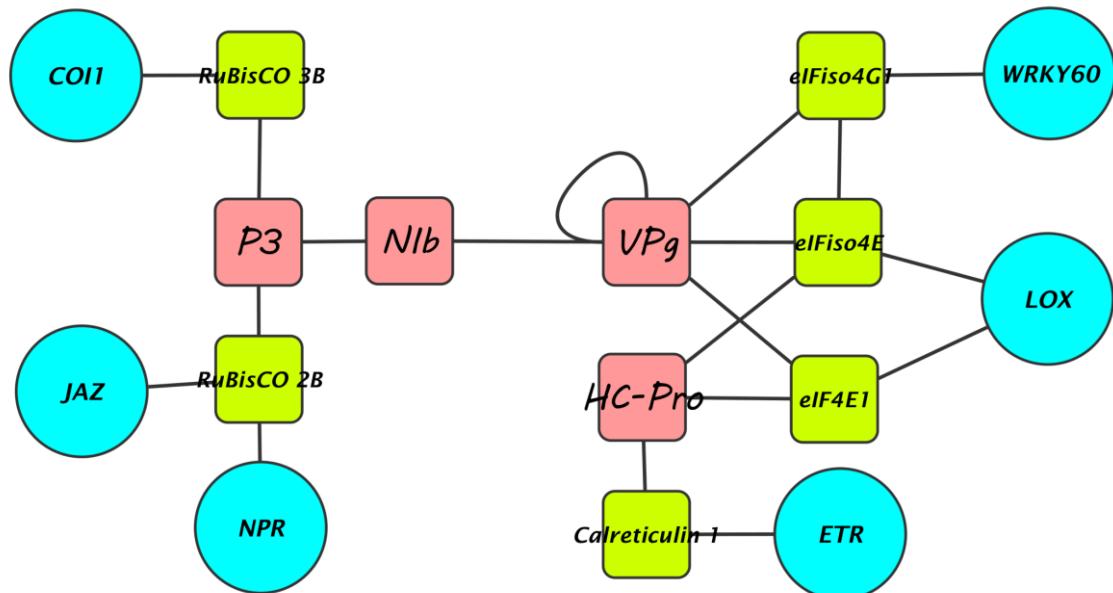
4.4.4 Analiza pripravljenih omrežij

V okviru analize pripravljenih omrežij smo si zadali več nalog:

- (1) iskanje novih povezav med komponentami v modelu rastlinske obrambe;
- (2) iskanje novih komponent, pomembnih za signalizacijo rastlinske obrambe;
- (3) iskanje novih poti, pomembnih za indukcijo obrambe.

4.4.4.1 Analiza vplivov virusnih proteinov na rastlinsko imunsko signalizacijo

Ker smo v omrežje PIS vključili informacijo o vezavah posameznih virusnih proteinov z geni navadnega repnjakovca (Elena in Rodrigo, 2012), smo lahko podrobnejše analizirali interakcije z že obstoječimi komponentami našega modela (Slika 44). Dodatno smo izboljšali resolucijo v modelu, saj je bil do vključitve te informacije virus predstavljen v abstraktni obliki (ni bil razširjen na vseh svojih 11-ih proteinov).



Slika 44: Interakcije med virusnimi in gostiteljivimi proteini.

Slika prikazuje povezave med virusnimi proteini (rdeci kvadrati) in proteini gostitelja, ki so ali lahko vključeni v nas model rastlinske obrambe (modri krogi) ali ne (zeleni kvadrati). Potyvirusi imajo sicer vsega skupaj 11 nastalih proteinov, ampak le štirje izbrani se povezujejo s komponentami našega modela.

Figure 44: Interactions between viral/host proteins.

The figure displays connections between viral proteins (red squares) and host proteins, which can either be included in our plant defense model (blue circles) or not (green squares). Even though Potyviruses have 11 total formed proteins, only the selected 4 have shown connections to any of our model components.

4.4.4.2 Izboljšava modela rastlinske obrambe

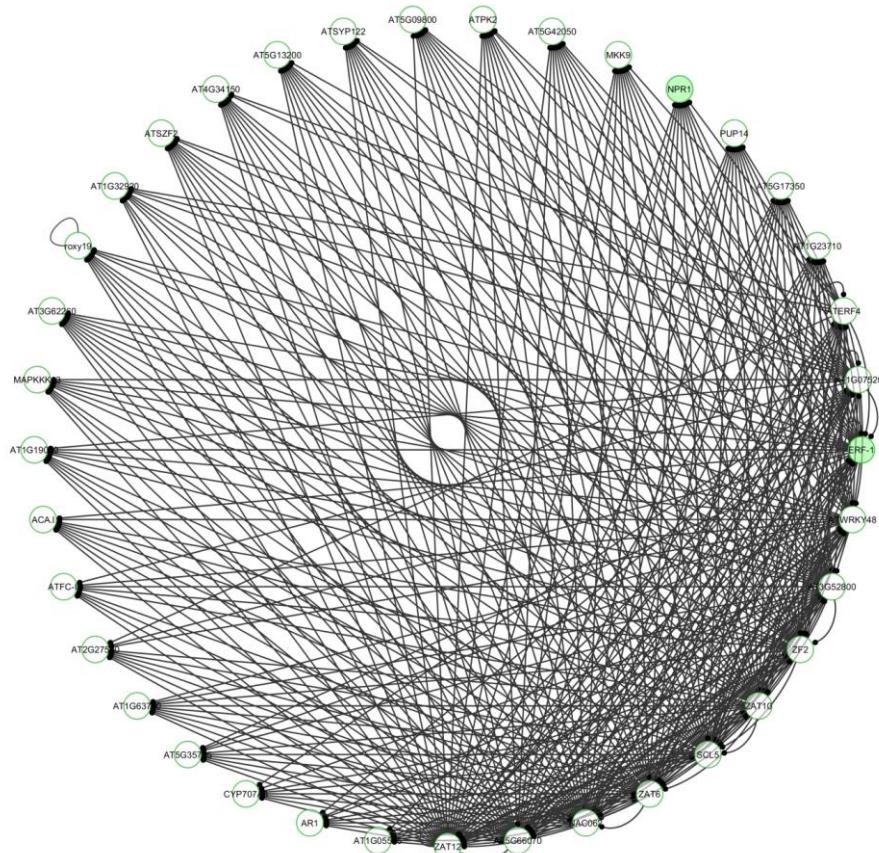
Obstoječi model rastlinske obrambe PIS navadnega repnjakovca (105 povezav) smo primerjali z omrežjem AtPIS-g0. Iz velikega omrežja znanja navadnega repnjakovca smo odstranili vse povezave iz modela PIS, kar je zmanjšalo njegovo velikost iz 53,880 na 53,280 povezav (glede na Preglednico 15). V tem zmanjšanem omrežju smo od 293 genov navadnega repnjakovca iz modela PIS lahko označili 243 genov z oznako ničte generacije (ostali v velikem omrežju niso bili prisotni). Za omrežje AtPIS-g0 obdržali le povezave med dvema genoma ničte generacije, t.j. 374 povezav na nivoju genov navadnega repnjakovca oz. 147 povezav AtPIS-g0 na nivoju genske družine.

V preseku med povezavami obstoječega modela rastlinske obrambe PIS navadnega repnjakovca ter povezavami omrežja AtPIS-g0 je bilo 19 reakcij (9 vezave med proteini, 6 aktivacijskih in 4 inhibicijskih). Od ostalih reakcij v obstoječem modelu PIS (86) za 51 nismo pričakovali vezave (29 posrednih oz. nepopolnih ter 22 reakcij katalize, translokacije ali vezave med metabolitom in proteinom), torej nam preostane še 35 reakcij modela PIS, ki niso v preseku s povezavami ostalih tehnologij.

Na nivoju družine smo iz omrežja AtPIS-g0 pridobili 128 potencialno novih povezav (izključujuč 19 povezav iz preseka) med že obstoječimi komponentami modela. Za njih bo pred vključitvijo potreben ročni pregled in potrditev glede na objave v literaturi oz. zasnovno novih poskusov, ki jih bodo potrdili. Nekatere od novih reakcij potekajo v okviru poti rastlinskega hormona (14 %), največji delež pa predstavljajo reakcije medsebojnega pogovora (ang. *cross-talk*) med potmi rastlinskih hormonov (40 %) oz. medsebojnega pogovora med rastlinskimi hormoni in splošnimi komponentami sistema (21 %). Zadnji del predstavlja motiv avtoregulacije, kjer protein vpliva sam nase (25 %).

4.4.4.3 Analiza biološke hipoteze

Nenadzorovane analize smo se lotili z uporabo algoritma MCODE (Bader in Hogue, 2003) implementiranega v orodju Cytoscape (Shannon in sod., 2003). Iz treh omrežij smo po uporabi algoritma dobili različna števila podomrežij, in sicer za vsako od BTP omrežij približno 60 (BTP-0,85 58 skupin; BTP-0,90 56 skupin; BTP-0,95 64 skupin). Nato smo iz velikega števila skupin izbrali nekatere najbolj zanimive, kot npr. skupino na Sliki 45. Ker je skupina MCODE nakazovala na povezavo med genoma ETR in NPR1 smo med njima tudi poiskali vse najkrajše poti ter vse sprehode do dolžine štirih korakov (Slika 46).

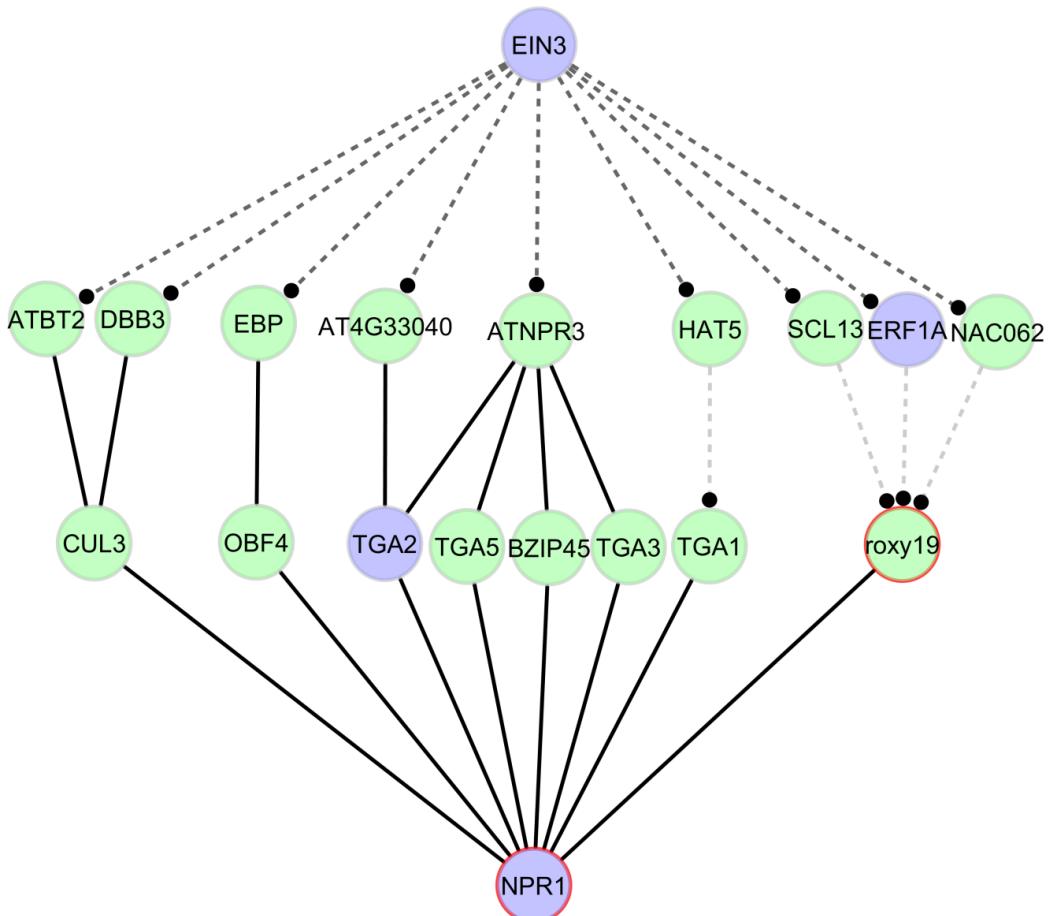


Slika 45: Izbrana skupina MCODE za nadaljne *in silico* analize.

Prikaz podomrežja MCODE prikazanega v orodju Cytoscape (Shannon in sod., 2003), kjer so vse točke omrežja urejene v krogu glede na njihovo število stopenj. V tej skupini je prisotna tudi povezava med genoma *erf1* in *npr1*, ki smo jo podrobnejše raziskali.

Figure 45: The selected MCODE cluster for further *in silico* analyses.

Subnetwork as created by MCODE, displayed in the Cytoscape tool (Shannon in sod., 2003), with a circle degree layout, where all the nodes are sorted according to their degree. This group contains a connection between the genes *ERF1* and *NPR1*, which we evaluated furhter.



Slika 46: Vse najkrajše poti med vozlišči *etr1* (pot etilena) in *npr1* (pot salicilne kisline).
Vijolične komponente so ene imed 16ih izbranih proteinov, ki smo jih uporabili v iskanju kratkih poti in sprehodov. Če je komponenta obkrožena z rdečo, pomeni nesposobnost prevajanja na sonde krompirja (v tem primeru smo se poslužili algoritma BLAST).

Figure 46: All shortest paths between *etr1* (ethylene) and *npr1* (salicylic acid) nodes.
Nodes coloured purple indicate belonging to the group of selected 16 target proteins used in short path and walk searching. If a node is coloured red, it implies that it could not be translated to potato via PLAZA-v3 orthologues (BLAST was used then to assign microarray probes).

Gen *npr1* navadnega repnjakovca (AT1G64280) je v aplikaciji GoMapMan na podlagi vključenih ortolognih tipov trenutno povezan s petimi ortolognimi skupinami (izključujuč naše pripravljene šibke komponente ter PLAZA-v3 družine/poddružine enokaličnic). Od teh dve ortogni skupini (ITAG_oMCL in ITAG_RSD) vsebujejo točno en gen krompirja (Sotub07g011600.1.1) ter ena družina PLAZA-v3 dvokaličnic še dva dodatna gena poleg izbranega. Prevod smo potrdili tudi z algoritmom BLASTp, kjer je bil ta gen najboljši zadelek gena navadnega repnjakovca proti bazi vseh genov krompirja (bitScore 1379, E-vrednost 10^{-152} , 50 % identičnih aminokislin, 68 % podobnih aminokislin, 4 % vrzeli).

Bližnji sorodniki gena v navadnem repnjakovcu vključujejo gene *npr3* (AT5G45110), *npr4* (AT4G19660) ter gen podoben genu *npr1* (AT4G26120). V krompirju pa poleg gena *npr1* (Sotub07g011600.1.1) najdemo še dva ostala proteina NPR (Sotub02g015550.1.1; Sotub07g016890.1.1) ter en ATP vezavni protein (Sotub07g016900.1.1). Na podlagi identifikatorja gena *npr1* krompirja smo v genomskem brskalniku SpudDB poiskali zaporedje pred začetkom njegovega odprtrega bralnega okvirja dolžine 2000 nukleotidov (Slika 47).

Slika 47: Zaporedje protomorske regije (2,000 nk) pred začetkom gena *npr1* iz genoma krompirja.
Figure 47: Sequence of a promotor region (2,000 bp) before the start of potato *npr1* gene.

Zaporedje smo analizirali v programih TransFac in PlantCare. Rezultati TransFac analize (Preglednica 23) so podali seznam večih vezavnih mest za transkripcijske faktorje udeležene predvsem v odgovor na vročinski stres ter seznam nekaterih ostalih družin. Iz druge strani so rezultati PlantCare programa (Preglednica 24) podali seznam možnih vezavnih mest predvsem za svetlobne odzive, pa tudi nekatere motive specifične za rastlinske hormone (abscizinsko kislino, etilen, giberelinsko ter jasmonsko kislino).

Preglednica 23: Rezultati TRANSFAC analize zaporedja promotorske regije gena *npr1*.

Table 23: TRANSFAC analysis results for the *npr1* gene promoter region sequence.

Gen navadnega repnjakovca (TAIR)	Tip transkripcijskega faktorja	Opis gena
AT1G04880	AT-bogat	high mobility group (HMG1/2) family protein
AT5G43170	C2H2	AZF3 (ARABIDOPSIS ZINC-FINGER PROTEIN 3)
AT3G20750	GATA	zinc finger (GATA type) family protein
AT3G51080	GATA	zinc finger (GATA type) family protein
AT1G13450	MYB	DNA binding protein GT-1
AT3G23250	MYB	MYB15 (MYB DOMAIN PROTEIN 15)
AT5G39610	NAC	ATNAC6 (NAC DOMAIN CONTAINING PROTEIN 6)
AT2G01930	transkripcija	BPC1 (BASIC PENTACYSTEINE1)
AT1G11570	vročinski stres	nuclear transport factor 2 (NTF2); putative
AT1G32330	vročinski stres	HEAT SHOCK TRANSCRIPTION FACTOR A1D
AT2G26150	vročinski stres	ATHSFA2# DNA binding / transcription factor
AT4G18880	vročinski stres	AT-HSFA4A# DNA binding / transcription factor

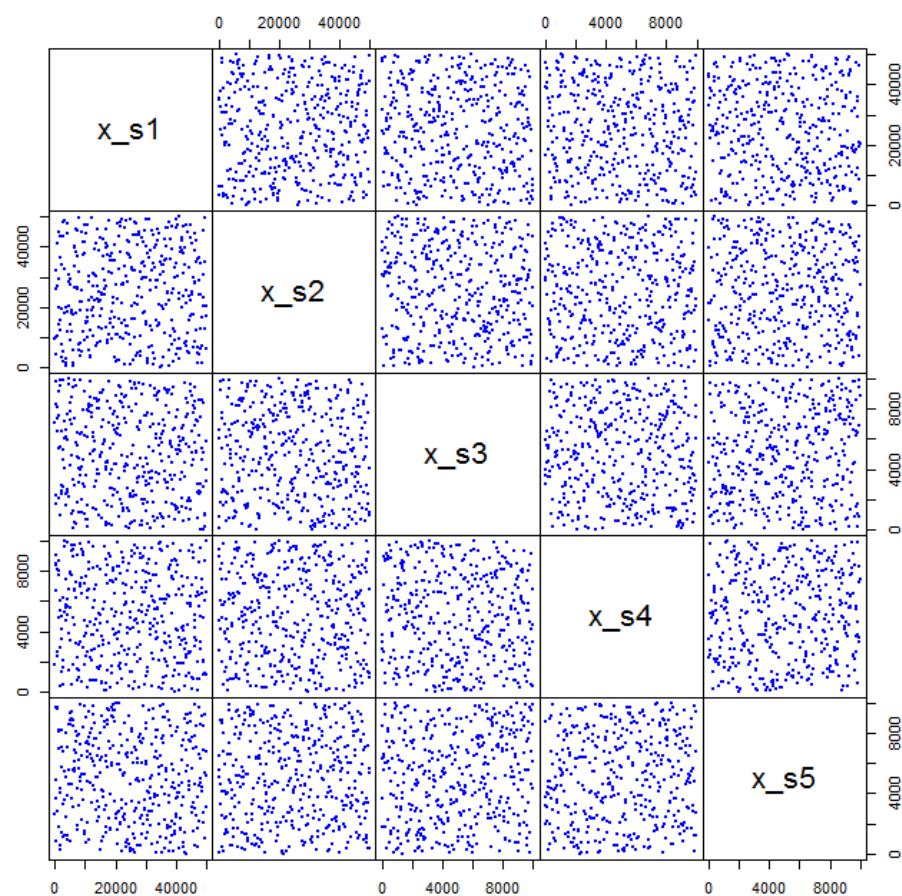
Preglednica 24: PlantCare analiza zaporedja promotorske regije gena *npr1*.

Table 24: PlantCare analysis results for the *npr1* gene promoter region sequence.

Ime motiva	Tip motiva	Opis motiva
ABRE	hormon (abscizinska k.)	cis-acting element in the abscisic acid responsiveness
WUN-motif	hormon (etilen)	wound-responsive element
P-box	hormon (giberelinska k.)	gibberellin-responsive element
CGTCA-motif	hormon (jasmonska k.)	cis-acting regulatory element in the MeJA-responsiveness
TGACG-motif	hormon (jasmonska k.)	cis-acting regulatory element in the MeJA-responsiveness
TC-rich repeats	obramba/stres	cis-acting element in defense/stress responsiveness
AE-box	odziv na svetlobo	part of a module for light response
AT1-motif	odziv na svetlobo	part of a light responsive module
Box 4	odziv na svetlobo	part of a conserved DNA module in light responsiveness
Box I	odziv na svetlobo	light responsive element
G-Box	odziv na svetlobo	cis-acting regulatory element in light responsiveness
G-box	odziv na svetlobo	cis-acting regulatory element in light responsiveness
GA-motif	odziv na svetlobo	part of a light responsive element
GATA-motif	odziv na svetlobo	part of a light responsive element
GT1-motif	odziv na svetlobo	light responsive element
LAMP-element	odziv na svetlobo	part of a light responsive element
Sp1	odziv na svetlobo	light responsive element
TCT-motif	odziv na svetlobo	part of a light responsive element
5UTR Py-rich	ostalo	cis-acting element conferring high transcription levels
ARE	ostalo	cis-acting regulatory element for anaerobic induction
CAAT-box	ostalo	common cis-acting element in promoter/enhancer regions
MBS	ostalo	MYB binding site involved in drought-inducibility
O2-site	ostalo	cis-acting regulatory element in zein metabolism regulation
Skn-1_motif	ostalo	cis-acting regulatory element for endosperm expression

4.5 RAZVOJ MATEMATIČNEGA MODELA RASTLINSKE OBRAMBE

Na podlagi začetnih eksperimentalnih vrednosti za 9 komponent in treh vrednosti določenih iz naše strani smo za spremembe v izražanju genov v kontrolnih vzorcih skozi čas zgradili latinsko hiperkocko (LH) velikosti 400 različnih kombinacij vrednosti, ki je pokrivala začetni prostor 33 neznanih parametrov zastavljenega modela. Slika 48 prikazuje razsevni diagram kombinacij za 5 izmed teh spremenljivk – tak pregled je potreben, da slučajno pri naključni porazdelitvi parametrov ne pride do pristranskosti.



