Univerza v Ljubljani
Fakulteta za računalništvo in informatiko

Robert Cvitkovič

Uporaba bioloških omrežij za izračun funkcijske obogatenosti skupine genov

diplomsko delo

visokošolski strokovni študijski program
prve stopnje
računalništvo in informatika

Mentor: doc. dr. Tomaž Curk

Ljubljana, 2017

Besedilo je oblikovano z urejevalnikom besedil LaTeX.

Fakulteta za računalništvo in informatiko izdaja naslednjo nalogo:

Tematika naloge:

Za razumevanje funkcijske vloge genov biologi izpostavljajo celice različnim eksperimentalnim pogojem in opazujejo odziv posameznih genov. Skupino genov, katerih aktivnost se spremeni v eksperimentalnem pogoju, poskušajo povezati z znanimi procesi in funkcijami genov v celici. Uveljavljeni pristopi za izračun funkcijske obogatenosti skupine genov temeljijo na znanih skupinah funkcijsko povezanih genov. Preučite novejše pristope, ki za izračun obogatenosti namesto znanih množic genov uporabljajo omrežja funkcijsko povezanih genov. Primer takšnega postopka je algoritem SANTA, avtorjev Cornish in Markowetz, PLoS Comput Biol, 2014. Primer omrežij sta zbirka BioGRID in STRING. Rešitev implementirajte kot razširitev obstoječega dodatka Orange Bioinformatics.

Zahvaljujem se mentorju doc. dr. Tomažu Curku za strokovno pomoč pri izdelavi diplome. Zahvaljujem se tudi svojim staršem, ki so mi omogočili študij.

Kazalo

Seznam uporabljenih kratic

kratica	angleško	slovensko
AUK	area under the K-curve	površina pod krivuljo funkcije K
BioGRID	The Biological General Repository for Interaction Datasets	splošni biološki repozitorij za zbirke interakcij
DNA	Deoxyribonucleic acid	deoksiribonukleinska kislina
RNA	Ribonucleic acid	ribonukleinska kislina
SANTA	Spatial Analysis of Network Associations	prostorska analiza omrežnih povezav
siRNA	small interfering RNA	mala interferenčna RNA

Povzetek

Naslov: Uporaba bioloških omrežij za izračun funkcijske obogatenosti skupine genov

Avtor: Robert Cvitkovič

Povzetek: Molekularni biologi in genetiki se dandanes ukvarjajo s proučevanjem zapisa DNA. Pri eksperimentih tipično dobijo podatke, ki se zaradi svoje narave težko interpretirajo. Klasične metode primerjajo eksperimentalno dobljeno skupino genov z znanimi, funkcijsko povezanimi skupinami genov, a pogosto niso uspešne. Izboljšava pristopa je metoda SANTA, ki vrednoti eksperimentalne rezultate na podlagi omrežja genov. Metodo SANTA smo implementirali v sklopu diplome, v dodatku Orange Bioinformatics. Orange je že dobro uveljavljen program za obdelavo podatkov. Dodatek Bioinformatics zavzema specializirane funkcionalnosti za obdelavo genskih podatkov. Dodatek smo obogatili tudi z izboljšanim dostopom do zbirk BioGRID in STRING, ki hranita podatke o eksperimentalno določenih povezavah med geni.

Ključne besede: Python, Orange, bioinformatika, omrežje, SANTA, STRING, BioGRID.

Abstract

Title: Gene set function enrichment analysis using biological networks

Author: Robert Cvitkovič

Abstract: Molecular biologists and geneticists are, nowadays, mostly focused on studying and understanding the DNA transcription. The gathered experimental data is difficult to interpret. Classical methods compare the discovered groups of genes with predefined gene sets. Unfortunately, these methods do not perform well. A solution is proposed by the SANTA method, which evaluates the experimental results on discovered gene sets based on known and extensive gene networks. In this thesis, we have implemented the method in Orange Bioinformatics. Orange is a well-known data-analysis programme. The Bioinformatics add-on includes specialised functionality for processing genomic data. Furthermore, we have also enriched its widgets for access to BioGRID and STRING, which store information on functional sets of genes and interactions.

Keywords: Python, Orange, bioinformatics, network, SANTA, STRING, BioGRID.

Poglavje 1

Uvod

Sodobno računalništvo znanstvenikom omogoča hitro in enostavno zbiranje in obdelavo eksperimentalnih podatkov. Na trgu lahko najdemo veliko strojne in programske opreme, ki so visoko prilagojene in pomagajo znanstvenikom pri njihovem delu. Primer take programske opreme je *Orange* [4], ki je bil razvit v laboratoriju za bioinformatiko na Fakulteti za računalništvo in informatiko, Univerze v Ljubljani. To je splošno namenski program za obdelavo in vizualizacijo podatkov. Od konkurence se loči po tem, da implementira uporabniku enostaven grafični vmesnik. Ta je brezplačen in odprtokoden. Napisan je v programskem jeziku Python, ki mu omogoča visoko modularnost, je visokonivojski in omogoča preprosto spreminjanje ali dodajanje nove kode in funkcionalnosti. Primer dodane funkcionalnosti je dodatek (ang. add-on) za *Orange*, poimenovan *Orange Bioinformatics*, ki je specializiran za delo z biološkimi podatki. Dodatek se tipično uporablja za obdelavo genomskih podatkov.

1.1 Motivacija

Biologi pri eksperimentih pogosto merijo izražanje genov in proteinov, pri različnih organizmih pod različnimi pogoji. Cilj takšnih eksperimentov je preučiti DNA posameznih organizmov in določiti funkcionalnosti genov. Geni

oziroma proteini, ki so izmerjeni v eksperimentih se večinoma primerjajo z definiranimi skupinami genov, kar ne prinaša zanesljivih rezultatov. Takšni in njim podobni algoritmi so že implementirani v okviru *Orange Bioinformatics*.

Vse izvedene meritve se hkrati beležijo v velike zbirke, kjer so genske interakcije predstavljene kot veliko omrežje. Primer takšnih zbirk sta *BioGRID* in *STRING*. Omrežje *BioGRID* je sestavljeno na podlagi eksperimentalnih meritev fizičnih in genskih interakcij med proteini (produkti genov), njihovih interakcij s kemikalijami in na podlagi podobnosti post-transkripcijskih modifikacij proteinov. Omrežje *STRING* je sestavljeno na podlagi fizičnih interakcij, sopojavitve kompleksov, podobnosti v genskih ekspresiji in na podlagi drugih analiz, kot sta, na primer, analiza znanstvenih člankov in analiza ohranjenosti genomskih zaporedij. Zaradi velikost in strukture takšnih zbirk ne obstaja veliko metod, s katerimi bi jih lahko proučevali. Ena izmed redkih je metoda *SANTA* (ang. *Spatial Analysis of Network Associations*), ki nam to omogoča s pomočjo prostorske analize omrežja. Gene oziroma skupine genov lahko povežemo s celičnimi funkcijami in fenotipi. *Orange Bioinformatics* dodatek v celoti ne podpira takšne vrste zbirk