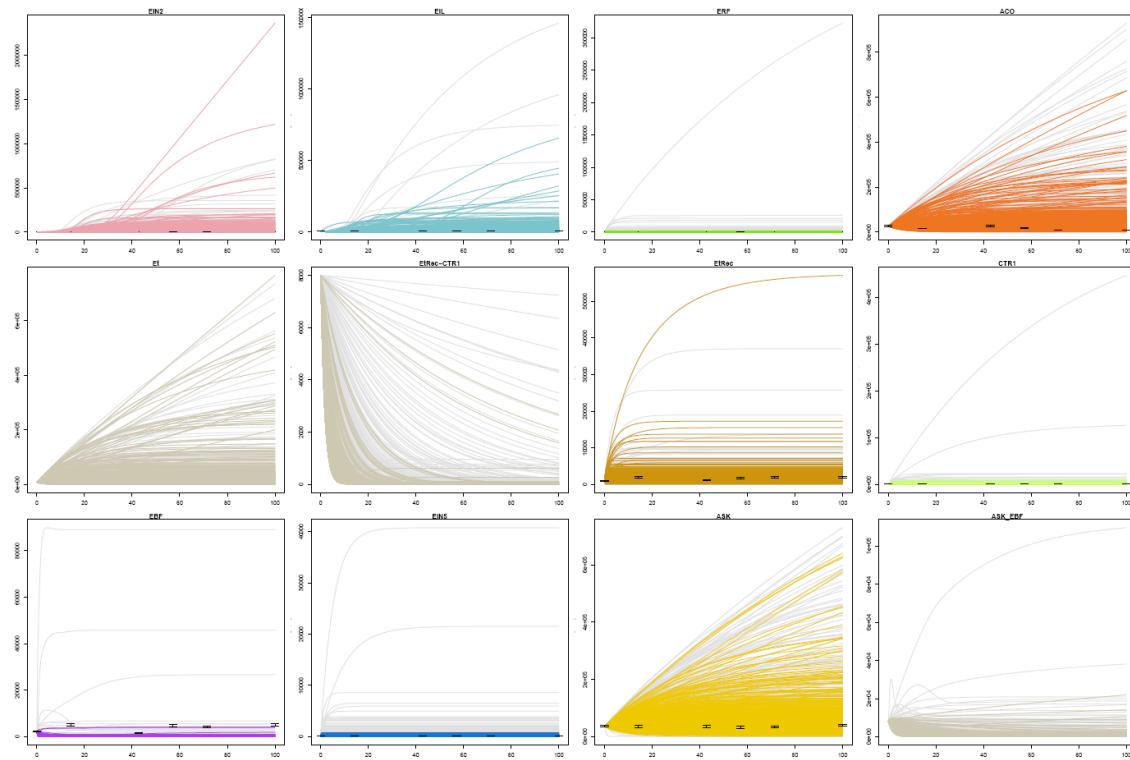
Slika 48: Porazdelitev parametrov v analizi sprememb izražanja genov kontrolnih vzorcev v zgrajeni latinski hiperkocki za pet parametrov sinteze.

Delna predstavitev LH zasnove za 400 različnih kombinacij parametrov v 33-dimenzionalnem prostoru, glede na razpon parametrov določenim na osnovi diferencialnih enačb.

Figure 48: Parameter distribution in the analysis of gene expression changes of control samples in the built LH for five synthesis parameters.

A partial representation of the Latin Hypercube design for 400 runs in 33-dimensional parameter space, given the parameter ranges defined on the basis of differential equations.

Te začetne kombinacije parametrov smo simulirali z uporabo *lsoda* algoritma iz paketa deSolve (Soetaert in sod., 2010) v programskem okolju R (v2.13.2), rezultati simulacij so na Sliki 49.



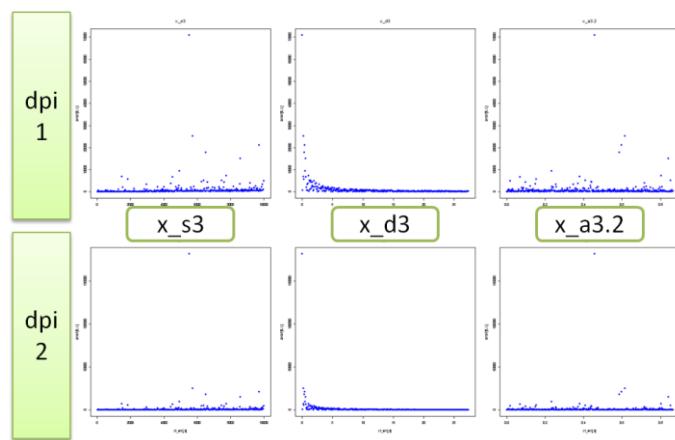
Slika 49: Rezultati numerične integracije za 12 komponent etilenskega podmodela za spremembe izražanja genov v kontrolnih listih.

Rezultati numerične integracije za vsako od dvajstih komponent modela, z uporabo 400 kombinacij parametrov pripravljenih v LH. Izrisane so tudi meritvene napake na podlagi eksperimentalnih vrednosti za 9 komponent.

Figure 49: Results of the numerical integration for 12 components of the ethylene submodel to see changes in gene expression in leaves of control samples.

Results of the numerical integration for each of the 12 model components, using the 400 parameter combinations generated with the LH. The error bars are given for the 9 chemicals that we actually measure.

Na podlagi simulacij lahko pripravimo linearne modele za vsako komponento, ki jo lahko merimo in časovno točko meritve (9×5). Od vseh 45 modelov izberemo tiste, ki najbolje pokrivajo naše eksperimentalne podatke (Slika 50).



Slika 50: Graf vrednosti rezultatov linearnih modelov glede na vrednosti parametrov za spremembe izražanja genov v kontrolnih listih.

Primer grafa za prvi in drugi dan po začetku eksperimenta, na osnovi katerega se odločimo (poleg statistične evaluacije) o pomembnosti nekega parametra za dani linearni model med komponento in časovno točko.

Figure 50: A plot of values of linear model outputs vs. parameter values to see changes in gene expression in leaves of control samples.

This is merely an example of plots for the first and second days after the experiment, on the basis of which we decide (apart from the statistical evaluation) the importance of a particular parameter for the given linear model of component/time point.

Izbrani linearni modeli predstavljajo emulator, ki ga uporabimo na 40,000 različnih kombinacij parametrov velike LH. Rezultati emulatorja nam podajo vrednosti, s katerimi lahko izračunamo neverjetnost vsake izmed 40,000 kombinacij parametrov in vsakega od opazovanih pogojev (45). Način izbora neverjetnosti za vsako kombinacijo parametrov oceni posameznik: kombinacije se lahko filtrira na podlagi prve največje neverjetnosti, vendar ima ta ponavadi velik vpliv negotovosti ocene linearnega modela. V našem primeru smo filtrirali po dveh kriterijih (Preglednica 25): prvi je upošteval vseh 45 opazovanih pogojev, drugi pa le prvo časovno točko (9 izračunanih neverjetnosti).

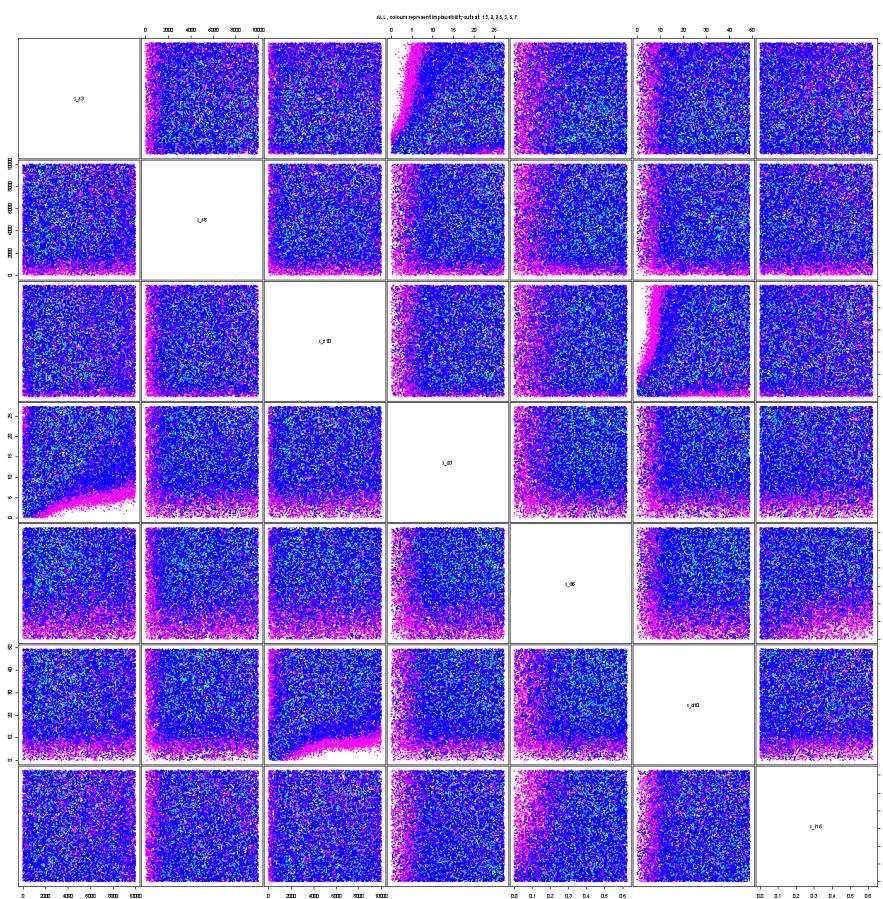
Preglednica 25: Zmanjšanje števila kombinacij parametrov v odvisnosti od izbora linearnih modelov. Prvi stolpec prikazuje izbrano prazno vrednost za izbor neverjetnih simulacij, ki je v našem primeru bila 3, 4 ali 5. Tриje sledeči stolpci prikazujejo število preostalih kombinacij parametrov v odvisnosti od prazne vrednosti, glede na kombinacijo vseh 45ih emulatorjev; zadnji trije stolpci pomenijo enako, le za kombinacijo devetih neverjetnosti prve izmerjene eksperimentalne časovne točke.

Table 25: The reduction in parameter combinations dependant on the linear model selection.

The first column displays the selected threshold value for selecting implausible runs, which can be 3, 4 or 5 in our case. The three columns that follow show the number of remaining parameter combinations (out of 40,000) when a particular threshold value was applied to the combination of all 45 emulators; the last three columns mean the same, only for the combination of improbabilities from the first measured experimental time point.

Pražna vrednost neverjetnosti	Izbor glede na vseh 45			Izbor glede na izbranih 9		
	Prva najvišja	Druga najvišja	Tretja najvišja	Prva najvišja	Druga najvišja	Tretja najvišja
3	2	6	18	985	5717	15576
4	62	226	469	4632	16538	29337
5	668	1594	2667	10458	26114	35980

Končno smo se odločili za ohranitev tistih kombinacij parametrov, kjer je bila najvišja neverjetnost nižja od prazne vrednosti 5, druga najvišja nižja od prazne vrednosti 3, ter tretja najvišja izračunana neverjetnost nižja od 2,5. S tem smo od 40,000 ohranili 3,728 kombinacij parametrov po prvem valu, kar pomeni v grobem 90 % zmanjšanje prostora parametrov glede na naše začetno določene meje (Slika 51).

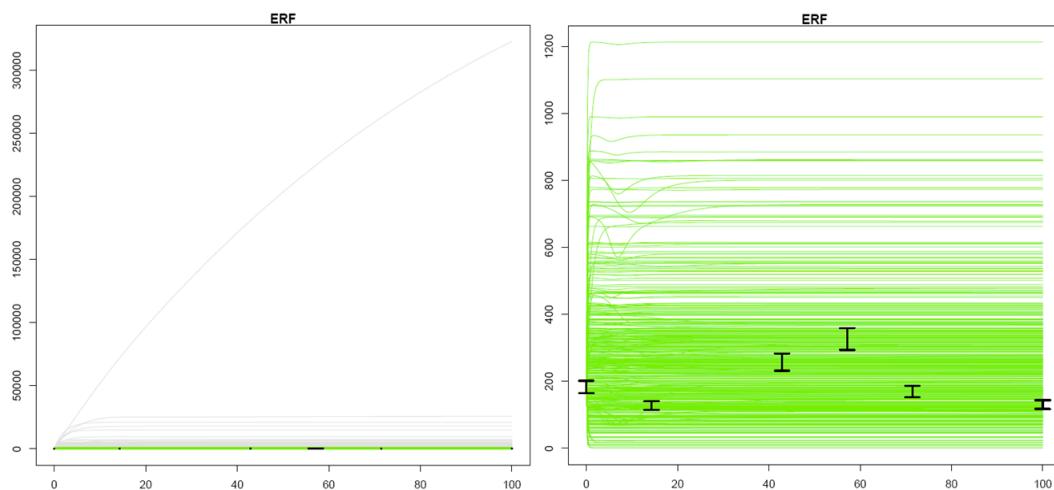


Slika 51: Rezultati zmanjšanja prostora parametrov po prvem valu Bayesove linearne metodologije. Prikazani so parametri sinteze in razkroja za tri komponente modela ter parameter inhibicije prve komponente. Vijolična barva prikazuje področja prostora parametrov, kjer je neverjetnost tako visoka, da je skrajno neverjetno kot rezultat dobiti dobro ujemanje med rezultati modela in eksperimentalnimi podatki.

Figure 51: Parameter space reduction results after the first wave of Bayes linear methodology. Displayed are the synthesis and degradation parameters for three model components and the inhibition parameter of the first chemical inhibition. The purple represents parameter space areas where implausibility is so high, where it is highly unlikely for us to obtain a good match between the model output and observed data.

Od preostalih kombinacij parametrov (3,728), za katere zaradi vrednosti njihovih neverjetnosti ne moremo trditi ali je kombinacija možna ali ne, spet izberemo podskupino 400ih in ponovno ponovimo korak numerične integracije (*lsoda*). To nas premakne v drugi (oz. naslednji) val metodologije, kjer se ponovijo enaki koraki: na podlagi 400 krivulj drugega valu zgradimo linearne modele za vsako komponento in časovno točko, izračunamo vrednosti neverjetnosti in izberemo ne-neverjetne kombinacije. S tem ponavljajočim se postopkom lahko nadaljujemo z zmanjševanjem prostora parametrov.

Primer zmanjšanja prostora parametrov je prikazan na Sliki 52 za komponento ERF, pri kateri smo z odstranitvijo neverjetnih kombinacij zmanjšali prostor do take mere, da smo se zelo dobro približali dejanskemu razponu vrednosti izražanja te komponente, nismo pa še pokrili dinamike izražanja, saj se bolj kompleksno obnašanje sistema pokaže šele v kasnejših valovih.



Slika 52: Približanje krivulj simulacij dejanskemu obnašanju molekule za ERF v drugem valu.
Slika prikazuje primerjavo med prvim (levo) in drugim (desno) valom Bayesove linearne metodologije, kjer smo se dobro približali pravemu razponu ERF komponente glede na eksperimentalne podatke. Nekatere od krivulj drugega vala tudi že kažejo bolj kompleksno obnašanje sistema (krivulje). Dodatno so na grafu prikazani tudi intervali napak glede na podatke, ki smo jih definirali kot $\pm 10\%$ od povprečja izmerjene točke komponente.

Figure 52: Simulation curves manage to capture the actual ERF behaviour better in the second wave.
The image displays the comparison between wave 1 (left) and wave 2 (right) curves, showing that we managed to zoom into the behaviour of ERF based on the experimental data. Some of the wave 2 curves are beginning to display the more complex system behaviour. Additionally, while plotting, we also plot the error bars, which we defined as $\pm 10\%$ of the mean of the measured time point of a component.

5 RAZPRAVA

Interakcije med gostiteljem in patogenom so zelo kompleksne in vodijo v velike spremembe na vseh nivojih biološke organizacije. Raziskovanje teh medsebojnih odnosov je velik izziv, katerega ni možno rešiti le z redukcionističnimi pristopi raziskovanja, ampak je nujno sistem proučevati celostno in pri tem uporabiti podatke visoko zmogljivih tehnologij generirane v zadnjih desetletjih. S tem se je razvilo relativno novo področje sistemsko biologije, ki preko integracije nebioloških disciplin (matematike, fizike, tehnike, statistike) predstavlja holističen način opazovanja sistemov.

V prvem delu doktorske naloge smo obdelali lastne podatke transkriptomskega poskusa, združili javno znana zaporedja v pan-genom krompirja ter pripravili in združili znanje razpršeno v različnih virih. Na podlagi tega smo pripravili veliko omrežje znanja v rastlini navadni repnjakovec in ko-ekspresijska omrežja, vzporedno pa razvili spletno aplikacijo GoMapMan. Ti rezultati prvega dela so nato v drugem delu omogočali analizo omrežij, nadgradnjo modela imunske signalizacije in vodili v tvorbo novih hipotez ter začetek dinamičnega modeliranja.

5.1 INTEGRACIJA PODATKOV ZAPOREDIJ GENOMOV OZ. TRANSKRIPTOMOV

V genetiki pan-genom (oz. supra-genom) opisuje komplement vseh genov v populaciji neke vrste (tipično se izraz uporablja v povezavi z bakterijami in arhejami) ali v našem primeru, vseh rastlinskih kultivarjev. Gradjo ga središčni genom (ang. *core genome*; geni prisotni v vseh rastlinskih kultivarjih), pogrešljivi genom (ang. *dispensable genome*; geni v dveh ali več kultivarjih) ter unikatni geni specifični za kultivar. V okviru doktorske naloge smo pripravili pan-genom krompirja z integracijo dveh genskih modelov krompirja (PGSC, ITAG) in dveh množic zaporedij unigene (POCI, StGI).

Združena zaporedja (Poglavlje 4.2.1, str. 77) omogočajo lažji pregled sekundarne informacije prisotne v teh skupinah podatkov, predvsem povezave z ontološkimi pojmi (GO, MapMan) ter eksperimentalno dokazane funkcionalne opise. Pomembna motivacija za pripravo genoma krompirja je, da, podobno kot Entrez (Maglott, 2004) služi za pretvorbo med eksperimentalnimi podatki sesalčjih vrst, služi za integracijo podatkov pridobljenih na različnih eksperimentalnih zasnovah. Dodatno, povezovanje pretvorniške tabele z ortolognimi skupinami različnih tipov omogoča lažji prenos znanja iz modelne rastline navadnega repnjakovca in ostalih kmetijsko pomembnih rastlin na krompir. Tabelo smo po vzoru orodja DAVID (Huang in sod., 2009) organizirali tudi v obliki aplikacije z bazo MySQL (interni dostop).

Veliko pač v pripravljanju pan-genomov ali pan-transkriptomov predstavlja uporaba samega algoritma BLAST. Ker le-ta deluje na podlagi lokalnih poravnav, lahko pride do velikega števila lažno pozitivnih zadetkov zaradi prekratkih poravnav med zaporedji. Velikega pomena je tudi izbor pravilnega algoritma BLAST, kjer je priporočena uporaba BLASTp (proteinsko proti proteinskemu) ter odsvetovana uporaba BLASTn (nukleotidno proti nukleotidnem zaporedju). Ker pri združevanju genskih modelov z množicama unigene na osnovi starejše tehnologije EST proteinska zaporedja niso vedno na voljo, predstavlja BLASTx primeren kompromis, saj prevede nukleotidna zaporedja v vseh šest bralnih okvirjev. Zaradi dejstva, da sta dve zaporedji lahko homologni, če je njuna proteinska enakost tudi le 25 % (za nukleotidna enakost pa ponavadi 70 %; Rost, 1999), tako s primerjavo proteinskih zaporedij močno povečamo robustnost in zanesljivost končnih rezultatov.

V našem primeru smo težave izhajajoč iz delovanja algoritma BLAST najprej rešili z zmanjšanjem redundancy med zaporedji z uporabo algoritma CD-HIT, nato pa še z učenjem odločitvenega drevesa v programu WEKA, ki je to veliko število lažno pozitivnih zadetkov (izhajajoč iz kratkih poravnav) algoritma BLAST primerno zmanjšal. Preizkusili smo tri algoritme, J48, ADTree in JRip, katerih natančnost na učni množici je bila pri vseh višja od 90 % (pri porazdelitvi razrednega atributa 30/70 za 'je ortolog'/'ni ortolog'). Ker je imel algoritom J48 imel najboljše razmerje med lažno pozitivnimi in lažno negativnimi odločitvami (Preglednica 10, str. 74), smo ga izbrali za filtriranje rezultatov.

V tem delu je primerno omeniti tudi podatke, s katerim smo pridobili učno množico za učenje odločitvenih dreves oz. pravil. Ker je priprava ortolognih skupin v bližnji povezavi s taksonomskim kontekstom (Chan in sod., 2013), smo za pripravo povezav naše učne množice uporabili dva tipa ortolognih skupin pridobljenih z algoritmom orthoMCL (Li in sod., 2003), eno s širšim filogenetskim razponom (12 rastlin) ter drugo z ožjim (5 rastlin). Rezultate algoritma RSD (Wall in Deluca, 2007) nismo uporabili, ker le-ta dejansko predstavlja najboljši zadetek algoritma BLAST v obe smeri primerjave zaporedij.

Tretji korak katerega smo se poslužili za manjšanje vpliva delovanja algoritma BLAST je bila priprava uteži za vsako od preostalih povezav (Slika 10, str. 41), s katerimi smo še dodatno zmanjšali vpliv lokalnih poravnav algoritma. Vrednost izračunane uteži je bila odvisna od dolžine poravnave dveh zaporedij (oz. pokritja) in vrednosti podobnosti zaporedij. Problem kratkih lokalnih poravnav algoritma BLAST bi lahko alternativno reševali tudi z izbiro globalnega algoritma poravnav, od najstarejšega algoritma globalnih poravnav med zaporedji Needleman-Wunsch (Needleman in Wunsch, 1970) do novejših (Chakraborty in Bandyopadhyay, 2013; Huang in Chao, 2003), pri čemer pa je potrebno upoštevati njihovo časovno zahtevnost.

Množico binarnih povezav smo na podlagi uteži nato združili v paralogne skupine v programu Pajek, s končnim številom 35,609 glavnih predstavnikov, kar je v okviru ocen števila genov krompirja in primerljivo npr. z genomom paradižnika (The Tomato Genome Consortium, 2012). Za primerjavo so Grimplet in sod. (2012) z uporabo le algoritma BLAST (in neupoštevanju omenjenih pomanjkljivosti) pri združevanju štirih genskih modelov in dveh množic zaporedij unigene za vinsko trto napovedali skoraj 60 tisoč genov.

Po začetku uporabe pripravljenih paralognih skupin je bilo pri večjih skupinah opaženo pojavljanje podskupin, ki so bile med seboj povezane z manjšim številom povezav in so pogosto predstavljalce ločene skupine dejansko paralognih genov. Popravljanje teh obstoječih paralogov bi ob učenju odločitvenih dreves lahko rešili preko uporabe drugih algoritmov priprave ortolognih skupin za učno množico. Ker uporabljeni algoritem orthoMCL deluje na osnovi algoritma BLAST ter spada v skupino metod, ki za inferenco ortologije uporablajo le podobnost zaporedij (Kuzniar in sod., 2008), lahko končne ortologne skupine vsebujejo tako ortologe kot izven-paraloge (ang. *out-parologue*, paralogni pari, ki so se razvili v vsaki rastlinski vrsti posebej, po podvojitvi gena). Poslužili bi se lahko algoritmov, ki za inferenco ortolognih razmerij vključujejo filogenetske evolucijske modele in delujejo na osnovi metod analize omrežij, npr. algoritmov RIO (Zmasek in Eddy, 2002) ali OrthoStrapper (Storm in Sonnhammer, 2002).

Drugo rešitev za ločevanje ortolognih in paralognih genov pa je dodaten razvoj analize omrežij našega postopka, kar bi lahko predstavljalo naslednji implementiran korak postopka – popravljanje vsake paralogne skupine z ozirom na nizko vmesnost vsake točke (ang. *betweenness centrality*), t.j. števila najkrajših poti med vsakim parom genov. V okviru doktorskega dela se take nadgradnje nismo lotili in za popravljanje izbranih paralognih skupin predlagali bolj časovno zahteven pristop filogenetske analize za posamezne bolj zanimive paralogne skupine (poravnava več zaporedij, priprava filogenetskega drevesa).

5.2 APLIKACIJA GOMAPMAN IN NJENE PREDNOSTI

Razvili smo spletno aplikacijo GoMapMan (Ramšak in sod., 2014), ki povezuje gene modelnih in agro-ekonomsko pomembnih rastlinskih vrst z ontologijo MapMan in ortologno funkcionalno informacijo v podatkovni bazi, s čimer omogoča lažje izboljševanje, povezovanje in vizualizacijo razumevanja in označevanja funkcij genov. Informacija vsebovana v bazi GoMapMan je rezultat kombinacije avtomatskih korakov pridobivanja genskih anotacij iz eksternih baz podatkov (razkropljenih po različnih virih), avtomatskih sprememb povezav genov z ontologijo (t.i. konsolidacija) preko informacije vsebovane v ortolognih skupinah ter ročnih popravkov teh povezav, ki so odvisne od ekspertov specifičnih področij (skrbnikov, ang. *curators*). S tem je olajšan prenos znanja o rastlinskih genih iz modelnih rastlin na poljščine. Baza vsebuje tudi povezave z mikromrežnimi identifikatorji ter raznovrstne druge javno dostopne tipe podatkov, ki dodatno opisujejo funkcijo genov, kar omogoča povezovanje z eksperimentalnimi podatkovnimi nizi ter povezovanje vsebovanega znanja z različnimi programi za interpretacijo podatkov visoko zmogljivih tehnologij (npr. Gene Set Enrichment Analysis GSEA ali MapMan) preko redno posodobljenih izvoznih datotek.

Zaradi fleksibilne zaslove ontologije MapMan je veliko različnih skupin asinhrono nadgrajevalo drevo svoje izbrane rastlinske vrste in bioloških poti. Dodatno je vrstno specifični razvoj ontološkega drevesa vodil v prisotnost nekaterih razdelkov samo pri nekaterih rastlinah, tako da uporaba ene same izvozne datoteke ne bi pokrila vseh procesov. Združitev in sinhronizacija ontoloških dreves med sabo je to težavo rešila, dodatno pa smo tudi prvič ontologijo definirali v strukturiranem OBO formatu (Smith in sod., 2007), kar bo olajšalo prihodnje nadgradnje ontologije.

Vnos podatkov v podatkovno bazo GoMapMan je odvisen od tipa sprememb. Manjše spremembe, ki spremenjajo manjše tabele entitet (seznam rastlin, genskih modelov, ortolognih tipov, anotacijskih tipov, mikromrežnih tipov) se lahko spreminja le v ukaznem okolju. Do sedaj je nadgradnja funkcionalnih opisov genov potekala ločeno za vsako rastlino in v odvisnosti od fokusa raziskovalne skupine, kar je vodilo do raznih protislovij in napak – s tem namenom smo razvili skupino po meri narejenih skript, katere se uporablja ročno (niso periodične) v ozadju aplikacije ob vsaki dodaji novih rastlinskih vrst ali novih ortolognih tipov. Vse skripte imajo vnaprej določeno strukturo, kar olajša

vodenje sprememb in pohitri vnos novih podatkov v bazo. Obenem jasna ločba nalog, ki jo vsaka skripta opravlja, ter njihova inherentna modularnost dovoljujeta razvoj bolj specializiranih funkcij v aplikaciji. Biološki ekspert izpolni vhodno datoteko s popravki gensko ontoloških anotacij, nato pa administrator sistema te popravke implementira. Možno je pa tudi ročno popravljanje anotacij genov neposredno v razvitem uporabniškem vmesniku.

Baumgartner in sod. (2007) so nakazali na nujnost vključitve metod tekstovnega rudarjenja v popravljanje funkcionalnih opisov, saj ročni popravki niso dovolj hitri v primerjavi s hitrostjo pridobivanja novih podatkov v današnji znanosti. Kot alternativo tekstovnemu rudarjenju smo razvili avtomatiziran postopek t.i. konsolidacije (Slika 13, str. 49), v katerem izrabljamo dejstvo, da imajo evolucijsko sorodni geni pogosto ohranljeno funkcijo v različnih organizmih (Du Plessis in sod., 2011). Postopek uporablja ortologno informacijo med večimi rastlinami hkrati in tako omogoča prenos znanja med njimi, posebej iz modelne rastline navadnega repnjakovca na ostale kmetijsko pomembne rastlinske vrste (Slika 34, str. 89).

Implementacija podatkov transkriptoma in mikromrežnih zasnov dodatno pripomore pri lažji biološki interpretaciji podatkov, saj vmesen korak pretvorbe ni potreben za končnega uporabnika. Dodatno nam povezava med geni in mikromrežnimi sondami omogoča, da za slednje ni potrebno ob vsaki nadgradnji na novo generirati funkcionalnih anotacij, saj MapMan opise podedujejo kar od starševskega gena. Tudi iz vidika vnosa novih povezav za že obstoječe rastline je proces lahek, dokler obstaja povezovalna tabela med geni in transkripti oz. geni in mikromrežnimi sondami (ki jo lahko pripravi raziskovalec sam z uporabo algoritma BLAST). Poleg uporabe tega znanja v programih analize (npr. GSEA ali MapMan) pa se take povezave lahko uporablja tudi v orodjih za iskanje novih povezav preko povezovanja ontološke informacije ter že znanih povezav z eksperimentalnimi podatki.

Ročno popravljanje genskih funkcionalnih anotacij predstavlja največji izziv zaradi časovne potratnosti postopka, ampak je na dolgi rok zelo uspešno kar se tiče doprinosa h kvaliteti analize podatkov. S tem namenom smo v aplikaciji GoMapMan razvili tudi vmesnik (Slika 36, str. 92), ki omogoča popravke v manjšem merilu, a je omejen na avtorizirane in o postopku podučene znanstvenike. Predlagani protokol ročnih popravkov vključuje pregled znane informacije o nekem genu, znane proteinske domene (npr. PFAM ali InterPro), vključenosti v biološke procese (npr. Gene Ontology ali KEGG) ter literature za sam gen. Veliko informacije se pridobi tudi s primerjavo z ostalimi rastlinskimi vrstami, še posebej modelno rastlino navadni repnjakovec, kjer je zanesljivost in gostota informacij pogosto dovoljšna za osnutek odločitve o funkciji gena. V primeru, da je gen v neki ortologni skupini, lahko zaključke izpeljemo iz cele ortologne skupine, zaradi predpostavke, da opravlajo vsi geni pripadajoči neki ortologni skupini enako funkcijo. Tako naj bi bili vsi geni opisani z enako skupino ontoloških vozlišč MapMan, na enak način deluje tudi avtomatski protokol konsolidacije. Končno lahko uporabimo tudi primerjavo zaporedij preko npr. algoritma BLAST, v skrajnem primeru pa se genu dodeli status neznane funkcije (funkcionalna kategorija 35.2).

5.3 INTEGRACIJA JAVNO DOSTOPNEGA IN EKSPERIMENTALNEGA ZNANJA

Priprava velikega omrežja znanja in s tem povezano zbiranje podatkov literature (Preglednica 15, str. 97) je predstavljala enega najbolj časovno zahtevnih delov doktorskega dela, saj je bilo potrebno vsak vir posebej obdelati pred vključitvijo v razumljivo celoto. Tudi izbor podatkov je za kvaliteto zelo pomemben: zaradi težav s katerimi se srečujejo algoritmi določanja vezavnih mest transkripcijskih faktorjev smo se odločili odstraniti vse računalniške napovedane povezave. Dodatno je bila potrebno filtriranje glede na dejansko identifikacijo gena kot transkripcijski faktor. Zapletlo se je tudi pri prevodu kvalitativnih modelov obrambe (Naseem in sod., 2012) na gene navadnega repnjakovca, kjer smo razvili posebno skripto, ki na podlagi seznama genov obeh genskih družin povezave pripravi vse kombinacije med njimi. A ker v primeru take ekspanzije lahko hitro pride do lažnih povezav, saj je osnovna povezava podana na nivoju družine in ne specifičnega gena, se je pri interpretaciji te težave potrebno zavedati.

Ob prevodu omrežja na gene krompirja smo uporabili ortologno informacijo aplikacije PLAZA (Proost in sod., 2014), in sicer pod-družin rastlin dvokaličnic (Preglednica 19, str. 101). V tem koraku lahko pride do težav z neprevajanjem, ko gen navadnega repnjakovca ni v ortologni skupini z genom krompirjem, kar povzroča izgubo resolucije. Če želimo omrežje ortolognih skupin razširiti na vse gene krompirja, je velikost omrežja krompirja zelo velika (kombinatorni izbruh povezav), kar lahko zmanjšamo s pripisom reprezentativnega gena krompirja vsaki ortologni skupini. Izbor genov, ki v krompirju opravljam enako funkcijo je prav tako zelo pomemben, saj ni nujno, da so vsi geni z evolucijo sploh še ohranili svoje funkcije (Weßling in sod., 2014).

Vključitev ko-ekspresijskih omrežij dodatno poveča končno velikost celotnega omrežja, zaradi lažno pozitivnih povezav, pa tudi zaradi slabega prekrivanja rezultatov različnih algoritmov inference gensko regulatornih omrežij. Prekrivanje med različnimi algoritmi inference in modelnim omrežjem iz javnih baz ni bilo veliko (Preglednica 21, str. 105), kar pa ni indikator slabe priprave podatkov, ampak bolj odraz težavnosti inference same ter razlik med samim delovanjem algoritmov (Windram in sod., 2014). Dodatno je potrebno upoštevati dejstvo, da opravljamo z nepopolnimi, a komplementarnimi, podatki, saj veliko omrežje znanja navadnega repnjakovca pokriva le okoli 50 % genske zasedbe rastline (Slika 40, str. 98). Tudi pri uporabi in analizi ko-ekspresijskih mrež se je potrebno zavedati, da korelacija vzorcev izražanja genov ne pomeni vzročnosti in tako podobni vzorci izražanja bolj nakazujejo, da si ti geni le delijo podobne regulatorje njihovega izražanja. Tem opozorilom navkljub, je napovedna vrednost povezav iz analize rezultatov različnih visoko zmogljivih tehnologij pomembne za potrjevanje pravilnosti postopka integracije in se njihovo pomembnost za opazovani sistem ocenjuje na podlagi pogostosti pojava in rezultatov ročnega pregleda literature.

Za medsebojno primerjavo prispevkov vsake od glavnih komponent končnih združenih omrežij znanja smo pripravili »zlati standard« povezav iz modela rastlinske obrambe PIS. V njem smo zadržali 30 reakcij med geni navadnega repnjakovca s prevodi v genomu krompirja glede na ortologne skupine PLAZA-v3 (Poglavlje 4.4.4.2, str. 109) od skupno 105 (izključili smo nepopolne ali posredne reakcije, reakcije katalize, translokacije ter vezave med metabolitom in proteinom). S tako pripravljenimi povezavami smo nato primerjali rezultate združenega velikega omrežja znanja navadnega repnjakovca ter oba rezultata ko-ekspresijskih mrež, za grobo oceno napovedne vrednosti integracije.

Surove vrednosti števila ujemajočih se povezav (Preglednica 22, str. 107) kažejo na najvišje ujemanje podomrežja PPI velikega omrežja znanja navadnega repnjakovca (12 ujemajočih povezav), nato omrežje transkripcijskih faktorjev prav tako iz velikega omrežja (7 povezav), medtem ko imata ko-ekspresijski omrežji 5 (BioLayout) oz. 3 (tarčni pristop) ujemajoče povezave. Ob upoštevanju velikosti samih omrežij povezav pa se najbolj izkaže tarčno ko-ekspresijsko omrežje, saj je vsebovalo le 1,757 povezav (1,7%), medtem ko so velikosti ostalih treh omrežij precej večje (okoli 10 tisoč; 1,1% za podomrežje PPI, 0,7% za podomrežje transkripcijskih faktorjev ter 0,4% za ko-ekspresijska omrežja programa BioLayout). Kljub temu povezave iz rezultatov ko-ekspresijskih omrežij niso zanemarljive, saj je porazdelitev ujemajočih se povezav enakomerno porazdeljena in se povezave vsakega od štirih omrežij povezujejo z različnimi reakcijami pripravljenega standarda.

Dodatno smo nadgrajeni model rastlinske imunske signalizacije PIS primerjali z podomrežjem AtPIS-g0 iz velikega podomrežja znanja (Poglavlje 4.4.4.2, str. 109). V preseku med obema je bilo 19 reakcij, kar na nivoju reakcij modela PIS pomeni 18 % celote, če pa dodatno zanemarimo še posredne in neposredne reakcije, reakcije katalize ali vezave med metabolitom in proteinom (ki jih naši zbrani podatki niso zajeli) pa 35 %. Te so posledica ali samih postopkov visoko zmogljivih tehnologij in s tem omejenega obsega podatkov v velikem omrežju znanja (le-to npr. ni vključevalo metod zaznave kompleksov ali vsebovalo omejeno število izmerjenih membranskih proteinov), kar se bo lahko nadgradilo v prihodnosti ob obstoju primernih eksperimentalnih podatkov.

Rezultati integracije podatkov različnih virov so tako potrdili nekatere od eksperimentlnih povezav modela rastlinske imunske signalizacije, dodatno pa smo našli nove povezave, ki v literaturi še niso znane oz. še niso eksperimentalno potrjene s tarčnimi poskusi. Tako smo od preostalih povezav v omrežju AtPIS-g0, ki niso bile v preseku z reakcijami nadgrajenega modela PIS pridobili 128 potencialno novih povezav med geni vključenimi v proces obrambe na okužbo, od katerih največji delež predstavlja reakcije medsebojnega pogovora, kar je v skladu z razvojem razvejenega omrežja povezav med signalizacijskimi potmi rastlinskih hormonov (Koornneef in Pieterse, 2008).

Ob primerjavi komponent našega modela rastlinske imunske signalizacije s publikacijo Elena in Rodrigo (2012) smo bili zmožni predvideti kako virus manipulira z imunsko signalizacijo rastline (Slika 44, str. 108). VPg ima veliko celičnih in viralnih vezavnih partnerjev, vključno z evkariontski iniciacijski faktor 4E (eIF4E) in njegove tri izoforme, HC-Pro, NIb ter samega sebe (Ivanov in sod., 2014). Vezava VPg na te iniciacijske faktorje je eden izmed možnih načinov s katerim virus izrablja translacijske komponente rastlinske celice (Quenouille in sod., 2013), saj so le-ti pripomorejo pri translaciji virusne RNA (Mäkinen in Hafrén, 2014; Moury in sod., 2013; Moury in sod., 2014).

Niti en od enajstih proteinov PVY glavnih komponent našega modela rastlinske obrambe ne veže direktno, marveč šele v drugem koraku (Slika 44, str. 108). S tem virus posredno vpliva na specifične hormonske poti salicilne (preko ključnega regulatornega gena *npr*) in jasmonske kisline (vpliv na sintezo preko gena lipoksigenaze *lox*, na signalizacijo pa preko gena COI1 ter JAZ TF). Ena od posrednih vezav vključuje protein kalretikulin, ki veže Ca^{2+} ione, slednji pa vpliva na etilenski receptor *etr*. Za HC-Pro papaje je bila dokazana tudi vezava na Ca^{2+} vezavni protein kalretikulin (Shen in sod., 2010). Veliko je tudi interakcij s faktorji začetka translacije ter fotosinteznimi podenotami (predvsem encimom RuBisCo), kjer so faktorji v skladu z teorijo virusne ugrabitve gostiteljevega sistema (Ivanov in sod., 2014), vpliv virusne okužbe pa je bilo prav tako že nakazan (Stare in sod., 2015).

5.4 TVORBA HIPOTEZ IN DINAMIČNO MODELIRANJE

Zaradi velikosti pripravljenih omrežij je sama analiza težavna brez uporabe metod, ki so sposobne na osnovi izbranih kriterijev razdeliti omrežje na manjše in bolj obvladljive enote. Tu smo se poslužili metod tvorjenja podomrežij na podlagi topoloških lastnosti, specifično na osnovi vrednosti medsebojne povezanosti točk, kar uporablja algoritem MCODE (Bader in Hogue, 2003). Nastala podomrežja smo pregledali in vključili transkriptomske podatke oddelka za vse tri genotipe. Obstoj visoko povezanih signalizacijskih poti različnih hormonov v obrambnem odgovoru rastline je dobro poznan, najbolj pogosto se izpostavlja antagonizem med jasmonsko in salicilno kislino (Koornneef in Pieterse, 2008; Van der Does in sod., 2013), vse bolj pa se kaže tudi pomen avksinov, abscizinske kisline, citokininov ter brassinosteroidov (Robert-Seilaniantz in sod., 2011). Etilen se pogosto omenja v povezavi s signalizacijsko potjo jasmonske kisline (Clarke in sod., 2000), medtem ko direktni vpliv etilenskih komponent na salicilno kislino ni posebej izpostavljan. Veliko novo najdenih povezav nakazuje medsebojne vplive v okviru ene hormonske poti, najdenih pa je bilo tudi precej povezav med različnimi hormoni, kar nakazuje na visoko povezano mrežo odgovorov, ki vključuje vse znane rastlinske hormone (Koornneef in Pieterse, 2008).

Ker direktni vpliv etilenskih komponent na salicilno kislino v znanstveni literaturi ni posebej izpostavljan, smo iz večjega razpona podomrežij algoritma MCODE za razvoj in testiranje biološke hipoteze izbrali tistega, ki je nakazoval na možno povezavo med potmi etilena ter salicilne kisline (Slika 45, str. 110). Dodatno smo za povezavo med genoma *ein3* (del etilenske poti) ter *npr1* (del salicilne poti) poiskali vse najkrajše poti (dolžine 4; Slika 46, str. 111). Dalje smo na odseku 2,000 nukleotidov pred začetkom odprtrega bralnega okvirja gena *npr1* opravili analizo promotorjev v dveh orodjih (Preglednica 23, str. 113; Preglednica 24, str. 113), od katere je možnost povezave ET in SA poti nakazoval le t.i. WUN-motiv, z biološkim opisom odgovora na poškodbe. Končno smo v laboratoriju zasnovali poskus, kjer se bo po tretiranju rastline z etilenom izmerilo vrednosti ekspresije gena *npr1* krompirja ter na tak način ob primerjavi tretiranih rastlin z netretiranimi potrdilo oz. ovrglo zastavljeno hipotezo. V vsakem primeru bomo rezultat lahko prenesli nazaj na osnovni model, s tem pa zaključili krog sistemsko biološkega raziskovanja – preko strukturnega modela bomo razjasnili pojme rastlinskega imunskega signalizacijskega omrežja.

Etilensko podomrežje iz modela rastlinske imunske signalizacije smo pripravili tudi za potrebe dinamičnega modeliranja. Za to smo pripravili skupino dvanajstih navadnih diferencialnih enačb (Slika 19, str. 63). Tem smo nato na podlagi transkriptomskih podatkov in ob uporabi Bayesove metode določanja (ene izmed metod ocenjevanja neznanih parametrov modela) zmanjšali parametrski prostor (Slika 51, str. 118).

To metodo smo izbrali zaradi njene sposobnosti naravno obravnavo stohastične narave bioloških sistemov in meritev, saj enega izmed rezultatov tega modelirnega pristopa predstavlja pogojne odvisnosti dogodkov. Rezultat Bayesove metode ni tako le ena kombinacija parametrov, marveč seznam kombinacij, ki se jih na podlagi eksperimentalnih podatkov ne da ovreči, zato pristop predstavlja primeren prvi začetni korak parametrizacije manjših modelov. Rezultate valov postopka (skupine vrednosti parameterskega prostora) pa se lahko nato naknadno lotimo z optimizacijo vsakega posebej in točnejše določitve. V primeru da bi namesto manjšega podomrežja etilenske signalizacije želeli parametrizirati celo omrežje pis (105 povezav), pa bi se bilo bolje poslužiti pristopov npr. Boolejevih omrežnih modelov.

Priprava modela je tako pomembna in odvisna od same formulacije – vsak algoritem ima svoje specifike in zmožnosti in imajo svoj razpon velikosti omrežij na katerih uspešno delujejo. Za vsako komponento smo lahko zaradi malega števila komponent ročno izbrali najbolj reprezentativno mikromrežno sondu, pod domnevo da lahko s transkriptomskimi vrednostmi izražanja dobimo dovolj dober približek dejanski reakciji, ki sicer poteka na nivoju proteinov. Ker so bili dani podatki na osnovi transkriptomike, nismo bili zmožni meriti vseh (proteinskih kompleksov ter metabolita). V modeliranju smo uporabili le transkriptomske podatke za kontrolne spodnje liste kultivarja Désirée, medtem ko bi bilo za dano problematiko bolj informativna primerjava med slepo/virusno inokuliranimi listi. To naloge smo pustili za prihodnost, saj je potrebno izbrati še strategijo in način primerjave, ki ga bomo uporabili. Na voljo je še nekaj eksperimentalnih časovnih serij, specifično ChiP-seq in RNA-seq podatki (An in sod., 2010) v rastlini navadni repnjakovec ter časovno resolved podatki iz paradižnika, krompirja in navadnega repnjakovca. Za dinamični model imamo v načrtu še nadgradnjo in razširitev vsaj v model, ki bi pokrival še signalizacijo ostalih dveh hormonov pomembnih za odgovor, obenem pa primerjati rezultate med različnimi genotipi in tretmaji.

6 SKLEPI

Namen doktorskega dela sta bili integracija in razvoj metod oz. postopkov integracije za izboljšanje razumevanja odgovora krompirja na okužbo s krompirjevim virusom Y. V ta namen smo uporabili razne tehnike raziskovanja pod okriljem sistemsko biološkega raziskovanja.

S pripravo aplikacije GoMapMan in s tem vzporedno pripravo pan-genoma krompirja ter pan-transkriptoma tobaka smo pokazali pomen prenosa funkcionalnih informacij iz bolj proučenih modelnih organizmov na poljščine.

- Za potrebe združevanja informacije o genih krompirja smo združili množice dveh genomskega modelov in dveh množic zaporedij unigene v skupen pan-genom krompirja. Ob tem smo razvili postopek združevanja na osnovi primerjave zaporedij z algoritmom BLAST, ki z obzirom na končne skupine dobro ublaži težave lokalnih primerjav zaporedij uporabljenega algoritma. S postopkom smo pridobili okoli 35 tisoč glavnih predstavnikov, kar je v okviru ocen števila genov krompirja in primerljivo s številom v krompirju bližnjem sorodniku paradižniku. Dodatno se je primernost postopka pokazala v kasnejši ročni biološki analizi nekaterih tarčnih skupin v okviru Oddelka. Obenem smo delovanje postopka preverili tudi na organizmu brez določenega genomskega zaporedja preko priprave pan-transkriptoma tobaka.
- Pan-genom krompirja in gene šestih dodatnih rastlinskih vrst (genome navadnega repnjakovca, paradižnika, riže, rdeče pese, kakava ter pan-transkriptom tobaka) smo združili v razviti aplikaciji GoMapMan (Ramšak in sod., 2014). Poleg povezav genov z ontologijo MapMan aplikacija te gene povezuje med seboj preko uporabe osmih različnih implementiranih ortolognih skupin iz različnih virov in skupin.
- Preko postopka avtomatske konsolidacije funkcionalnih opisov genov smo v aplikaciji GoMapMan omogočili prenos informacije iz modelne rastline navadnega repnjakovca na funkcionalno slabše anotirane rastlinske vrste. Preko zasnove uporabniškega vmesnika smo omogočili tudi ročne popravke povezav med geni in njihovimi funkcionalnimi opisi. S tem smo v okviru skupine že popravili tako strukturo ontologije MapMan (preko izboljšave funkcijskih skupin opisuječ biotski stres) ter popravkov funkcionalnih opisov nekaterih skupin genov.

- Sledljivost, ki je implementirana v aplikaciji omogoča lažje sklepanje o kvaliteti funkcionalnega opisa ter s tem olajša delo interpretirajočim znanstvenikom. Način širjenja vsebovane informacije v bazi se je izkazal za primerno izbranega, saj so nekatere znanstvene skupine informacijo v teh izvoznih datotekah tudi že uporabile (glede na citate objavljenega članka). Dodatno se interes za aplikacijo kaže tudi v zunanjih sodelavah, saj danes aplikacija vsebuje sedem rastlinskih vrst, tri od teh so bile dodane glede na želje tujih znanstvenikov.

Preko izgradnje velikega omrežja iz rezultatov eksperimentov visoko zmogljivih tehnologij smo z analizo omrežij razširili trenutno znanje o rastlinskem obrambnem sistemu. Matematično modeliranje manjšega podmodela nam je potrdilo pomen dobre osnovne zasnove modela. Tako analize zgrajenega strukturnega modela kot analize rezultatov modeliranja so vodile v nove ideje ter zasnovano nekaterih novih hipotez, nekatere od njih pa so bile tudi že prenesene iz *in silico* testiranja v laboratorij.

- Že obstoječi model rastlinske imunske signalizacije smo prevedli iz formalizma CSML v formalizem SBML, saj je slednji omogočen v orodju CellDesigner, katerega zmožnosti dinamičnega modeliranja so bolje podprte preko povezave z orodjem COPASI. Vzporedno s prenosom med formalizmi smo obstoječi model tudi nadgradili, tako iz vidika obstoječih komponent kot tudi iz vidika vsebovanih reakcij. Tako smo dodali 22 novih komponent ter 32 novih reakcij na nivoju genskih družin navadnega repnjakovca.
- Heterogene tipe iz javno dostopnih virov podatkov smo združili v velikem omrežju rastline navadni repnjakovec, ki vsebuje povezave proteinskih interakcij, transkripcijske regulacije ter regulacije preko malih RNA. V veliko omrežje smo dodali tudi povezave iz nadgrajenega modela rastlinske imunske signalizacije. Končno omrežje pokriva okoli 50 % vseh genov navadnega repnjakovca (skoraj 16 tisoč), med katerimi obstoji skoraj 60 tisoč povezav.
- Iz primerno obdelanih lastnih transkriptomskih podatkov smo z metodami inference omrežij pripravili ne-tarčno ko-ekspresijsko omrežje programa BioLayout ter tudi tarčno omrežje povezav.

- Analize pripravljenih omrežij so nakazale kvaliteto povečane informacije preko integracije heterogenih tipov podatkov, preko podrobnejših analiz modela: obogatitev omrežja za specifične funkcije, ročnem pregledu sodelujočih komponent PVY in proteinov krompirja ter iskanjem najkrajših poti v omrežju.
- Uporaba linearne Bayesove metodologije ocenjevanja parametrov diferencialnih enačb matematičnega modela je že v samo enem uporabljenem valu znatno zmanjšala prostor možnih spremeljivk sistema in se tako izkazala za primerno metodo, ki se jo bo lahko ponovno uporabilo, enkrat ko se izboljša model ponovno pripravi za analizo z metodo.
- Končno omrežje rastlinske obrambe in z njim povezan model rastlinske obrambe sta se izkazala za uspešnega pri generiranju novih hipotez, preko katerih smo bili zmožni na grobo oceniti napovedno vrednost in kvaliteto modela. Za nekatere od postavljenih hipotez so bili tudi že zastavljeni poskusi v laboratoriju.

7 POVZETEK (Summary)

7.1 POVZETEK

Okužba rastline s patogenom sproži kompleksen niz interakcij med obema igralcema, kar vodi do sprememb v kompleksnem signalizacijskem omrežju rastline, genske aktivnosti in reprogramiranja celičnega metabolizma. Namen doktorskega dela je bil integracija in razvoj metod oz. postopkov integracije za izboljšanje razumevanje odgovora krompirja na okužbo s krompirjevim virusom Y. V ta namen smo uporabili razne tehnike raziskovanja pod okriljem sistemsko biološkega raziskovanja. S pripravo aplikacije GoMapMan in s tem vzporedno pripravo pan-genoma krompirja ter pan-transkriptoma tobaka smo pokazali pomen prenosa funkcionalnih informacij iz bolj proučenih modelnih organizmov na poljščine. Preko izgradnje velikega omrežja iz rezultatov eksperimentov visoko zmogljivih tehnologij smo močno razširili trenutno stanje znanja o rastlinskem obrambnem sistemu. Matematično modeliranje manjšega podmodela nam je potrdilo pomen dobre osnovne zasnove modela. Tako analize zgrajenega strukturnega modela kot analize rezultatov modeliranja pa so vodile v nove ideje ter zasnovo nekaterih novih hipotez, nekatere od njih pa so bile tudi že prenesene iz *in silico* testiranja v laboratorij.

V prvem koraku smo za potrebe aplikacije GoMapMan in kasnejše izboljšave vsebinskih analiz podatkov visoko zmogljivih tehnologij pripravili priprave pan-genoma krompirja na osnovi dveh genomskeh modelov in dveh izbranih skupin EST zaporedij. Razviti postopek je vključeval več vsebinskih korakov, od primerjave zaporedij z algoritmom BLAST, do vmesne priprave odločitvenega drevesa z metodo strojnega učenja, do končne analize v programu za analizo omrežij Pajek. Prilagojen postopek smo uporabili tudi pri pripravi pan-transkriptoma tobaka. Ročne analize nekaterih pridobljenih paralognih skupin so se izkazale za biološko smiselne, s čimer smo posredno lahko validirali uspešnost razvitega postopka.

Integracija heterogenih tipov podatkov v aplikaciji GoMapMan je uspešno doprinesla k izboljšanju interpretacij rezultatov eksperimentov visoko zmogljivih tehnologij, predvsem transkriptomike, pa tudi metabolomike. Postopek konsolidacije, razvit v okviru doktorskega dela je očitno pripomogel k izboljšavi ontoloških anotacij ne samo za samo navadni repnjakovec, ampak za tudi več kmetijsko pomembnih poljščin (i.e. krompirja, paradižnika, tobaka, sladkorne pese). Obenem preko priprave izvoznih datotek aplikacija omogoča implementacijo v aplikacijah v sodelovanju z Institutom Jožef Stefan, kjer je končnim uporabnikom olajšana analiza preko različnih vizualizacij rezultatov.

Nadgradnja modela rastlinske obrambe in s tem povezano združevanje javno dostopnih večjih podatkovnih nizov v veliko omrežje rastline navadnega repnjakovca se je izkazalo za uspešno pri tvorbi novih hipotez. Dejansko večjo časovno obremenitev predstavlja tvorba novih hipotez pod okriljem znanstvenika, ki mora rezultate analize omrežij pregledati ročno. Tudi matematično modeliranje, četudi trenutno aplicirano v precej omejenem obsegu, je izpostavilo pomembne lastnosti sistema, katerih se za produkcijo uspešnih rezultatov ne sme zanemariti v prihodnosti. V kratkem, družni rezultati analize omrežij in matematičnega modeliranja manjšega sistema so se izkazali za primeren pristop analize kompleksnega sistema okužbe rastline s patogenom, tako iz vidika potrjevanja že znanih odnosov med biološkimi akterji, kot tudi tvorbe in potrjevanja novih povezav.

7.2 SUMMARY

An infection of a plant with a patogen triggers a complex set of interactions between both of the involved parties, which leads to changes in the complex signalization network of the plant, gene activity and reprogramming of the cell metabolism. The task posed in this doctoral study was the integration and development of methods (integration procedures) to improve our understanding of the potato response to infection with potato virus Y. For this reason we used various research techniques under the term of systems biology research. By preparing the GoMapMan application and the parallel preparation of the potato pan-genome and tobacco pan-transcriptome, we show the importance of functional annotation transfer from better studied model organisms to agriculturally important crops. By building the large network from high-throughput technologies experimental results, we vastly expanded the current knowledge state about the plant immune signalling. Mathematical modeling of a smaller subset confirmed the importance of a good basic model design. Both analyses of the large structural model as the analyses of modeling results have led to new ideas and new hypothesis design, some of which have already been transferred from *in silico* analyses to wet-lab testing.

As part of GoMapMan application requirements and following improvements of biological data interpretations, we first prepared the pan-genome of potato on the basis of two genomic models and two selected EST datasets. The developed procedure included several steps, from sequence comparison using the BLAST algorithm, to the intermediate preparation of a decision tree using machine learning approaches, and finally analysis in the network analysis program Pajek. A slightly adjusted procedure was used when preparing the tobacco pan-transcriptome. Manual analyses of some of the resulting parologue groups have shown to have biologically relevant results, through which we were able to validate the efficiency of the used procedure.

Integration of heterogeneous data types in GoMapMan application has successfully improved the interpretation of high-throughput experiment results, mostly transcriptomics, and some metabolomics. The developed consolidation procedure has visibly improved the ontology annotation of not only model plant *A. thaliana*, but also several agriculturally important crops (i.e. potato, tomato, tobacco, sugarbeet). At the same time GoMapMan allows implementation in other applications in cooperation with Jozef Stefan Institute, which further ease the analyses via means of various data visualisation procedures.

Upgrading the plant defense model and with that connected integration of publically available large datasets in a large knowledge network for *A. thaliana* has proven to be successful at generating new hypotheses. In fact, actual manual checking of the generated potential hypotheses (by a biological expert) represents a greater time sink, than the actual generation of them. Also mathematical modelling, even though it was currently applied in a quite limited scope, has exposed the importance of some system properties, which should not be ignored in order to produce results of quality. In short, the joint results of network analyses and mathematical modeling of a smaller system have been proven to be an efficient approach to analyse the complex system of plant pathogen infection, from the confirmation of already knowledge viewpoint, as well as from the aspect of generation and validation of new connections.

8 VIRI

- Abe, H., Tomitaka, Y., Shimoda, T., Seo, S., Sakurai, T., Kugimiya, S., Tsuda, S., Kobayashi, M. 2012. Antagonistic plant defense system regulated by phytohormones assists interactions among vector insect, thrips and a Tospovirus. *Plant & Cell Physiology*, 53, 1: 204–212
- Adams, M. J., Antoniw, J. F., Beaudoin, F. 2005. Overview and analysis of the polyprotein cleavage sites in the family Potyviridae. *Molecular Plant Pathology*, 6, 4: 471–487
- Äijö, T., Lähdesmäki, H. 2009. Learning gene regulatory networks from gene expression measurements using non-parametric molecular kinetics. *Bioinformatics*, 25, 22: 2937–2944
- Albert, R. 2005. Scale-free networks in cell biology. *Journal of Cell Science*, 118: 4947–4957
- Albert, R., Barabási, A.-L. 2002. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74, 47: doi:10.1103/RevModPhys.74.47: 54 str.
- Albert, R., Jeong, H., Barabási, A.-L. 2000. Error and attack tolerance of complex networks. *Nature*, 406, 6794: 378–382
- Alcaide-Loridan, C., Jupin, I. 2012. Ubiquitin and plant viruses, let's play together! *Plant Physiology*, 160, 1: 72–82
- Alemani, D., Pappalardo, F., Pennisi, M., Motta, S., Brusic, V. 2012. Combining cellular automata and lattice Boltzmann method to model multiscale avascular tumor growth coupled with nutrient diffusion and immune competition. *Journal of Immunological Methods*, 376: 55–68
- Alon, U. 2007. Network motifs: theory and experimental approaches. *Nature Reviews. Genetics*, 8: 450–461
- Altenhoff, A. M., Dessimoz, C. 2009. Phylogenetic and functional assessment of orthologs inference projects and methods. *PLoS Computational Biology*, 5, 1: doi:10.1371/journal.pcbi.1000262: 11 str.
- Altschul, S. F., Madden, T. L., Schäffer, 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, 17: 3389–3402

- An, F., Zhao, Q., Ji, Y., Li, W., Jiang, Z., Yu, X., Zhang, C., Han, Y., He, W., Liu, Y., Zhang, S., Ecker, J. R., Guo, H. 2010. Ethylene-induced stabilization of ETHYLENE INSENSITIVE3 and EIN3-LIKE1 is mediated by proteasomal degradation of EIN3 binding F-Box 1 and 2 that requires EIN2 in Arabidopsis. *The Plant Cell*, 22, 7: 2384–2401
- Arabidopsis Interactome Mapping Consortium. 2011. Evidence for network evolution in an Arabidopsis interactome map. *Science*, 333, 6042: 601–607
- Arnaud, E., Cooper, L., Shrestha, R. 2012. Towards a reference plant trait ontology for modeling knowledge of plant traits and phenotypes. V: 4th International Conference on Knowledge Engineering and Ontology Development, Barcelona, Spain, 4. - 7. oktober 2012: doi:10.5220/0004138302200225: 7 str.
- Asai, T., Tena, G., Plotnikova, J., Willmann, M. R., Chiu, W.-L., Gomez-Gomez, L., Boller, T., Ausubel, F. M., Sheen, J. 2002. MAP kinase signalling cascade in Arabidopsis innate immunity. *Nature*, 415: 977–983
- Auffray, C., Imbeaud, S., Roux-Rouquié, M., Hood, L. 2003. From functional genomics to systems biology: concepts and practices. *Comptes Rendus - Biologies*, 326, 10-11: 879–892
- Azé, J., Gentils, L., Toffano-Nioche, C., Loux, V., Gibrat, J.-F., Bessières, P., Rouveiro, C., Poupon, A., Froidevaux, C. 2008. Towards a semi-automatic functional annotation tool based on decision-tree techniques. *BMC Proceedings*, 2, Suppl. 4:S3: doi:10.1186/1753-6561-2-s4-s3: 11 str.
- Bader, G. D., Hogue, C. W. V. 2003. An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics*, 4, 2: doi:10.1186/1471-2105-4-2: 27 str.
- Baebler, Š., Krečič-Stres, H., Rotter, A., Kogovšek, P., Cankar, K., Kok, E. J., Gruden, K., Kovač, M., Žel, J., Pompe-Novak, M., Ravnikar, M. 2009. PVYNTN elicits a diverse gene expression response in different potato genotypes in the first 12 h after inoculation. *Molecular Plant Pathology*, 10, 2: 263–275
- Baebler, Š., Stare, K., Kovač, M., Blejec, A., Prezelj, N., Stare, T., Kogovšek, P., Pompe-Novak, M., Rosahl, S., Ravnikar, M., Gruden, K. 2011. dynamics of responses in compatible potato - potato virus Y interaction are modulated by salicylic acid. *PLOS ONE*, 6, 12: doi:10.1371/journal.pone.0029009: 12 str.
- Baebler, Š., Witek, K., Petek, M., Stare, K., Tušek-Žnidarič, M., Pompe-Novak, M., Renaut, J., Szajko, K., Strzelczyk-Zyta, D., Marczewski, W., Morgiewicz, K., Gruden, K., Hennig, J. 2014. Salicylic acid is an indispensable component of the Ny-1 resistance-gene-mediated response against potato virus Y infection in potato. *Journal of Experimental Botany*, 65, 4: 1095–1109.

- Baier, G., Sahle, S. 1998. Homogeneous and spatio-temporal chaos in biochemical reactions with feedback inhibition. *Journal of Theoretical Biology*, 193, 2: 233–242
- Bailer, S. M., Haas, J. 2009. Connecting viral with cellular interactomes. *Current Opinion in Microbiology*, 12, 4: 453–459
- Baker, M. 2013. Big biology: The 'omes puzzle. *Nature*, 494: 416–419
- Bansal, M., Belcastro, V., Ambesi-Impiombato, A., di Bernardo, D. 2007. How to infer gene networks from expression profiles. *Molecular Systems Biology*, 3, 78: doi:10.1038/msb4100158: 10 str.
- Barabasi, A.-L., Albert, R. 1999. Emergence of scaling in random networks. *Science*, 286, 5439: 509–512
- Barabási, A.-L., Oltvai, Z. N. 2004. Network biology: Understanding the cell's functional organization. *Nature Reviews: Genetics*, 5: 101–113
- Batada, N. N., Reguly, T., Breitkreutz, A., Boucher, L., Breitkreutz, B. J., Hurst, L. D., Tyers, M. 2006. Stratus not altocumulus: A new view of the yeast protein interaction network. *PLoS Biology*, 4, 10: 1720–1731
- Batagelj, V., Mrvar, A. 1998. Pajek – Program for large network analysis. *Connections*, 21, 2: 47–57
- Bauer-Mehren, A., Furlong, L. I., Sanz, F. 2009. Pathway databases and tools for their exploitation: Benefits, current limitations and challenges. *Molecular Systems Biology*, 5, 290: doi:10.1038/msb.2009.47: 13 str.
- Baumgartner, W. A. J., Cohen, K. B., Fox, L. M., Acquaah-Mensah, G., Hunter, L. 2007. Manual curation is not sufficient for annotation of genomic databases. *Bioinformatics*, 23, 13: 41–48
- Beczner, L., Horváth, J., Romhányi, I., Förster, H. 1984. Studies on the etiology of tuber necrotic ringspot disease in potato. *Potato Research*, 27, 3: 339–352
- Benjamini, Y., Hochberg, Y. 2009. Controlling the False Discovery Rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, 57, 1: 289–300
- Benson, D. A., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., Sayers, E. W. 2014. GenBank. *Nucleic Acids Research*, 42, D1: 32–37
- Bhalla, U. S. 2003. Understanding complex signaling networks through models and metaphors. *Progress in Biophysics and Molecular Biology*, 81, 1: 45–65
- Bhalla, U. S., Ram, P. T., Iyengar, R. 2002. MAP kinase phosphatase as a locus of flexibility in a mitogen-activated protein kinase signaling network. *Science*, 297, 5583: 1018–1023

- Bilgin, D. D., Zavala, J. A., Zhu, J., Clough, S. J., Ort, D. R., DeLucia, E. H. 2010. Biotic stress globally downregulates photosynthesis genes. *Plant, Cell and Environment*, 33, 10: 1597–1613
- Boller, T., Felix, G. 2009. A renaissance of elicitors: Perception of microbe-associated molecular patterns and danger signals by pattern-recognition receptors. *Annual Review of Plant Biology*, 60: 379–406
- Bolstad, B. M., Irizarry, R. A., Astrand, M., Speed, T. P. 2003. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19, 2: 185–193
- Boquel, S., Ameline, A., Giordanengo, P. 2011. Assessing aphids potato virus Y-transmission efficiency: A new approach. *Journal of Virological Methods*, 178: 63–67
- Bourgon, R., Gentleman, R., Huber, W. 2010. Independent filtering increases detection power for high-throughput experiments. *Proceedings of the National Academy of Sciences*, 107, 21: 9546–9551
- Brady, S. M., Orlando, D. A., Lee, J.-Y., Wang, J. Y., Koch, J., Dinneny, J. R., Mace, D., Ohler, U., Benfey, P. N. 2007. A high-resolution root spatiotemporal map reveals dominant expression patterns. *Science*, 318, 5851: 801–806
- Brandão, M. M., Dantas, L. L., Silva-Filho, M. C. 2009. AtPIN: *Arabidopsis thaliana* protein interaction network. *BMC Bioinformatics*, 10, 454: doi:10.1186/1471-2105-10-454: 7 str.
- Bray, D. 2003. Molecular networks: The top-down view. *Science*, 301, 5641: 1864–1865
- Breeze, E., Harrison, E., McHattie, S., Hughes, L., Hickman, R., Hill, C., Kiddle, S., Kim, Y., Penfold, C. A., Jenkins, D., Zhang, C., Morris, K., Jenner, C., Jackson, S., Thomas, B., Tabrett, A., Legaie, R., Moore, J. D., Wild, D. L., Ott, S., Rand, D., Beynon, J., Denby, K., Mead, A., Buchanan-Wollaston, V. 2011. High-resolution temporal profiling of transcripts during *Arabidopsis* leaf senescence reveals a distinct chronology of processes and regulation. *The Plant Cell*, 23: 873–894
- Buttigieg, P. L., Morrison, N., Smith, B., Mungall, C. J., Lewis, S. E. 2013. The environment ontology: contextualising biological and biomedical entities. *Journal of Biomedical Semantics*, 4, 43: doi:10.1186/2041-1480-4-43: 9 str.
- Cai, B., Li, C.-H., Huang, J. 2014. Systematic identification of cell-wall related genes in *populus* based on analysis of functional modules in co-expression network. *PLoS ONE*, 9, 4: doi:10.1371/journal.pone.0095176: 9 str.

- Cao, J., Schneeberger, K., Ossowski, S., Günther, T., Bender, S., Fitz, J., Koenig, D., Lanz, C., Stegle, O., Lippert, C., Wang, X., Ott, F., Müller, J., Alonso-Blanco, C., Borgwardt, K., Schmid, K. J., Weigel, D. 2011. Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. *Nature Genetics*, 43, 10: 956–963
- Carrera, J., Rodrigo, G., Jaramillo, A., Elena, S. F. 2009. Reverse-engineering the *Arabidopsis thaliana* transcriptional network under changing environmental conditions. *Genome Biology*, 10, 9: doi:10.1186/gb-2009-10-9-r96: 15 str.
- Carroll, A. J., Badger, M. R., Millar, A. H. 2010. The MetabolomeExpress project: enabling web-based processing, analysis and transparent dissemination of GC/MS metabolomics datasets. *BMC Bioinformatics*, 11: doi:10.1186/1471-2105-11-376: 13 str.
- Chakraborty, A., Bandyopadhyay, S. 2013. FOGSAA: Fast optimal global sequence alignment algorithm. *Scientific Reports*, 3: doi:10.1038/srep01746: 9 str.
- Chan, C. X., Mahboub, M., Ragan, M. A. 2013. clustering evolving proteins into homologous families. *BMC Bioinformatics*, 14: doi:10.1186/1471-2105-14-120: 11 str.
- Chang, K. N., Zhong, S., Weirauch, M. T., Hon, G., Pelizzola, M., Li, H., Carol Huang, S. S., Schmitz, R. J., Urich, M. a., Kuo, D., Nery, J. R., Qiao, H., Yang, A., Jamali, A., Chen, H., Ideker, T., Ren, B., Bar-Joseph, Z., Hughes, T. R., Ecker, J. R. 2013. Temporal transcriptional response to ethylene gas drives growth hormone cross-regulation in *Arabidopsis*. *eLife*, 2: doi:10.7554/eLife.00675: 20 str.
- Chatterjee, S., Simonoff, J. S. 2012. Nonlinear regression. V: Handbook of regression analysis. Chatterjee S. in Simonoff J. S. (eds.). New York, John Wiley & Sons: 215–225
- Chen, F., Mackey, A. J., Vermunt, J. K., Roos, D. S. 2007. Assessing performance of orthology detection strategies applied to eukaryotic genomes. *PLoS ONE*, 2, 4: doi:10.1371/journal.pone.0000383: 12 str.
- Chen, J., Wu, H. 2008. Efficient local estimation for time-varying coefficients in deterministic dynamic models with applications to HIV-1 dynamics. *Journal of the American Statistical Association*, 103, 481: 369–384
- Childs, K. L., Hamilton, J. P., Zhu, W., Ly, E., Cheung, F., Wu, H., Rabinowicz, P. D., Town, C. D., Buell, C. R., Chan, A. P. 2007. The TIGR plant transcript assemblies database. *Nucleic Acids Research*, 35: 846–851
- Chung, B. Y. W., Miller, W. A., Atkins, J. F., Firth, A. E. 2008. An overlapping essential gene in the Potyviridae. *Proceedings of the National Academy of Sciences of the United States of America*, 105, 15: 5897–5902

- Clarke, J. D., Volko, S. M., Ledford, H., Ausubel, F. M., Dong, X. 2000. Roles of salicylic acid, jasmonic acid, and ethylene in cpr-induced resistance in *Arabidopsis*. *The Plant Cell*, 12, 11: 2175–2190
- Cohen, W. W. 1994. Fast effective rule induction. V: Machine learning: Proceedings of the twelfth International conference (10 str.).
<http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.50.8204> (16. sep. 2015)
- Cooper, B., Clarke, J. D., Budworth, P., Kreps, J., Hutchison, D., Park, S., Guimil, S., Dunn, M., Luginbühl, P., Ellero, C., Goff, S. A., Glazebrook, J. 2003. A network of rice genes associated with stress response and seed development. *Proceedings of the National Academy of Sciences of the United States of America*, 100, 8: 4945–4950
- Courtot, M., Juty, N., Knüpfer, C., Waltemath, D., Zhukova, A., Dräger, A., Dumontier, M., Finney, A., Golebiewski, M., Hastings, J., Hoops, S., Keating, S., Kell, D. B., Kerrien, S., Lawson, J., Lister, A., Lu, J., Machne, R., Mendes, P., Pocock, M., Rodriguez, N., Villeger, A., Wilkinson, D. J., Wimalaratne, S., Laibe, C., Hucka, M., Le Novère, N. 2011. controlled vocabularies and semantics in systems biology. *Molecular Systems Biology*, 7: doi:10.1038/msb.2011.77: 12 str.
- Craig, P. S., Goldstein, M., Rougier, J. C., Seheult, A. H. 2001. Bayesian forecasting for complex systems using computer simulators. *Journal of the American Statistical Association*, 96, 454: 717–729
- Craig, P. S., Goldstein, M., Seheult, A. H., Smith, J. A. 1997. pressure matching for hydrocarbon reservoirs: A case study in the use of Bayes linear strategies for large computer experiments. *Lecture Notes in Statistics*, 121: 37–93
- Cui, J., Li, P., Li, G., Xu, F., Zhao, C., Li, Y., Yang, Z., Wang, G., Yu, Q., Li, Y., Shi, T. 2008. AtPID: *Arabidopsis thaliana* protein interactome database - An integrative platform for plant systems biology. *Nucleic Acids Research*, 36: 999–1008
- Cusick, M. E., Klitgord, N., Vidal, M., Hill, D. E. 2005. Interactome: Gateway into systems biology. *Human Molecular Genetics*, 14, 2: 171–181
- De Backer, P., De Waele, D., van Speybroeck, L. 2010. Ins and outs of systems biology vis-a-vis molecular biology: Continuation or clear cut? *Acta Biotheoretica*, 58: 15–49
- De Bodt, S., Carvajal, D., Hollunder, J., Van den Cruyce, J., Movahedi, S., Inzé, D. 2010. CORNET: A user-friendly tool for data mining and integration. *Plant Physiology*, 152, 3: 1167–1179

- De Bodt, S., Proost, S., Vandepoele, K., Rouzé, P., Van de Peer, Y. 2009. Predicting protein-protein interactions in *Arabidopsis thaliana* through integration of orthology, gene ontology and co-expression. BMC Genomics, 10: doi:10.1186/1471-2164-10-288: 15 str.
- De Boor, C., Swartz, B. 1973. Collocation at Gaussian points. Siam Journal on Numerical Analysis, 5, 3: 530–558
- DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., Philippakis, A. A., Del Angel, G., Rivas, M. A., Hanna, M., McKenna, A., Fennell, T. J., Kernytsky, A. M., Sivachenko, A. Y., Cibulskis, K., Gabriel, S. B., Altshuler, D., Daly, M. J. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature Genetics, 43, 5: 491–498
- Dhar, P. K. 2007. The next step in biology: A periodic table? Journal of Biosciences, 32, 5: 1005–1008
- Di Carli, M., Benvenuto, E., Donini, M. 2012. Recent insights into plant-virus interactions through proteomic analysis. Journal of Proteome Research, 11, 10: 4765–4780
- Ding, S. W., Voinnet, O. 2007. antiviral immunity directed by small RNAs. Cell, 130, 3: 413–426
- Dodds, P. N., Rathjen, J. P. 2010. Plant immunity: Towards an integrated view of plant-pathogen interactions. Nature Reviews: Genetics, 11, 8: 539–548
- Du Plessis, L., Skunca, N., Dessimoz, C. 2011. the what, where, how and why of gene ontology: A primer for bioinformaticians. Briefings in Bioinformatics, 12, 6: 723–735
- Dutkowski, J., Kramer, M., Surma, M. A., Balakrishnan, R., Cherry, J. M., Krogan, N. J., Ideker, T. 2013. A gene ontology inferred from molecular networks. Nature Biotechnology, 31, 1: 38–45
- Duvick, J., Fu, A., Muppirala, U., Sabharwal, M., Wilkerson, M. D., Lawrence, C. J., Lushbough, C., Brendel, V. 2008. PlantGDB: A resource for comparative plant genomics. Nucleic Acids Research, 36: 959–965
- Edgar, R., Domrachev, M., Lash, A. E. 2002. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Research, 30, 1: 207–210
- Edwards, K. D., Bombarély, A., Story, G. W., Allen, F., Mueller, L. A., Coates, S. A., Jones, L. 2010. TobEA: An atlas of tobacco gene expression from seed to senescence. BMC Genomics, 11: doi:10.1186/1471-2164-11-142: 15 str.

- Elena, S. F., Carrera, J., Rodrigo, G. 2011. A systems biology approach to the evolution of plant-virus interactions. *Current Opinion in Plant Biology*, 14, 4: 372–377
- Elena, S. F., Rodrigo, G. 2012. Towards an integrated molecular model of plant-virus interactions. *Current Opinion in Virology*, 2, 6: 713–718
- Endy, D., Brent, R. 2001. Modelling cellular behaviour. *Nature*, 409, 6818: 391–395
- Eronen, L., Toivonen, H. 2012. Biomine: Predicting links between biological entities using network models of heterogeneous databases. *BMC Bioinformatics*, 13: doi:10.1186/1471-2105-13-119: 21 str.
- FAOSTAT. 2013. Food and Agriculture Organization of the United Nations database. <http://faostat3.fao.org/home/E/> (7. maj 2015)
- Farcot, E., Lavedrine, C., Vernoux, T. 2015. A modular analysis of the auxin signalling network. *PLOS ONE*, 10, 3: doi:10.1371/journal.pone.0122231: 27 str.
- Fernandez-Pozo, N., Menda, N., Edwards, J. D., Saha, S., Tecle, I. Y., Strickler, S. R., Bombarely, A., Fisher-York, T., Pujar, A., Foerster, H., Yan, A., Mueller, L. A. 2014. The Sol Genomics Network (SGN) - from genotype to phenotype to breeding. *Nucleic Acids Research*, 43, D1: 1036–1041
- Freund, Y., Mason, L. 1999. The alternating decision tree learning algorithm. Proceeding of the Sixteenth International conference on machine learning: doi:10.1007/s13398-014-0173-7.2: 10 str.
- Fu, L., Niu, B., Zhu, Z., Wu, S., Li, W. 2012. CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics*, 28, 23L: 3150–3152
- Funahashi, A., Morohashi, M., Kitano, H., Tanimura, N. 2003. CellDesigner: A process diagram editor for gene-regulatory and biochemical networks. *BIOSILICO*, 1, 5: 159–162
- Galbraith, S. J., Tran, L. M., Liao, J. C. 2006. Transcriptome network component analysis with limited microarray data. *Bioinformatics*, 22, 15: 1886–1894
- Gao, P., Honkela, A., Rattray, M., Lawrence, N. D. 2008. Gaussian process modelling of latent chemical species: Applications to inferring transcription factor activities. *Bioinformatics*, 24, 16: 70–75
- Ge, H., Walhout, A. J. M., Vidal, M. 2003. Integrating “omic” information: A bridge between genomics and systems biology. *Trends in Genetics*, 19, 10: 551–560
- Gehan, M. A., Greenham, K., Mockler, T. C., McClung, C. R. 2015. Transcriptional networks - crops, clocks, and abiotic stress. *Current Opinion in Plant Biology*, 24: 39–46
- Gillespie, D. T. 2001. Approximate accelerated stochastic simulation of chemically reacting systems. *Journal of Chemical Physics*, 115, 4: 1716–1733

- Godsey, B. 2013. Improved inference of gene regulatory networks through integrated Bayesian clustering and dynamic modeling of time-course expression data. *PLoS ONE*, 8, 7: doi:10.1371/journal.pone.0068358: 6 str.
- Gómez-Gómez, L., Felix, G., Boller, T. 1999. A single locus determines sensitivity to bacterial flagellin in *Arabidopsis thaliana*. *Plant Journal*, 18, 3: 277–284
- González, I., Le Cao, K.-A., Davis, M. J., Déjean, S. 2012. Visualising associations between paired “omics” data sets. *BioData Mining*, 5: doi:10.1186/1756-0381-5-19: 19 str.
- Goodstein, D. M., Shu, S., Howson, R., Neupane, R., Hayes, R. D., Fazo, J., Mitros, T., Dirks, W., Hellsten, U., Putnam, N., Rokhsar, D. S. 2012. Phytozome: A comparative platform for green plant genomics. *Nucleic Acids Research*, 40: 1178–1186
- Grimplet, J., Van Hemert, J., Carbonell-Bejerano, P., Díaz-Riquelme, J., Dickerson, J., Fennell, A., Pezzotti, M., Martínez-Zapater, J. M. 2012. Comparative analysis of grapevine whole-genome gene predictions, functional annotation, categorization and integration of the predicted gene sequences. *BMC Research Notes*, 5, 213: doi:10.1186/1756-0500-5-213: 10 str.
- Gu, H., Zhu, P., Jiao, Y., Meng, Y., Chen, M. 2011. PRIN: A predicted rice interactome network. *BMC Bioinformatics*, 12: doi:10.1186/1471-2105-12-161: 13 str.
- Guo, D., Rajamäki, M.-L., Saarma, M., Valkonen, J. P. T. 2001. Towards a protein interaction map of potyviruses: Protein interaction matrixes of two Potyviruses based on the yeast two-hybrid system. *Journal of General Virology*, 82, 4: 935–939
- Gutiérrez, R. A., Lejay, L. V., Dean, A., Chiaromonte, F., Shasha, D. E., Coruzzi, G. M. 2007. Qualitative network models and genome-wide expression data define carbon/nitrogen-responsive molecular machines in *Arabidopsis*. *Genome Biology*, 8, 1: doi:10.1186/gb-2007-8-1-r7: 13 str.
- Gutiérrez, R. A., Shasha, D. E., Coruzzi, G. M. 2005. Systems biology for the virtual plant. *Plant Physiology*, 138, 2: 550–554
- Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I. H. 2009. The WEKA data mining software: An update. *ACM SIGKDD Explorations Newsletter*, 11, 1: 10–18
- Hartung, M., Gross, A., Rahm, E. 2012. CODEX: Exploration of semantic changes between ontology versions. *Bioinformatics*, 28, 6: 895–896
- Heinlein, M. 2015. Plant virus replication and movement. *Virology*, 479: 657–671

- Helmy, M., Sugiyama, N., Tomita, M., Ishihama, Y. 2012. The rice proteogenomics database OryzaPG-DB: Development, expansion, and new features. *Frontiers in Plant Science*, 3: doi:10.3389/fpls.2012.00065: 6 str.
- Heyndrickx, K. S., Vandepoele, K. 2012. Systematic identification of functional plant modules through the integration of complementary data sources. *Plant Physiology*, 159, 3: 884–901
- Himmelblau, D. M., Jones, C. R., Bischoff, K. B. 1967. Determination of rate constants for complex kinetics models. *Industrial in Engineering Chemistry Fundamentals*, 6, 4: 539–543
- Hinrichs, J., Berger, S., Shaw, J. G. 1998. A hypersensitive response-like mechanism is involved in resistance of potato plants bearing the Ry(sto) gene to the Potyviruses potato virus Y and tobacco etch virus. *Journal of General Virology*, 79, 1: 167–176
- Hirai, M. Y., Sugiyama, K., Sawada, Y., Tohge, T., Obayashi, T., Suzuki, A., Araki, R., Sakurai, N., Suzuki, H., Aoki, K., Goda, H., Nishizawa, O. I., Shibata, D., Saito, K. 2007. Omics-based identification of *Arabidopsis* Myb transcription factors regulating aliphatic glucosinolate biosynthesis. *Proceedings of the National Academy of Sciences of the United States of America*, 104, 15: 6478–6483
- Hoops, S., Gauges, R., Lee, C., Pahle, J., Simus, N., Singhal, M., Xu, L., Mendes, P., Kummer, U. 2006. COPASI - A COmplex PAthway SImulator. *Bioinformatics*, 22, 24: 3067–3074
- Horvath, S., Dong, J. 2008. Geometric interpretation of gene coexpression network analysis. *PLoS Computational Biology*, 4, 8: doi: 10.1371/journal.pcbi.1000117: 27 str.
- Hsu, S.-D., Tseng, Y.-T., Shrestha, S., Lin, Y.-L., Khaleel, A., Chou, C.-H., Chu, C.-F., Huang, H.-Y., Lin, C.-M., Ho, S.-Y., Jian, T.-Y., Lin, F.-M., Chang, T.-H., Weng, S.-L., Liao, K.-W., Liao, I.-E., Liu, C.-C., Huang, H.-D. 2014. miRTarBase update 2014: An information resource for experimentally validated miRNA-target interactions. *Nucleic Acids Research*, 42: 78–85
- Hu, Q. N., Zhu, H., Li, X., Zhang, M., Deng, Z., Yang, X., Deng, Z. 2012. Assignment of EC numbers to enzymatic reactions with reaction difference fingerprints. *PLoS ONE*, 7, 12: doi:10.1371/journal.pone.0052901: 6 str.
- Huang, D. W., Sherman, B. T., Lempicki, R. A. 2009. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols*, 4, 1: 44–57
- Huang, X., Chao, K.-M. 2003. A generalized global alignment algorithm. *Bioinformatics* (Oxford, England), 19, 2: 228–233

- Huang, Y., Liu, D., Wu, H. 2006. Hierarchical Bayesian methods for estimation of parameters in a longitudinal HIV dynamic system. *Biometrics*, 62, 2: 413–423
- Huber, W., Von Heydebreck, A., Sültmann, H., Poustka, A., Vingron, M. 2002. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, 18, Suppl. 1: 96–104
- Humphries, M. D., Gurney, K. 2008. Network “small-world-ness”: A quantitative method for determining canonical network equivalence. *PLoS ONE*, 3, 4: doi:10.1371/journal.pone.0002051: 10 str.
- Hyduke, D. R., Palsson, B. Ø. 2010. Towards genome-scale signalling network reconstructions. *Nature Reviews: Genetics*, 11, 4: 297–307
- Itkin, M., Heinig, U., Tzfadia, O., Bhide, A. J., Shinde, B., Cardenas, P. D., Bocobza, S. E., Unger, T., Malitsky, S., Finkers, R., Tikunov, Y., Bovy, A., Chikate, Y., Singh, P., Rogachev, I., Beekwilder, J., Giri, A. P., Aharoni, A. 2013. Biosynthesis of antinutritional alkaloids in solanaceous crops is mediated by clustered genes. *Science*, 341, 6142: 175–179
- Ivanov, K. I., Eskelin, K., Lõhmus, A., Mäkinen, K. 2014. Molecular and cellular mechanisms underlying Potyvirus infection. *The Journal of General Virology*, 95: 1415–1429
- Jauhainen, A., Nerman, O., Michailidis, G., Jörnsten, R. 2012. Transcriptional and metabolic data integration and modeling for identification of active pathways. *Biostatistics*, 13, 4: 748–761
- Jiao, Y., Lawler, K., Patel, G. S., Purushotham, A., Jones, A. F., Grigoriadis, A., Tutt, A., Ng, T., Teschendorff, A. E. 2011. DART: Denoising algorithm based on relevance network topology improves molecular pathway activity inference. *BMC Bioinformatics*, 12: doi:10.1186/1471-2105-12-403: 16 str.
- Jones, A. M., Xuan, Y., Xu, M., Wang, R.-S., Ho, C.-H., Lalonde, S., You, C. H., Sardi, M. I., Parsa, S. A., Smith-Valle, E., Su, T., Frazer, K. A., Pilot, G., Pratelli, R., Grossmann, G., Acharya, B. R., Hu, H.-C., Engineer, C., Villiers, F., Ju, C., Takeda, K., Su, Z., Dong, Q., Assmann, S. M., Chen, J., Kwak, J. M., Schroeder, J. I., Albert, R., Rhee, S. Y., Frommer, W. B. 2014. Border control - A membrane-linked interactome of *Arabidopsis*. *Science*, 344, 6185: 711–716
- Jones, J. D. G., Dangl, J. L. 2006. The plant immune system. *Nature*, 444, 7117: 323–329
- Jordan, I. K., Mariño-Ramírez, L., Wolf, Y. I., Koonin, E. V. 2004. Conservation and coevolution in the scale-free human gene coexpression network. *Molecular Biology and Evolution*, 21, 11: 2058–2070

- Kang, Y. H., Kirik, V., Hulskamp, M., Nam, K. H., Hagely, K., Lee, M. M., Schiefelbein, J. 2009. The MYB23 gene provides a positive feedback loop for cell fate specification in the *Arabidopsis* root epidermis. *The Plant Cell*, 21, 4: 1080–1094
- Karasev, A. V., in Gray, S. M. 2013. Continuous and emerging challenges of potato virus Y in potato. *Annual Review of Phytopathology*, 51: 571–86
- Karlebach, G., Shamir, R. 2008. Modelling and analysis of gene regulatory networks. *Nature Reviews: Molecular Cell Biology*, 9, 10: 770–780
- Katerji, N., Van Hoorn, J. W., Hamdy, A., Mastorilli, M. 2000. Salt tolerance classification of crops according to soil salinity and to water stress day index. *Agricultural Water Management*, 43, 1: 99–109
- Kerrien, S., Aranda, B., Breuza, L., Bridge, A., Broackes-Carter, F., Chen, C., Duesbury, M., Dumousseau, M., Feuermann, M., Hinz, U., Jandrasits, C., Jimenez, R. C., Khadake, J., Mahadevan, U., Masson, P., Pedruzzi, I., Pfeiffenberger, E., Porras, P., Raghunath, A., Roechert, B., Orchard, S., Hermjakob, H. 2012. The IntAct molecular interaction database in 2012. *Nucleic Acids Research*, 40: 841–846
- Kersey, P. J., Allen, J. E., Christensen, M., Davis, P., Falin, L. J., Grabmueller, C., Hughes, D. S. T., Humphrey, J., Kerhornou, A., Khobova, J., Langridge, N., McDowall, M. D., Maheswari, U., Maslen, G., Nuhn, M., Ong, C. K., Paulini, M., Pedro, H., Toneva, I., Tuli, M. A., Walts, B., Williams, G., Wilson, D., Youens-Clark, K., Monaco, M. K., Stein, J., Wei, X., Ware, D., Bolser, D. M., Howe, K. L., Kulesha, E., Lawson, D., Staines, D. M. 2014. Ensembl Genomes 2013: Scaling up access to genome-wide data. *Nucleic Acids Research*, 42, D1: D546–D552
- Keurentjes, J. J. B., Angenent, G. C., Dicke, M., Dos Santos, V. A. P. M., Molenaar, J., Van der Putten, W. H., De Ruiter, P. C., Struik, P. C., Thomma, B. P. H. J. 2011. Redefining plant systems biology: From cell to ecosystem. *Trends in Plant Science*, 16, 4: 183–190
- Khouri, C. K., Bjorkman, A. D., Dempewolf, H., Ramirez-Villegas, J., Guarino, L., Jarvis, A., Rieseberg, L. H., Struik, P. C. 2014. Increasing homogeneity in global food supplies and the implications for food security. *Proceedings of the National Academy of Sciences of the United States of America*, 111, 11: 4001–4006
- Kiddle, S. J., Windram, O. P. F., McHattie, S., Mead, A., Beynon, J., Buchanan-Wollaston, V., Denby, K. J., Mukherjee, S. 2009. Temporal clustering by affinity propagation reveals transcriptional modules in *Arabidopsis thaliana*. *Bioinformatics*, 26, 3: 355–362
- Kiehl, T. R., Mattheyses, R. M., Simmons, M. K. 2004. Hybrid simulation of cellular behavior. *Bioinformatics*, 20, 3: 316–322

- Kirk, P. D. W., Stumpf, M. P. H. 2009. Gaussian process regression bootstrapping: Exploring the effects of uncertainty in time course data. *Bioinformatics*, 25, 10: 1300–1306
- Kitano, H. 2002. Systems biology: A brief overview. *Science*, 295, 5560: 1662–1664
- Klie, S., Nikoloski, Z. 2012. The choice between MapMan and Gene Ontology for automated gene function prediction in plant science. *Frontiers in Genetics*, 3, 115: doi:10.3389/fgene.2012.00115: 14 str.
- Kloosterman, B., De Koeyer, D., Griffiths, R., Flinn, B., Steuernagel, B., Scholz, U., Sonnewald, S., Sonnewald, U., Bryan, G. J., Prat, S., Bánfalvi, Z., Hammond, J. P., Geigenberger, P., Nielsen, K. L., Visser, R. G. F., Bachem, C. W. B. 2008. Genes driving potato tuber initiation and growth: Identification based on transcriptional changes using the POCI array. *Functional and Integrative Genomics*, 8, 4: 329–340
- Kogovšek, P., Pompe-Novak, M., Baebler, Š., Rotter, A., Gow, L., Gruden, K., Foster, G. D., Boonham, N., Ravnikar, M. 2010. Aggressive and mild potato virus Y isolates trigger different specific responses in susceptible potato plants. *Plant Pathology*, 59, 6: 1121–1132
- Kolesnikov, N., Hastings, E., Keays, M., Melnichuk, O., Tang, Y. A., Williams, E., Dylag, M., Kurbatova, N., Brandizi, M., Burdett, T., Megy, K., Pilicheva, E., Rustici, G., Tikhonov, A., Parkinson, H., Petryszak, R., Sarkans, U., Brazma, A. 2014. ArrayExpress update: Simplifying data submissions. *Nucleic Acids Research*, 43: 1113–1116
- Koornneef, A., Pieterse, C. M. J. 2008. Cross talk in defense signaling. *Plant Physiology*, 146, 3: 839–844
- Kovač, M., Müller, A., Milovanovič Jarh, D., Milavec, M., Düchting, P., Ravnikar, M. 2009. Multiple hormone analysis indicates involvement of jasmonate signalling in the early defence of potato to potato virus Y^{NTN}. *Biologia Plantarum*, 53, 1: 195–199
- Kraak, A. 1992. Industrial applications of potato starch products. *Industrial Crops and Products*, 1: 107–112
- Krecic-Stres, H., Vučak, C., Ravnikar, M., Kovac, M. 2005. Systemic potato virus Y^{NTN} infection and levels of salicylic and gentisic acids in different potato genotypes. *Plant Pathology*, 54, 4: 441–447
- Krouk, G., Lingeman, J., Colon, A. M., Coruzzi, G., Shasha, D. 2013. Gene regulatory networks in plants: learning causality from time and perturbation. *Genome Biology*, 14, 6: doi:10.1186/gb-2013-14-6-123: 7 str.

- Kuzniar, A., van Ham, R. C. H. J., Pongor, S., Leunissen, J. a M. 2008. The quest for orthologs: finding the corresponding gene across genomes. *Trends in Genetics*, 24, 11: 539–551
- Lamesch, P., Berardini, T. Z., Li, D., Swarbreck, D., Wilks, C., Sasidharan, R., Muller, R., Dreher, K., Alexander, D. L., Garcia-Hernandez, M., Karthikeyan, A. S., Lee, C. H., Nelson, W. D., Ploetz, L., Singh, S., Wensel, A., Huala, E. 2012. The *Arabidopsis* Information Resource (TAIR): Improved gene annotation and new tools. *Nucleic Acids Research*, 40: 1202–1210
- Langenberg, W. G., Zhang, L. 1997. Immunocytology shows the presence of tobacco etch virus P3 protein in nuclear inclusions. *Journal of Structural Biology*, 118, 3: 243–247
- Langfelder, P., Horvath, S. 2008. WGCNA: An R package for weighted correlation network analysis. *BMC Bioinformatics*, 9: doi:10.1186/1471-2105-9-559: 13 str.
- Lee, I., Ambaru, B., Thakkar, P., Marcotte, E. M., Rhee, S. Y. 2010. Rational association of genes with traits using a genome-scale gene network for *Arabidopsis thaliana*. *Nature Biotechnology*, 28, 2: 149–156
- Lee, J. M., Sonnhammer, E. L. L. 2003. Genomic gene clustering analysis of pathways in eukaryotes. *Genome Research*, 13, 5: 875–882
- Lescot, M., Déhais, P., Thijs, G., Marchal, K., Moreau, Y., Van de Peer, Y., Rouzé, P., Rombauts, S. 2002. PlantCARE, a database of plant cis-acting regulatory elements and a portal to tools for in silico analysis of promoter sequences. *Nucleic Acids Research*, 30: 325–327
- Li, F., Ding, S.-W. 2006. Virus counterdefense: Diverse strategies for evading the RNA-silencing immunity. *Annual Review of Microbiology*, 60: 503–531
- Li, L., Stoeckert, C. J., Roos, D. S. 2003. OrthoMCL: Identification of ortholog groups for eukaryotic genomes. *Genome Research*, 13, 9: 2178–2189
- Li, S., Assmann, S. M., Albert, R. 2006. Predicting essential components of signal transduction networks: A dynamic model of guard cell abscisic acid signaling. *PLoS Biology*, 4, 10: 1732–1748
- Li, W., Godzik, A. 2006. CD-HIT: A Fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22, 13: 1658–1659
- Liang, C., Jaiswal, P., Hebbard, C., Avraham, S., Buckler, E. S., Casstevens, T., Hurwitz, B., Mccouch, S., Ni, J., Pujar, A., Ravenscroft, D., Ren, L., Spooner, W., Tecle, I., Thomason, J., Tung, C. W., Wei, X., Yap, I., Youens-Clark, K., Ware, D., Stein, L. 2008. Gramene: A growing plant comparative genomics resource. *Nucleic Acids Research*, 36: 947–953

- Liao, J. C., Boscolo, R., Yang, Y.-L., Tran, L. M., Sabatti, C., Roychowdhury, V. P. 2003. Network component analysis: Reconstruction of regulatory signals in biological systems. *Proceedings of the National Academy of Sciences of the United States of America*, 100, 26: 15522–15527
- Likhoshvai, V. A., Omel'yanchuk, N. A., Mironova, V. V., Fadeev, S. I., Mjolsness, E. D., Kolchanov, N. A. 2007. Mathematical model of auxin distribution in the plant root. *Russian Journal of Developmental Biology*, 38, 6: 374–382
- Lin, L., Shi, Y., Luo, Z., Lu, Y., Zheng, H., Yan, F., Chen, J., Chen, J., Adams, M. J., Wu, Y. 2009. Protein-protein interactions in two Potyviruses using the yeast two-hybrid system. *Virus Research*, 142, 1-2: 36–40
- Lin, S. M., Du, P., Huber, W., Kibbe, W. A. 2008. Model-based variance-stabilizing transformation for Illumina microarray data. *Nucleic Acids Research*, 36, 2: doi:10.1093/nar/gkm1075: 9 str.
- Lister, R., O'Malley, R. C., Tonti-Filippini, J., Gregory, B. D., Berry, C. C., Millar, A. H., Ecker, J. R. 2008. Highly integrated single-base resolution maps of the epigenome in *Arabidopsis*. *Cell*, 133, 3: 523–536
- Locke, J. C. W., Kozma-Bognár, L., Gould, P. D., Fehér, B., Kevei, E., Nagy, F., Turner, M. S., Hall, A., Millar, A. J. 2006. Experimental validation of a predicted feedback loop in the multi-oscillator clock of *Arabidopsis thaliana*. *Molecular Systems Biology*, 2: doi:10.1038/msb4100102: 6 str.
- Lohse, M., Nagel, A., Herter, T., May, P., Schröda, M., Zrenner, R., Tohge, T., Fernie, A. R., Stitt, M., Usadel, B. 2013. Mercator: A fast and simple web server for genome scale functional annotation of plant sequence data. *Plant, Cell and Environment*, 37, 5: 1250–1258
- Lopez-Romero, P. 2012. Agi4x44Preprocess: 29 str. <http://www.bioconductor.org/packages/2.13/bioc/manuals/Agi4x44PreProcess/man/Agi4x44PreProcess.pdf> (junij 2015)
- Ma, L., Chen, C., Liu, X., Jiao, Y., Su, N., Li, L., Wang, X., Cao, M., Sun, N., Zhang, X., Bao, J., Li, J., Pedersen, S., Bolund, L., Zhao, H., Yuan, L., Gane, K. S. W., Wang, J., Xing, W. D., Wang, J. 2005. A microarray analysis of the rice transcriptome and its comparison to *Arabidopsis*. *Genome Research*, 15, 9: 1274–1283
- Ma, S., Bohnert, H. J. 2007. Integration of *Arabidopsis thaliana* stress-related transcript profiles, promoter structures, and cell-specific expression. *Genome Biology*, 8, 4: doi:10.1186/gb-2007-8-4-r49: 22 str.
- Ma, S., Gong, Q., Bohnert, H. J. 2007. An *Arabidopsis* gene network based on the graphical Gaussian model. *Genome Research*, 17, 11: 1614–1625

- Maglott, D. 2004. Entrez Gene: Gene-centered information at NCBI. *Nucleic Acids Research*, 33, D1: D54–D58
- Mäkinen, K., Hafrén, A. 2014. Intracellular coordination of Potyviral RNA functions in infection. *Frontiers in Plant Science*, 5: doi:10.3389/fpls.2014.00110: 12 str.
- Mandadi, K. K., Scholthof, K.-B. G. 2013. Plant immune responses against viruses: How does a virus cause disease? *The Plant Cell*, 25, 5: 1489–1505
- Mao, L., Van Hemert, J. L., Dash, S., Dickerson, J. A. 2009. Arabidopsis gene co-expression network and its functional modules. *BMC Bioinformatics*, 10: doi:10.1186/1471-2105-10-346: 24 str.
- Marbach, D., Prill, R. J., Schaffter, T., Mattiussi, C., Floreano, D., Stolovitzky, G. 2010. Revealing strengths and weaknesses of methods for gene network inference. *Proceedings of the National Academy of Sciences of the United States of America*, 107, 14: 6286–6291
- Massa, A. N., Childs, K. L., Buell, C. R. 2013. Abiotic and biotic stress responses in group Phureja DM1-3 516 R44 as measured through whole transcriptome sequencing. *Plant Genome*, 6, 3: doi:10.3835/plantgenome2013.05.0014: 10 str.
- Matsuda, F., Hirai, M. Y., Sasaki, E., Akiyama, K., Yonekura-Sakakibara, K., Provart, N. J., Sakurai, T., Shimada, Y., Saito, K. 2010. AtMetExpress development: A phytochemical atlas of *Arabidopsis* development. *Plant Physiology*, 152, 2: 566–578
- Matys, V., Kel-Margoulis, O. V., Fricke, E., Liebich, I., Land, S., Barre-Dirrie, A., Reuter, I., Chekmenev, D., Krull, M., Hornischer, K., Voss, N., Stegmaier, P., Lewicki-Potapov, B., Saxel, H., Kel, a E., Wingender, E. 2006. TRANSFAC and its module TRANSCompel: Transcriptional gene regulation in eukaryotes. *Nucleic Acids Research*, 34: 108–110
- May, P., Wienkoop, S., Kempa, S., Usadel, B., Christian, N., Rupprecht, J., Weiss, J., Recuenco-Munoz, L., Ebenhöh, O., Weckwerth, W., Walther, D. 2008. Metabolomics- and proteomics-assisted genome annotation and analysis of the draft metabolic network of *Chlamydomonas reinhardtii*. *Genetics*, 179, 1: 157–166
- Mazzocchi, F. 2012. Complexity and the reductionism-holism debate in systems biology. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, 4, 5: 413–427
- Milavec, M., Gruden, K., Ravnikar, M., Kovač, M. 2008. Peroxidases in the early responses of different potato cultivars to infection by potato virus Y(^{NTN}). *Plant Pathology*, 57, 5: 861–869
- Miljkovic, D., Stare, T., Mozetic, I., Podpecan, V., Petek, M., Witek, K., Dermastia, M., Lavrac, N., Gruden, K. 2012. Signalling network construction for modelling plant defence response. *PLoS ONE*, 7, 12: doi:10.1371/journal.pone.0051822: 18 str.

- Mochida, K., Shinozaki, K. 2011. Advances in omics and bioinformatics tools for systems analyses of plant functions. *Plant and Cell Physiology*, 52, 12: 2017–2038
- Moreau, Y., Tranchevent, L.-C. 2012. Computational tools for prioritizing candidate genes: Boosting disease gene discovery. *Nature Reviews Genetics*, 13, 8: 523–536
- Moreno-Risueno, M. A., Busch, W., Benfey, P. N. 2010. Omics meet networks - using systems approaches to infer regulatory networks in plants. *Current Opinion in Plant Biology*, 13, 2: 126–131
- Motion, G. B., Amaro, T. M. M. M., Kulagina, N., Huitema, E. 2015. Nuclear processes associated with plant immunity and pathogen susceptibility. *Briefings in Functional Genomics*: doi:10.1093/bfgp/elv013: 10 str.
- Motta, S., Pappalardo, F. 2013. Mathematical modeling of biological systems. *Briefings in Bioinformatics*, 14, 4: 411–422
- Moury, B., Charron, C., Janzac, B., Simon, V., Gallois, J. L., Palloix, A., Caranta, C. 2013. Evolution of plant eukaryotic initiation factor 4E (eIF4E) and Potyvirus genome-linked protein (VPg): A game of mirrors impacting resistance spectrum and durability. *Infection, Genetics and Evolution*, 27: 472–480
- Moury, B., Janzac, B., Ruellan, Y., Simon, V., Ben Khalifa, M., Fakhfakh, H., Fabre, F., Palloix, A. 2014. Interaction patterns between potato virus Y and eIF4E-mediated recessive resistance in the Solanaceae. *Journal of Virology*, 88, 17: 9799–9807
- Movahedi, S., Van de Peer, Y., Vandepoele, K. 2011. Comparative network analysis reveals that tissue specificity and gene function are important factors influencing the mode of expression evolution in *Arabidopsis* and rice. *Plant Physiology*, 156, 3: 1316–1330
- Mukhtar, M. S., Carvunis, A.-R., Dreze, M., Epple, P., Steinbrenner, J., Moore, J., Tasan, M., Galli, M., Hao, T., Nishimura, M. T., Pevzner, S. J., Donovan, S. E., Ghamsari, L., Santhanam, B., Romero, V., Poulin, M. M., Gebreab, F., Gutierrez, B. J., Tam, S., Monachello, D., Boxem, M., Harbort, C. J., McDonald, N., Gai, L., Chen, H., He, Y., Vandenhaute, J., Roth, F. P., Hill, D. E., Ecker, J. R., Vidal, M., Beynon, J., Braun, P., Dangl, J. L. 2011. Independently evolved virulence effectors converge onto hubs in a plant immune system network. *Science*, 333, 6042: 596–601
- Mur, L. A. J., Kenton, P., Lloyd, A. J., Ougham, H., Prats, E. 2008. The hypersensitive response; The centenary is upon us but how much do we know? *Journal of Experimental Botany*, 59, 3: 501–520
- Muraro, D., Byrne, H. M., King, J. R., Bennett, M. J. 2012. Mathematical modelling plant signalling networks. *Mathematical Modelling of Natural Phenomena*, 7, 2: 32–48

- Mustroph, A., Lee, S. C., Oosumi, T., Zanetti, M. E., Yang, H., Ma, K., Yaghoubi-Masihi, A., Fukao, T., Bailey-Serres, J. 2010. Cross-kingdom comparison of transcriptomic adjustments to low-oxygen stress highlights conserved and plant-specific responses. *Plant Physiology*, 152, 3: 1484–1500
- Mutwil, M., Usadel, B., Schütte, M., Loraine, A., Ebenhöh, O., Persson, S. 2010. Assembly of an interactive correlation network for the *Arabidopsis* genome using a novel heuristic clustering algorithm. *Plant Physiology*, 152, 1: 29–43
- Nagasaki, M., Saito, A., Jeong, E., Li, C., Kojima, K., Ikeda, E., Miyano, S. 2010. Cell illustrator 4.0: A computational platform for systems biology. *In Silico Biology*, 10, 1-2: 5–26
- Nakahara, K. S., Masuta, C. 2014. Interaction between viral RNA silencing suppressors and host factors in plant immunity. *Current Opinion in Plant Biology*, 20: 88–95
- Naseem, M., Philippi, N., Hussain, A., Wangorsch, G., Ahmed, N., Dandekar, T. 2012. Integrated Systems View on networking by hormones in *Arabidopsis* immunity reveals multiple crosstalk for cytokinin. *The Plant Cell*, 24, 5: 1793–1814
- Needleman, S. B., Wunsch, C. D. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48, 3: 443–453
- Newman, M. E. J. 2004. Power laws, Pareto distributions and Zipf's law. *Contemporary Physics*, 46, 5: 323–351
- Noble, D. 2002. Modeling the heart – from genes to cells to the whole organ. *Science*, 295, 5560: 1678–1682
- Noble, D. 2006. The music of life: Biology beyond the genome. Oxford University Press, USA: 168 str.
- O'Malley, M. A., Dupré, J. 2005. Fundamental issues in systems biology. *BioEssays*, 27, 12: 1270–1276
- Oksman-Caldentey, K.-M., Saito, K. 2005. Integrating genomics and metabolomics for engineering plant metabolic pathways. *Current Opinion in Biotechnology*, 16, 2: 174–179
- Pacheco, R., García-Marcos, A., Manzano, A., de Lacoba, M. G., Camañes, G., García-Agustín, P., Díaz-Ruiz, J. R., Tenllado, F. 2012. comparative analysis of transcriptomic and hormonal responses to compatible and incompatible plant-virus interactions that lead to cell death. *Molecular Plant-Microbe Interactions*, 25, 5: 709–723

- Padmanabhan, M. S., Dinesh-Kumar, S. P. 2010. All hands on deck – the role of chloroplasts, endoplasmic reticulum, and the nucleus in driving plant innate immunity. *Molecular Plant-Microbe Interactions*, 23, 11: 1368–1380
- Palaniswamy, S. K., James, S., Sun, H., Lamb, R. S., Davuluri, R. V., Grotewold, E. 2006. AGRIS and AtRegNet. A platform to link cis-regulatory elements and transcription factors into regulatory networks. *Plant Physiology*, 140, 3: 818–829
- Palukaitis, P., Groen, S. C., Carr, J. P. 2013. The Rumsfeld paradox: Some of the things we know that we don't know about plant virus infection. *Current Opinion in Plant Biology*, 16, 4: 513–519
- Patel, R. V., Nahal, H. K., Breit, R., Provart, N. J. 2012. BAR expressolog identification: Expression profile similarity ranking of homologous genes in plant species. *The Plant Journal*, 71, 6: 1038–1050
- Penfold, C. A., Buchanan-Wollaston, V., Denby, K. J., Wild, D. L. 2012. Nonparametric Bayesian inference for perturbed and orthologous gene regulatory networks. *Bioinformatics*, 28, 12: 233–241
- Penfold, C. A., Wild, D. L. 2011. How to infer gene networks from expression profiles, revisited. *Interface Focus*, 1, 6: 857–870
- Peng, X., Chan, E. Y., Li, Y., Diamond, D. L., Korth, M. J., Katze, M. G. 2009. Virus-host interactions: From systems biology to translational research. *Current Opinion in Microbiology*, 12, 4: 432–438
- Pieterse, C. M. J., Leon-Reyes, A., Van der Ent, S., Van Wees, S. C. M. 2009. Networking by small-molecule hormones in plant immunity. *Nature Chemical Biology*, 5, 5: 308–316
- Piron, P. G. M. 1986. New aphid vectors of potato virus Y^N. *Netherlands Journal of Plant Pathology*, 92, 5: 223–229
- Podpečan, V., Lavrač, N., Mozetič, I., Novak, P. K., Trajkovski, I., Langohr, L., Kulovesi, K., Toivonen, H., Petek, M., Motaln, H., Gruden, K. 2011. SegMine workflows for semantic microarray data analysis in Orange4WS. *BMC Bioinformatics*, 12: doi:10.1186/1471-2105-12-416: 16 str.
- Pokhilko, A., Fernández, A. P., Edwards, K. D., Southern, M. M., Halliday, K. J., Millar, A. J. 2012. The clock gene circuit in *Arabidopsis* includes a repressilator with additional feedback loops. *Molecular Systems Biology*, 8: doi:10.1038/msb.2012.6: 13 str.

- Pompe-Novak, M., Gruden, K., Baebler, Š., Krečič-Stres, H., Kovač, M., Jongsma, M., Ravnikar, M. 2006. Potato virus Y induced changes in the gene expression of potato (*Solanum tuberosum* L.). *Physiological and Molecular Plant Pathology*, 67, 3-5: 237–247
- Postnikova, O. A., Nemchinov, L. G. 2012. Comparative analysis of microarray data in *Arabidopsis* transcriptome during compatible interactions with plant viruses. *Virology Journal*, 9: doi:10.1186/1743-422X-9-101: 12 str.
- Powell, S., Forslund, K., Szklarczyk, D., Trachana, K., Roth, A., Huerta-Cepas, J., Gabaldón, T., Rattei, T., Creevey, C., Kuhn, M., Jensen, L. J., von Mering, C., Bork, P. 2014. eggNOG v4.0: Nested orthology inference across 3686 organisms. *Nucleic Acids Research*, 42: 231–239
- Proost, S., Fostier, J., De Witte, D., Dhoedt, B., Demeester, P., Van De Peer, Y., Vandepoele, K. 2012. i-ADHoRe 3.0: Fast and sensitive detection of genomic homology in extremely large data sets. *Nucleic Acids Research*, 40, 2: doi:10.1093/nar/gkr955: 11 str.
- Proost, S., Van Bel, M., Vaneechoutte, D., Van de Peer, Y., Inzé, D., Mueller-Roeber, B., Vandepoele, K. 2014. PLAZA 3.0: An access point for plant comparative genomics. *Nucleic Acids Research*, 43: 974–981
- Quackenbush, J., Cho, J., Lee, D., Liang, F., Holt, I., Karamycheva, S., Parvizi, B., Pertea, G., Sultana, R., White, J. 2001. The TIGR gene indices: Analysis of gene transcript sequences in highly sampled eukaryotic species. *Nucleic Acids Research*, 29, 1: 159–164
- Quenouille, J., Vassilakos, N., Moury, B. 2013. Potato virus Y: A major crop pathogen that has provided major insights into the evolution of viral pathogenicity. *Molecular Plant Pathology*, 14, 5: 439–452
- Quinlan, J. R. 1994. C4.5: Programs for machine learning. *Machine Learning*, 16: 235–240
- Rajamäki, M.-L., Valkonen, J. P. T. 2009. Control of nuclear and nucleolar localization of nuclear inclusion protein a of picorna-like potato virus A in *Nicotiana* species. *The Plant Cell*, 21, 8: 2485–2502
- Ramsay, J. O., Hooker, G., Campbell, D., Cao, J. 2007. Parameter estimation for differential equations: A generalized smoothing approach. *Journal of the Royal Statistical Society, Series B*, 69, 5: 741–770
- Ramšak, Ž., Baebler, Š., Rotter, A., Korbar, M., Mozetic, I., Usadel, B., Gruden, K. 2014. GoMapMan: Integration, consolidation and visualization of plant gene annotations within the MapMan ontology. *Nucleic Acids Research*, 42, D1: 1167–1175

- Rentzsich, R., Orengo, C. A. 2009. Protein function prediction - the power of multiplicity. *Trends in Biotechnology*, 27, 4: 210–219
- Riedel, D., Lesemann, D. E., Maiss, E. 1998. Ultrastructural localization of nonstructural and coat proteins of 19 Potyviruses using antisera to bacterially expressed proteins of plum pox Potyvirus. *Archives of Virology*, 143, 11: 2133–2158
- Robert-Seilaniantz, A., Grant, M., Jones, J. D. G. 2011. Hormone crosstalk in plant disease and defense: More Than just jasmonate-salicylate antagonism. *Annual Review of Phytopathology*, 49: 317–343
- Rodrigo, G., Carrera, J., Ruiz-Ferrer, V., del Toro, F. J., Llave, C., Voinnet, O., Elena, S. F. 2012. A meta-analysis reveals the commonalities and differences in *Arabidopsis thaliana* response to different viral pathogens. *PLoS ONE*, 7, 7: doi:10.1371/journal.pone.0040526: 14 str.
- Rogers, S., Khanin, R., Girolami, M. 2007. Bayesian model-based inference of transcription factor activity. *BMC Bioinformatics*, 8, Suppl. 2: doi:10.1186/1471-2105-8-S2-S2: 11 str.
- Roossinck, M. J. 2011. Plant virus metagenomics: Biodiversity and ecology. *Annual Review of Genetics*, 46: 359–369
- Rosenfeld, N., Elowitz, M. B., Alon, U. 2002. Negative autoregulation speeds the response times of transcription networks. *Journal of Molecular Biology*, 323, 5: 785–793
- Rost, B. 1999. Twilight zone of protein sequence alignments. *Protein Engineering*, 12, 2: 85–94
- Rotter, A., Hren, M., Baebler, S., Blejec, A., Gruden, K. 2008. Finding differentially expressed genes in two-channel DNA microarray datasets – How to increase reliability of data preprocessing. *OMICS: A Journal Of Integrative Biology*, 12, 3: 171–182
- Rotter, A., Usadel, B., Baebler, S., Stitt, M., Gruden, K. 2007. Adaptation of the MapMan ontology to biotic stress responses: Application in Solanaceous species. *Plant Methods*, 3: doi:10.1186/1746-4811-3-10: 9 str.
- Roy, S. 2014. Perspectives in systems biology. *Systems and Synthetic Biology*, 8: 1–2
- Ruepp, A., Zollner, A., Maier, D., Albermann, K., Hani, J., Mokrejs, M., Tetko, I., Güldener, U., Mannhaupt, G., Münsterkötter, M., Mewes, H. W. 2004. The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes. *Nucleic Acids Research*, 32, 18: 5539–5545

- Saddic, L. A., Huvermann, B., Bezhani, S., Su, Y., Winter, C. M., Kwon, C. S., Collum, R. P., Wagner, D. 2006. The LEAFY target LMI1 is a meristem identity regulator and acts together with LEAFY to regulate expression of CAULIFLOWER. *Development*, 133, 9: 1673–1682
- Sakurai, T., Yamada, Y., Sawada, Y., Matsuda, F., Akiyama, K., Shinozaki, K., Hirai, M. Y., Saito, K. 2013. PRIME Update: Innovative content for plant metabolomics and integration of gene expression and metabolite accumulation. *Plant and Cell Physiology*, 54, 2: doi:10.1093/pcp/pcs184: 8 str.
- Scardoni, G., Petterlini, M., Laudanna, C. 2009. Analyzing biological network parameters with CentiScaPe. *Bioinformatics*, 25, 21: 2857–2859
- Scholthof, K.-B. G., Adkins, S., Czosnek, H., Palukaitis, P., Jacquot, E., Hohn, T., Hohn, B., Saunders, K., Candresse, T., Ahlquist, P., Hemenway, C., Foster, G. D. 2011. Top 10 plant viruses in molecular plant pathology. *Molecular Plant Pathology*, 12, 9: 938–954
- Schulz, M. H., Kohler, S., Bauer, S., Robinson, P. N. 2011. Exact score distribution computation for ontological similarity searches. *BMC Bioinformatics*, 12: doi:10.1186/1471-2105-12-441: 12 str.
- Schwessinger, B., Ronald, P. C. 2012. Plant innate immunity: Perception of conserved microbial signatures. *Annual Review of Plant Biology*, 63: 451–482
- Seo, Y. S., Chern, M., Bartley, L. E., Han, M., Jung, K. H., Lee, I., Walia, H., Richter, T., Xu, X., Cao, P., Bai, W., Ramanan, R., Amonpant, F., Arul, L., Canlas, P. E., Ruan, R., Park, C. J., Chen, X., Hwang, S., Jeon, J. S., Ronald, P. C. 2011. Towards establishment of a rice stress response interactome. *PLoS Genetics*, 7, 4: doi:10.1371/journal.pgen.1002020: 12 str.
- Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., Ideker, T. 2003. Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Research*, 13, 11: 2498–2504
- Shen, W., Yan, P., Gao, L., Pan, X., Wu, J., Zhou, P. 2010. Helper component-proteinase (HC-Pro) protein of Papaya ringspot virus interacts with papaya calreticulin. *Molecular Plant Pathology*, 11, 3: 335–346
- Shrestha, R., Matteis, L., Skofic, M., Portugal, A., McLaren, G., Hyman, G., Arnaud, E. 2012. Bridging the phenotypic and genetic data useful for integrated breeding through a data annotation using the Crop Ontology developed by the crop communities of practice. *Frontiers in Physiology*, 3: doi:10.3389/fphys.2012.00326: 10 str.

- Sierro, N., Battey, J. N. D., Ouadi, S., Bakaher, N., Bovet, L., Willig, A., Goepfert, S., Peitsch, M. C., Ivanov, N. V. 2014. The tobacco genome sequence and its comparison with those of tomato and potato. *Nature Communications*, 5: doi:10.1038/ncomms4833: 9 str.
- Smedley, D., Haider, S., Ballester, B., Holland, R., London, D., Thorisson, G., Kasprzyk, A. 2009. BioMart - Biological queries made easy. *BMC Genomics*, 10: doi:10.1186/1471-2164-10-22: 12 str.
- Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L. J., Eilbeck, K., Ireland, A., Mungall, C. J., Leontis, N., Rocca-Serra, P., Ruttenberg, A., Sansone, S.-A., Scheuermann, R. H., Shah, N., Whetzel, P. L., Lewis, S. 2007. The OBO Foundry: Coordinated evolution of ontologies to support biomedical data integration. *Nature Biotechnology*, 25, 11: 1251–1255
- Smyth, G. K. 2004. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. V: *Statistical Applications in Genetics and Molecular Biology*, Vol. 3. Stumpf M. P. H. (ur.). Berlin, De Gruyter: 1–26
- Soetaert, K., Petzoldt, T., Setzer, R. W. 2010. Package deSolve : solving initial value differential equations in R. *Journal Of Statistical Software*, 33, 9: 1–25
- Soitamo, A. J., Jada, B., Lehto, K. 2011. HC-Pro silencing suppressor significantly alters the gene expression profile in tobacco leaves and flowers. *BMC Plant Biology*, 11: doi:10.1186/1471-2229-11-68: 16 str.
- Spooner, D. M., Hijmans, R. J. 2001. Potato systematics and germplasm collecting, 1989–2000. *American Journal of Potato Research*, 78, 4: 237–268
- Srivastava, G. P., Li, P., Liu, J., Xu, D. 2010. Identification of transcription factor's targets using tissue-specific transcriptomic data in *Arabidopsis thaliana*. *BMC Systems Biology*, 4, Suppl. 2: doi:10.1186/1752-0509-4-S2-S2: 12 str.
- Stare, T., Ramšak, Ž., Blejec, A., Stare, K., Turnšek, N., Weckwerth, W., Wienkoop, S., Vodnik, D., Gruden, K. 2015. Bimodal dynamics of primary metabolism-related responses in tolerant potato-potato virus Y interaction. *BMC Genomics*, 16: doi:10.1186/s12864-015-1925-2: 17 str.
- Stark, C., Breitkreutz, B.-J., Reguly, T., Boucher, L., Breitkreutz, A., Tyers, M. 2006. BioGRID: A general repository for interaction datasets. *Nucleic Acids Research*, 34: 535–539
- Stegle, O., Denby, K. J., Cooke, E. J., Wild, D. L., Ghahramani, Z., Borgwardt, K. M. 2010. A robust Bayesian two-sample test for detecting intervals of differential gene expression in microarray time series. *Journal of Computational Biology*, 17, 3: 355–367

- Storm, C. E. V., Sonnhammer, E. L. L. 2002. Automated ortholog inference from phylogenetic trees and calculation of orthology reliability. *Bioinformatics* (Oxford, England), 18, 1: 92–99
- Street, N. R., Sjödin, A., Bylesjö, M., Gustafsson, P., Trygg, J., Jansson, S. 2008. A cross-species transcriptomics approach to identify genes involved in leaf development. *BMC Genomics*, 9: doi:10.1186/1471-2164-9-589: 18 str.
- Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., Mesirov, J. P. 2005. Gene Set Enrichment Analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 102, 43: 15545–15550
- Subramanian, N., Torabi-Parizi, P., Gottschalk, R. A., Germain, R. N., Dutta, B. 2015. Network representations of immune system complexity. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, 7, 1: 13–38
- Sun, Q., Zyballov, B., Majeran, W., Friso, G., Olinares, P. D. B., van Wijk, K. J. 2009. PPDB, the plant proteomics database at Cornell. *Nucleic Acids Research*, 37: 969–974
- Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., Simonovic, M., Roth, A., Santos, A., Tsafou, K. P., Kuhn, M., Bork, P., Jensen, L. J., Von Mering, C. 2014. STRING v10: Protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Research*, 43: 447–452
- Tan, S.-L., Ganji, G., Paeper, B., Proll, S., Katze, M. G. 2007. Systems biology and the host response to viral infection. *Nature Biotechnology*, 25, 12: 1383–1389
- Tatusov, R. L., Fedorova, N. D., Jackson, J. D., Jacobs, A. R., Kiryutin, B., Koonin, E. V., Krylov, D. M., Mazumder, R., Mekhedov, S. L., Nikolskaya, A. N., Rao, B. S., Smirnov, S., Sverdlov, A. V., Vasudevan, S., Wolf, Y. I., Yin, J. J., Natale, D. A. 2003. The COG database: An updated version includes eukaryotes. *BMC Bioinformatics*, 4: doi:10.1186/1471-2105-4-41: 14 str.
- The Gene Ontology Consortium. 2000. Gene Ontology: Tool for the unification of biology. *Nature Genetics*, 25: 25–29
- The Gene Ontology Consortium. 2013. Gene Ontology annotations and resources. *Nucleic Acids Research*, 41: 530–535
- The Potato Genome Sequencing Consortium. 2011. Genome sequence and analysis of the tuber crop potato. *Nature*, 475, 7355: 189–195
- The Tomato Genome Consortium. 2012. The tomato genome sequence provides insights into fleshy fruit evolution. *Nature*, 485, 7400: 635–641

- Theocharidis, A., Van Dongen, S., Enright, A. J., Freeman, T. C. 2009. Network visualization and analysis of gene expression data using BioLayout Express(3D). *Nature Protocols*, 4, 10: 1535–1550
- Thimm, O., Blasing, O., Gibon, Y., Nagel, A., Meyer, S., Kruger, P., Selbig, J., Muller, L. A., Rhee, S. Y., Stitt, M. 2004. MAPMAN: A user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. *The Plant Journal*, 37, 6: 914–939
- Trajkovski, I., Lavrac, N., Tolar, J. 2008. SEGS: Search for enriched gene sets in microarray data. *Journal of Biomedical Informatics*, 41, 4: 588–601
- Tsuda, K., Ito, Y., Sato, Y., Kurata, N. 2011. Positive autoregulation of a KNOX gene is essential for shoot apical meristem maintenance in rice. *The Plant Cell*, 23, 12: 4368–4381
- Tsuda, K., Sato, M., Stoddard, T., Glazebrook, J., Katagiri, F. 2009. Network properties of robust immunity in plants. *PLoS Genetics*, 5, 12: doi:10.1371/journal.pgen.1000772: 16 str.
- Tsuda, K., Somssich, I. E. 2015. Transcriptional networks in plant immunity. *New Phytologist*, 206, 3: 932–947
- Tuikkala, J., Vähämaa, H., Salmela, P., Nevalainen, O. S., Aittokallio, T. 2012. A multilevel layout algorithm for visualizing physical and genetic interaction networks, with emphasis on their modular organization. *BioData Mining*, 5: doi:10.1186/1756-0381-5-2: 16 str.
- Tully, J. P., Hill, A. E., Ahmed, H. M. R., Whitley, R., Skjellum, A., Mukhtar, M. S. 2014. Expression-based network biology identifies immune-related functional modules involved in plant defense. *BMC Genomics*, 15: doi:10.1186/1471-2164-15-421: 14 str.
- Tyson, J. J., Chen, K., Novak, B. 2001. Network dynamics and cell physiology. *Nature Reviews: Molecular Cell Biology*, 2, 12: 908–916
- Tzfadia, O., Amar, D., Bradbury, L. M. T., Wurtzel, E. T., Shamir, R. 2012. The MORPH algorithm: Ranking candidate genes for membership in *Arabidopsis* and tomato pathways. *The Plant Cell*, 24, 11: 4389–4406
- Ubeda-Tomás, S., Beemster, G. T. S., Bennett, M. J. 2012. Hormonal regulation of root growth: Integrating local activities into global behaviour. *Trends in Plant Science*, 17, 6: 326–331

- Van der Does, D., Leon-Reyes, A., Koornneef, A., Van Verk, M. C., Rodenburg, N., Pauwels, L., Goossens, A., Körbes, A. P., Memelink, J., Ritsema, T., Van Wees, S. C. M., Pieterse, C. M. J. 2013. Salicylic acid suppresses jasmonic acid signaling downstream of SCFCOI1-JAZ by targeting GCC promoter motifs via transcription factor ORA59. *The Plant Cell*, 25, 2: 744–761
- Van Dongen, S. M. 2000. Graph clustering by flow simulation. Doktorska disertacija. Utrecht, Utrecht University, Faculteit Wiskunde en Informatica.
- Van Loon, L. C., Van Strien, E. A. 1999. The families of pathogenesis-related proteins, their activities, and comparative analysis of PR-1 type proteins. *Physiological and Molecular Plant Pathology*, 55: 85–97
- Vernon, I., Goldstein, M., Bowerz, R. G. 2010. Galaxy formation: A Bayesian uncertainty analysis. *Bayesian Analysis*, 5, 4: 619–669
- Vidal, M., Cusick, M. E., Barabási, A.-L. 2011. Interactome networks and human disease. *Cell*, 144, 6: 986–998
- Wall, D. P., Deluca, T. 2007. Ortholog detection using the reciprocal smallest distance algorithm. V: *Methods in Molecular Biology* Vol. 396. Walker J. M. (ur.). New York, Springer: 95-110
- Wang, D., Amornsiripanitch, N., Dong, X. 2006. A genomic approach to identify regulatory nodes in the transcriptional network of systemic acquired resistance in plants. *PLoS Pathogens*, 2, 11: 1042–1050
- Wang, R. S., Wang, Y., Zhang, X. S., Chen, L. 2007. Inferring transcriptional regulatory networks from high-throughput data. *Bioinformatics*, 23, 22: 3056–3064
- Watts, D. J., Strogatz, S. H. 1998. Collective dynamics of “small-world” networks. *Nature*, 393, 6684: 440–442
- Weßling, R., Epple, P., Altmann, S., He, Y., Yang, L., Henz, S. R., McDonald, N., Wiley, K., Bader, K. C., Gläßer, C., Mukhtar, M. S., Haigis, S., Ghamsari, L., Stephens, A. E., Ecker, J. R., Vidal, M., Jones, J. D. G., Mayer, K. F. X., Ver Loren van Themaat, E., Weigel, D., Schulze-Lefert, P., Dangl, J. L., Panstruga, R., Braun, P. 2014. convergent targeting of a common host protein-network by pathogen effectors from three kingdoms of life. *Cell Host & Microbe*, 16, 3: 364–375
- Williams, T. C. R., Poolman, M. G., Howden, A. J. M., Schwarzlander, M., Fell, D. A., Ratcliffe, R. G., Sweetlove, L. J. 2010. A genome-scale metabolic model accurately predicts fluxes in central carbon metabolism under stress conditions. *Plant Physiology*, 154, 1: 311–323
- Williamson, D., Goldstein, M., Blaker, A. 2012. Fast linked analyses for scenario-based hierarchies. *Journal of the Royal Statistical Society, Series C*, 61, 5: 665–691

- Windram, O., Madhou, P., McHattie, S., Hill, C., Hickman, R. J., Cooke, E. J., Jenkins, D. J., Penfold, C. A., Baxter, L., Breeze, E., Kiddle, S. J., Rhodes, J., Atwell, S., Kim, Y., Stegle, O., Borgwardt, K. M., Zhang, C., Tabrett, A., Legaie, R., Moore, J. D., Finkenstädt, B., Wild, D. L., Mead, A., Rand, D. A., Ott, S., Buchanan-Wollaston, V., Denby, K. J. 2012. Arabidopsis defense against *Botrytis cinerea*: Chronology and regulation deciphered by high-resolution temporal transcriptomic analysis. *Plant Cell*, 24, 9: 3530–3557
- Windram, O., Penfold, C. A., Denby, K. J. 2014. Network modeling to understand plant immunity. *Annual Review of Phytopathology*, 52: 93–111
- Wise, R. P., Moscou, M. J., Bogdanove, A. J., Whitham, S. a. 2007. Transcript profiling in host-pathogen interactions. *Annual Review of Phytopathology*, 45: 329–369
- Wittkop, T., Teravest, E., Evani, U. S., Fleisch, K. M., Berman, A. E., Powell, C., Shah, N. H., Mooney, S. D. 2013. STOP Using just GO: A multi-ontology hypothesis generation tool for high throughput experimentation. *BMC Bioinformatics*, 14: doi:10.1186/1471-2105-14-53: 10 str.
- Wolkenhauer, O. 2001. Systems biology: The reincarnation of systems theory applied in biology? *Briefings in Bioinformatics*, 2, 3: 258–270
- Yang, H., Huang, Y., Zhi, H., Yu, D. 2011. Proteomics-based analysis of novel genes involved in response toward soybean mosaic virus infection. *Molecular Biology Reports*, 38, 1: 511–521
- Yang, K., Zhang, L. 2008. Performance comparison of gene family clustering methods with expert curated gene family data set in *Arabidopsis thaliana*. *Planta*, 228, 3: 439–447
- Yuan, B. Z., Nishiyama, S., Kang, Y. 2003. Effects of different irrigation regimes on the growth and yield of drip-irrigated potato. *Agricultural Water Management*, 63, 3: 153–167
- Zeilinger, M. N., Farre, E. M., Taylor, S. R., Kay, S. A., Doyle, F. J. 2006. A novel computational model of the circadian clock in Arabidopsis that incorporates PRR7 and PRR9. *Molecular Systems Biology*, 2: doi:10.1038/msb4100101: 13 str.
- Zhang, Z., Yu, J., Li, D., Zhang, Z., Liu, F., Zhou, X., Wang, T., Ling, Y., Su, Z. 2010. PMRD: Plant microRNA database. *Nucleic Acids Research*, 38: 806–813
- Zheng, Z.-L., Zhao, Y. 2013. Transcriptome comparison and gene coexpression network analysis provide a systems view of citrus response to “*Candidatus Liberibacter asiaticus*” infection. *BMC Genomics*, 14: doi:10.1186/1471-2164-14-27: 21 str.
- Zhou, X., Wang, G., Zhang, W. 2007. UV-B responsive microRNA genes in *Arabidopsis thaliana*. *Molecular Systems Biology*, 3: doi:10.1038/msb4100143: 10 str.

- Zipfel, C., Kunze, G., Chinchilla, D., Caniard, A., Jones, J. D. G., Boller, T., Felix, G. 2006. Perception of the bacterial PAMP EF-Tu by the receptor EFR restricts Agrobacterium-mediated transformation. *Cell*, 125, 4: 749–760
- Zmasek, C. M., Eddy, S. R. 2002. RIO: Analyzing proteomes by automated phylogenomics using resampled inference of orthologs. *BMC Bioinformatics*, 3: doi:10.1186/1471-2105-3-14: 19 str.
- Zvereva, A. S., Pooggin, M. M. 2012. Silencing and innate immunity in plant defense against viral and non-viral pathogens. *Viruses*, 4: 2578–2597
- Žitnik, M., Zupan, B. 2014. Gene network inference by probabilistic scoring of relationships from a factorized model of interactions. *Bioinformatics*, 30, 12: 246–254

ZAHVALA

"*It takes a village ...*"

S tem ne želim implicirati, da je doktorsko delo po težavnosti podobno vzgoji otroka, drži pa, da vas je za nastanek tega dela zaslužnih veliko.

Kristina, hvala za vse dane nasvete in vodstvo, vse pogovore (tako profesionalne kot humorne) ter za učenje plavanja v znanstvenih vodah po tvojem *baby steps* sistemu, ob učenju veščin mreženja predvsem (hopefully this gullibility of mine will not go away and your system keeps being efficient).

Najlepša hvala članom komisije za njihov *peer review* z vsemi danimi komentarji in vprašanji, ki niso samo bolje izoblikovali in izboljšali trenutnega doktorskega dela, ampak bodo služili za naprej.

Ob prehodu na člane *fito* oddelka bi se najprej zahvalila Maji in Jani, ki s Kristino tvorita vodstveno trojico, ki oddelek vodi v teh vodah odličnosti in kvalitete, obenem pa še vedno ohranja ta sproščeni duh potreben za uspešno delovanje.

Navdih imena naslednje množice (ang. *cluster*) sem poimenovala po filmu iz leta mojega rojstva, *the lunch club* – my fellow niñas m. Anna in Tjaša, Anja, Ion, Denis in Nejc – without our *conversaciones banales*, the world would be a much duller place to live in. Larisa, AnaM in Nataša – brez medsebojno sinhroniziranih nevronskih mrež, nabritega smisla za humor in delovne pripravljenosti na pogostno *tmv* izpostavljenost, for welcome leisure-related activities, that brighten the day to day existence.

Podskupini omičarjev, *the geeky bioinformaticians* – my dear fellow Joint First Author Špeli, AnaR in MarkoP – living proof, that cooperation and sometimes impassioned debates do promote synergies and successes. Prof. Blejec Andreju za pomoč in matematičnih, statističnih in vsakodnevnih tematik pogovora (lokalni ribiči, ki podajo začetne vrednosti verjetnosti Palomares incidenta je poseben primer združka vseh treh tem). Enchu za pomoč pri spoznavanju moči root dostopa v sistemu, poučevanja računalniških konceptov biotehnologinji in vso s smehom prežeto komunikacijo ob primerinem kofeinskem napitku.

Najožjemu DŠIWSO skupku – Dejanu, Špeli, Tjaši – pa tudi razširitvi – Aleksandri, Klavdiji, Mojci in Tini – sometimes it is good to have a common denominator to unite us in jokes! Cene, da še kdaj skušala reševati svet preko debat o načinu ekološkega varčevanja z lesnimi dobrinami. Manci za zaupanje z oskrbo računalnikov; Tanji za druženje z istim tipom pogleda na smešnost sveta; Lidiji za nov razvoj na nivoju opazovanja – dnevne statistike števila objemov; Vesni za razumevanje ob razbitju steklovine (pa če prinaša srečo). Vsem ostalim fitovcem in fitovkam, preteklim in trenutnim, katerih unija naredi naš oddelek to kar je.

Sodelavcem iz Odseka za tehnologije znanja Instituta Jožef Stefan: Nadi, Igorju, Vidu in Dragani – za prijetno sodelovanje pri razvoju večih aplikacij. To Michael Goldstein and Ian Vernon for kindly accepting me to Durham University and dispelling of the fears of differential equations.

Anki, Iztoku in Doroteji za deljene jutranje kave in pogovore; elemelonama Roku (*light of my life*) in Mitji za popoldanske kave in prihodnostni tarok. Skupini ISS-Piran – Ajdi, Andreju, Drejcu, Mojci in Urški – the PhD bench is slowly and steadily increasing, while spam patiently awaits. CETO skupini – AniK, Neži, Moniki, Matjažu, MarkuP – da se nam še kdaj uspe uskladiti pod ikono zlate gazele.

Last, but not least – družini – ati, mami in Matjaž, za vso podporo, potrpljenje, oskrbo in pomoč od malih nog, preko pubertetniških časov, do danes.

PRILOGE

Priloga A: Seznam vseh reakcij nadgrajenega modela rastlinske imunske signalizacije (PIS).

V prvem stolpcu so rastlinski hormoni označeni s kraticami: etilen ET, jasmonska kislina JA, salicilna kislina SA. V drugem stolpcu so opisani tipi reakcij glede na zastavljenou ontologijo reakcij (Slika 14, str. 54): A pomeni aktivacijo, I inhibicijo in V vezavo. Zapis reakcije v tretjem stolpcu je odvisen od tipa reakcije, v zadnjem stolpcu pa S pomeni reakcijo iz osnovne objave Miljkovic in sod. (2012), N pa novo vključeno reakcijo.

Appendix A: All reactions of the upgraded plant immune signalisation model (PIS).

First column contains the process type descriptions, with the following hormone abbreviations: ethylene ET, jasmonic acid JA, salicylic acid SA and auxins AUX. Second column describes the reaction type as defined by the reaction ontology (Figure 14, pp.54): A denotes activation, I inhibitory and B binding reactions. The reaction notation of the third column depends on the reaction type, while the last column S denotes a reaction from the Miljkovic in sod. (2012) publication and N a newly added reaction.

Proces	Tip reakcije	Reakcija	Vir
ET (signalizacija)	[A] translokacija	Cu > Cu(ER); HMA	S
ET (signalizacija)	[A] proteinska aktivacija	ETR(i) > ETR(a); Cu(ER)	S
ET (signalizacija)	[V] proteinska vezava	ETR(a) + CTR(a) = ETR(a)-CTR(a)	S
ET (signalizacija)	[A] proteinska aktivacija	RTE1 [activates] CTR	N
ET (signalizacija)	[V] vezava metabolita in proteina	ETR(a) + ET = ETR(a)-ET	S
ET (signalizacija)	[A] (de)fosforilacija	EIN2 [activates] EIN2	S
ET (signalizacija)	[I] (de)fosforilacija	CTR [inhibits] EIN2	S
ET (signalizacija)	[V] proteinska vezava	SCF + ETP = SCF-ETP	N
ET (signalizacija)	[A] (de)fosforilacija	ETP [activates] EIN2	N
ET (signalizacija)	[I] translacija/transkripcija	EIN5 [inhibits] EBF	S
ET (signalizacija)	[A] translacija/transkripcija	EIN3(like) [activates] EBF	S
ET (signalizacija)	[I] razpad	ET [degrades] EBF	N
ET (signalizacija)	[A] translacija/transkripcija	EIN2(a) [activates] EIN3(like)	S
ET (signalizacija)	[V] proteinska vezava	SCF + EBF = SCF-EBF	S
ET (signalizacija)	[I] razpad	EBF [degrades] EIN3(like)	S
ET (signalizacija)	[A] translacija/transkripcija	EIN3(like) [activates] ERF/EDF	S
ET (signalizacija)	[A] translacija/transkripcija	EIN3(like) [activates] PR3/4	S
ET (signalizacija)	[A] translacija/transkripcija	ERF/EDF [activates] PDF1.2	S
ET (sinteza)	[A] kataliza	L-Met > SAMe; SAM	S
ET (sinteza)	[A] kataliza	SAMe > ACC; ACS	S
ET (sinteza)	[A] kataliza	ACC > ET; ACO	S
JA (signalizacija)	[V] proteinska vezava	SCF + COI1 = SCF-COI1	S
JA (signalizacija)	[V] vezava metabolita in proteina	SCF-COI1 + JA-Ile = SCF-COI1-JA-Ile	S
JA (signalizacija)	[I] razpad	COI1 [degrades] JAZ	N
JA (signalizacija)	[A] translacija/transkripcija	X2 [activates] JAZ	S
JA (signalizacija)	[A] translacija/transkripcija	MYC [activates] JAZ	S
JA (signalizacija)	[V] proteinska vezava	JAZ + MYC = JAZ-MYC	S
JA (signalizacija)	[I] translacija/transkripcija	JAM [inhibits] MYC	N
JA (signalizacija)	[A] translacija/transkripcija	MYC [activates] CLH	S
JA (signalizacija)	[A] translacija/transkripcija	MYC [activates] JR1	S
JA (signalizacija)	[A] translacija/transkripcija	MYC [activates] PR13	S
JA (signalizacija)	[A] translacija/transkripcija	MYC [activates] VSP	S

se nadaljuje

nadaljevanje Priloge A: Seznam vseh reakcij nadgrajenega modela rastlinske imunske signalizacije (PIS).

Proces	Tip reakcije	Reakcija	Vir
JA (signalizacija)	[I] translacija/transkripcija	JAZ [inhibits] CLH	S
JA (signalizacija)	[I] translacija/transkripcija	JAZ [inhibits] JR1	S
JA (signalizacija)	[I] translacija/transkripcija	JAZ [inhibits] PR13	S
JA (signalizacija)	[I] translacija/transkripcija	JAZ [inhibits] VSP	S
JA (signalizacija)	[A] translacija/transkripcija	WRKY33 [activates] CLH	N
JA (signalizacija)	[A] translacija/transkripcija	WRKY33 [activates] JR1	N
JA (signalizacija)	[A] translacija/transkripcija	WRKY33 [activates] PR13	N
JA (signalizacija)	[A] translacija/transkripcija	WRKY33 [activates] VSP	N
JA (signalizacija)	[virus]	CLH [inhibits] virus	N
JA (signalizacija)	[virus]	JR1 [inhibits] virus	N
JA (signalizacija)	[virus]	PR13 [inhibits] virus	N
JA (signalizacija)	[virus]	VSP [inhibits] virus	N
JA (sinteza)	[A] kataliza	PUFA:LA > 13-HPOT; LOX	S
JA (sinteza)	[A] kataliza	13-HPOT > 12,13-EOT; AOS	S
JA (sinteza)	[A] kataliza	12,13-EOT > OPDA(c); AOC	S
JA (sinteza)	[A] translokacija	OPDA(c) > OPDA(p); CTS	S
JA (sinteza)	[A] kataliza	OPDA(p) > OPC8; OPR	S
JA (sinteza)	[A] kataliza	OPC8 > OPC8-CoA; OPCL	S
JA (sinteza)	[A] kataliza	OPC8-CoA > OPC6-CoA; ACX	S
JA (sinteza)	[A] kataliza	OPC6-CoA > OPC4-CoA; MFP	S
JA (sinteza)	[A] kataliza	OPC4-CoA > JA-CoA; KAT	S
JA (sinteza)	[A] kataliza	JA-CoA > JA(p); ACH	N
JA (sinteza)	[A] translokacija	JA(p) > JA; X4	S
JA (sinteza)	[A] kataliza	JA > MeJA; JMT	S
JA (sinteza)	[A] kataliza	JA > JA-Ile; JAR	S
SA (signalizacija)	[A] translacija/transkripcija	EDS5 [activates] ICS	S
SA (signalizacija)	[V] proteinska vezava	EDS1 + PAD = EDS1-PAD	N
SA (signalizacija)	[I] proteinske aktivacije	NPR [inhibits] EDS1-PAD	S
SA (signalizacija)	[A] translacija/transkripcija	EDS1-PAD [activates] EDS5	S
SA (signalizacija)	[V] proteinska vezava	NPR > NPR-oligo; GSNO	N
SA (signalizacija)	[A] translacija/transkripcija	SA [activates] TRX-H	N
SA (signalizacija)	[I] razpad	NPR-oligo > NPR; TRX-H	S
SA (signalizacija)	[A] translokacija	NPR > NPR-nuc; MOS	N
SA (signalizacija)	[A] translacija/transkripcija	NPR-nuc [activates] WRKY70	S
SA (signalizacija)	[V] proteinska vezava	NIMIN + NPR = NIMIN-NPR	N
SA (signalizacija)	[V] proteinska vezava	NIMIN-NPR + TGA = NIMIN-NPR-TGA	S
SA (signalizacija)	[A] translacija/transkripcija	NPR-NIMIN-TGA [activates] PR1/2/5	S
SA (signalizacija)	[virus]	PR1/2/5 [inhibits] virus	N
SA (sinteza)	[A] kataliza	Chor > IsoChor; ICS	S
SA (sinteza)	[A] kataliza	IsoChor > SA(c); IPL	S
SA (sinteza)	[A] translokacija	SA(c) > SA; X3	S
SA (sinteza)	[A] kataliza	Chor > Prep; CM	S
SA (sinteza)	[A] kataliza	Prep > PhePyr; PA	S
SA (sinteza)	[A] kataliza	PhePyr > PheAla; ADT	S
SA (sinteza)	[A] kataliza	PheAla > TransCA; PAL	S
SA (sinteza)	[A] kataliza	TransCA > BA; AAO	S
SA (sinteza)	[A] kataliza	BA > SA; BA2H(a)	S

se nadaljuje

nadaljevanje Priloge A: Seznam vseh reakcij nadgrajenega modela rastlinske imunske signalizacije (PIS).

Proces	Tip reakcije	Reakcija	Vir
SA (sinteza)	[A] kataliza	TransCA > OrthoCA; AAO	S
SA (sinteza)	[A] kataliza	OrthoCA > SA; X1	S
SA (sinteza)	[A] kataliza	SA > MeSA; SAMC	N
SA (sinteza)	[A] kataliza	MeSA > SA; MES	N
SA (hramba)	[A] kataliza	SA > SGE; SAGT	S
SA (hramba)	[A] kataliza	SA > SAG; SAGT	S
medsebojni pogovor	[I] translacija/transkripcija	JAZ [inhibits] EIN3(like)	N
medsebojni pogovor	[A] translacija/transkripcija	MYC [activates] PR3/4	S
medsebojni pogovor	[virus]	PR3/4 [inhibits] virus	N
medsebojni pogovor	[A] translacija/transkripcija	MPK4(a) [activates] LOX	S
medsebojni pogovor	[A] translacija/transkripcija	MYC [activates] LOX	N
medsebojni pogovor	[I] translacija/transkripcija	SA [inhibits] OPDA	S
medsebojni pogovor	[I] proteinske aktivacije	MPK4 [inhibits] EDS1-PAD	S
medsebojni pogovor	[A] proteinska aktivacija	MPK3 [activates] EDS1-PAD	S
medsebojni pogovor	[A] proteinska aktivacija	MPK6 [activates] EDS1-PAD	S
medsebojni pogovor	[A] translacija/transkripcija	SA [activates] PAD4	N
medsebojni pogovor	[A] kataliza	PrEROS > ROS; ROS-production	S
medsebojni pogovor	[V] proteinska vezava	RBX + CUL = RBX-CUL	N
medsebojni pogovor	[V] proteinska vezava	RBX-CUL + ASK = RBX-CUL-ASK	N
medsebojni pogovor	[A] translacija/transkripcija	TAS3 > TAS3; MIR390	N
medsebojni pogovor	[I] razpad	TAS3 [degrades] ARF3/4	N
medsebojni pogovor	[A] kataliza	JA-Ile > 12-OH-JA-Ile; CYP94	N
medsebojni pogovor	[A] proteinska aktivacija	BA2H(i) > BA2H(a); ROS	S
medsebojni pogovor	[A] (de)fosforilacija	MPK3(i) > MPK3(a); HRT + ROS	S
medsebojni pogovor	[A] (de)fosforilacija	MPK6(i) > MPK6(a); HRT + ROS	S
medsebojni pogovor	[virus]	virus [activates] HRT	S

Priloga B: Seznam novih komponent v nadgrajenem modelu rastlinske imunske signalizacije (PIS).
Opisi novih komponent sledijo zastavljeni taksonomiji komponent (Slika 15, str. 54). Preglednica vsebuje seznam 113 novo vključenih genov ali RNA, katere smo združili s seznamom 151 genov ali RNA v objavi Miljkovic in sod. (2012). Vse smo opisali z družino (tretji stolpec) in kratkim imenom komponente ter genskim identifikatorjem v navadnem repnjakovcu (peti stolpec). Za vse (tudi že obstoječe komponente iz objave) smo definirali proces (drugi stolpec) ter pod-družino (četrти stolpec). V drugem stolpcu rastlinski hormoni označeni s kraticami: etilen ET, jasmonska kislina JA, salicilna kislina SA, avksini AUX.

Appendix B: New components the upgraded plant immune signalisation model (PIS).

New component descriptions follow the formulated komponent taxonomy (Figure 15, pp.54). The list contains 113 newly added genes or RNA, which were merged with the list of 151 existing genes or RNA in the publication of Miljkovic in sod. (2012). All new components were described with their gene family (third column) and the combination of components short name and gene identifier in *A. thaliana* (fifth column). We then defined the process (second column) and subfamily for all components (including the 151 previously existing components). Hormone abbreviations in the second column are the following: ethylene ET, jasmonic acid JA, salicylic acid SA and auxins AUX.

Skupina	Proces	Družina	Pod-družina	Kratko ime komponente
protein	ET (sinteza)	ACS	ACS1,2,6	ACS1 (AT3G61510)
protein	ET (sinteza)	ACS	ACS3	ACS3 (AT5G28360)
protein	ET (sinteza)	ACS	ACS10	ACS10 (AT1G62960)
protein	ET (sinteza)	ACS	ACS12	ACS12 (AT5G51690)
protein	ET (signalizacija)	ETP	ETP	ETP2 (AT3G18910)
protein	ET (signalizacija)	ETP	ETP	ETP1 (AT3G18980)
protein	ET (signalizacija)	RTE1	RTE1	RTE1 (AT2G26070)
protein	ET (odgovor)	PDF1.2	PDF1.2	PDF1.2b (AT2G26020)
protein	ET (odgovor)	PDF1.2	PDF1.2	PDF1.2c (AT5G44430)
protein	JA (sinteza)	CTS	CTS	CTS (AT4G39850)
protein	JA (sinteza)	OPR	OPR	OPR1 (AT1G76680)
protein	JA (sinteza)	OPR	OPR	OPR2 (AT1G76690)
protein	JA (sinteza)	OPCL	OPCL	OPCL (AT1G20480)
protein	JA (sinteza)	OPCL	OPCL	OPCL (AT1G20500)
protein	JA (sinteza)	OPCL	OPCL	OPCL (AT5G38120)
protein	JA (sinteza)	MFP	MFP	MFP2 (AT3G06860)
protein	JA (sinteza)	MFP	MFP	KCR1 (AT1G67730)
protein	JA (sinteza)	MFP	MFP	KCR2 (AT1G24470)
protein	JA (sinteza)	MFP	MFP	MFP (AT3G15290)
protein	JA (sinteza)	ACH	ACH	ACH (AT5G48370)
protein	JA (sinteza)	ACH	ACH	ACH (AT2G30720)
protein	JA (sinteza)	JMT	JMT	JMT (AT4G36470)
protein	JA (sinteza)	JAR	JAR	DFL2 (AT4G03400)
protein	JA (sinteza)	CYP94	CYP94	CYP94B1 (AT5G63450)
protein	JA (sinteza)	CYP94	CYP94	CYP94B2 (AT3G01900)
protein	JA (sinteza)	CYP94	CYP94	CYP94B3 (AT3G48520)
protein	JA (sinteza)	CYP94	CYP94	CYP94C1 (AT2G27690)
protein	JA (sinteza)	CYP94	CYP94	CYP94D1 (AT1G34540)
protein	JA (sinteza)	CYP94	CYP94	CYP94D2 (AT3G56630)

se nadaljuje

nadaljevanje Priloge B: Seznam novih komponent v nadgrajenem modelu rastlinske imunske signalizacije (PIS).

Skupina	Proces	Družina	Pod-družina	Kratko ime komponente
protein	JA (signalizacija)	JAM	JAM1	JAM1 (AT2G46510)
protein	JA (signalizacija)	JAM	JAM2	JAM2 (AT1G01260)
protein	JA (signalizacija)	JAM	JAM3	JAM3 (AT4G16430)
protein	JA (signalizacija)	MYC	MYC	MYC3 (AT5G46760)
protein	JA (signalizacija)	MYC	MYC	MYC4 (AT4G17880)
protein	SA (sinteza)	ADT	ADT	ADT3 (AT2G27820)
protein	SA (sinteza)	ADT	ADT	ADT4 (AT3G44720)
protein	SA (sinteza)	ADT	ADT	ADT6 (AT1G08250)
protein	SA (sinteza)	ADT	ADT	ADT1 (AT1G11790)
protein	SA (sinteza)	ADT	ADT	ADT2 (AT3G07630)
protein	SA (sinteza)	AAO	AAO	AAO1 (AT5G20960)
protein	SA (sinteza)	AAO	AAO	AAO3 (AT2G27150)
protein	SA (sinteza)	AAO	AAO	AAO2 (AT3G43600)
protein	SA (hramba)	SAMC	SAMC	SAMC1 (AT4G39460)
protein	SA (hramba)	SAMC	SAMC	SAMC2 (AT1G34065)
protein	SA (hramba)	SAGT	SAGT	SAGT1 (AT2G43820)
protein	SA (hramba)	SAGT	SAGT	UGT74F1 (AT2G43840)
protein	SA (hramba)	MES	MES	MES2/8 (AT2G23600)
protein	SA (hramba)	MES	MES	MES1 (AT2G23620)
protein	SA (hramba)	MES	MES	MES9 (AT4G37150)
protein	SA (hramba)	MES	MES	MES7 (AT2G23560)
protein	SA (hramba)	MES	MES	MES3 (AT2G23610)
protein	SA (hramba)	MES	MES	MES4 (AT2G23580)
protein	SA (signalizacija)	TRX-H	TRX-H	TRX3 (AT5G42980)
protein	SA (signalizacija)	TRX-H	TRX-H	TRX5 (AT1G45145)
protein	SA (signalizacija)	GSNO	GSNO	GSNOR (AT5G43940)
protein	SA (signalizacija)	MOS	MOS	MOS3 (AT1G80680)
protein	SA (signalizacija)	MOS	MOS	MOS6 (AT4G02150)
protein	SA (signalizacija)	MOS	MOS	MOS7 (AT5G05680)
protein	SA (signalizacija)	MOS	MOS	IMPA1 (AT3G06720)
protein	SA (signalizacija)	MOS	MOS	IMPA2 (AT4G16143)
protein	SA (signalizacija)	MOS	MOS	IMPA4 (AT1G09270)
protein	SA (signalizacija)	MOS	MOS	IMPA5 (AT5G49310)
protein	SA (signalizacija)	MOS	MOS	IMPA6 (AT1G02690)
protein	SA (signalizacija)	MOS	MOS	IMPA7 (AT3G05720)
protein	SA (odgovor)	PR1	PR1	PRB1 (AT2G14580)
protein	AUX (signalizacija)	ARF3-4	ARF3-4	ARF3 (AT2G33860)
protein	AUX (signalizacija)	ARF3-4	ARF3-4	ARF4 (AT5G60450)

se nadaljuje

nadaljevanje Priloge B: Seznam novih komponent v nadgrajenem modelu rastlinske imunske signalizacije (PIS).

Skupina	Proces	Družina	Pod-družina	Kratko ime komponente
protein	JA (signalizacija)	JAM	JAM1	JAM1 (AT2G46510)
protein	medsebojni pogovor	MPK3	MPK3	MPK3 (AT3G45640)
protein	medsebojni pogovor	MPK6	MPK6	MPK6 (AT2G43790)
protein	medsebojni pogovor	proizvodnja ROS	CAT	CAT1 (AT1G20630)
protein	medsebojni pogovor	proizvodnja ROS	CAT	CAT2 (AT4G35090)
protein	medsebojni pogovor	proizvodnja ROS	CAT	CAT3 (AT1G20620)
protein	medsebojni pogovor	proizvodnja ROS	GPX	GPX1 (AT2G25080)
protein	medsebojni pogovor	proizvodnja ROS	GPX	GPX2 (AT2G31570)
protein	medsebojni pogovor	proizvodnja ROS	GPX	GPX3 (AT2G43350)
protein	medsebojni pogovor	proizvodnja ROS	GPX	GPX4 (AT2G48150)
protein	medsebojni pogovor	proizvodnja ROS	GPX	GPX5 (AT3G63080)
protein	medsebojni pogovor	proizvodnja ROS	GPX	GPX6 (AT4G11600)
protein	medsebojni pogovor	proizvodnja ROS	GPX	GPX7 (AT4G31870)
protein	medsebojni pogovor	proizvodnja ROS	GPX	GPX8 (AT1G63460)
protein	medsebojni pogovor	proizvodnja ROS	GPX	GPX9 (AT2G30860)
protein	medsebojni pogovor	proizvodnja ROS	NADPH-oxidase	RBOHA (AT5G07390)
protein	medsebojni pogovor	proizvodnja ROS	NADPH-oxidase	RBOHB (AT1G09090)
protein	medsebojni pogovor	proizvodnja ROS	NADPH-oxidase	RBOHD (AT5G47910)
protein	medsebojni pogovor	proizvodnja ROS	NADPH-oxidase	RBOHC (AT5G51060)
protein	medsebojni pogovor	proizvodnja ROS	NADPH-oxidase	RBOHF (AT1G64060)
protein	medsebojni pogovor	proizvodnja ROS	NADPH-oxidase	RBOH (AT1G19230)
protein	medsebojni pogovor	proizvodnja ROS	NADPH-oxidase	RBOH (AT4G11230)
protein	medsebojni pogovor	proizvodnja ROS	NADPH-oxidase	RBOH (AT4G25090)
protein	medsebojni pogovor	proizvodnja ROS	NADPH-oxidase	RBOH (AT3G45810)
protein	medsebojni pogovor	proizvodnja ROS	NADPH-oxidase	RBOH (AT5G60010)
protein	medsebojni pogovor	proizvodnja ROS	SOD	CSD1 (AT1G08830)
protein	medsebojni pogovor	proizvodnja ROS	SOD	CSD2 (AT2G28190)
protein	medsebojni pogovor	proizvodnja ROS	SOD	CSD3 (AT5G18100)
protein	medsebojni pogovor	proizvodnja ROS	SOD	CCS (AT1G12520)
protein	medsebojni pogovor	proizvodnja ROS	SOD	FSD1 (AT4G25100)
protein	medsebojni pogovor	proizvodnja ROS	SOD	FSD2 (AT5G51100)
protein	medsebojni pogovor	proizvodnja ROS	SOD	FSD3 (AT5G23310)
protein	medsebojni pogovor	proizvodnja ROS	SOD	FSD (AT4G00651)
protein	medsebojni pogovor	proizvodnja ROS	SOD	MSD1 (AT3G10920)
protein	medsebojni pogovor	proizvodnja ROS	SOD	SOD (AT3G56350)
protein	medsebojni pogovor	CUL	CUL	CUL1 (AT4G02570)
protein	medsebojni pogovor	CUL	CUL	CUL2 (AT1G02980)
protein	medsebojni pogovor	CUL	CUL	CUL (AT1G43140)
protein	medsebojni pogovor	RBX	RBX	RBX1 (AT5G20570)
protein	medsebojni pogovor	RBX	RBX	RBX (AT3G42830)
protein	medsebojni pogovor	ASK	ASK	ASK1 (AT1G75950)
protein	medsebojni pogovor	ASK	ASK	ASK2 (AT5G42190)
protein	medsebojni pogovor	WRKY33	WRKY33	WRKY33 (AT2G38470)
miRNA	medsebojni pogovor	MIR390	MIR390a	MIR390a (AT2G38325)
miRNA	medsebojni pogovor	MIR390	MIR390b	MIR390b (AT5G58465)
ta-siRNA	medsebojni pogovor	TAS3	TAS3	TAS3 (AT3G17185)
ta-siRNA	medsebojni pogovor	TAS3	TAS3b	TAS3b (AT5G49615)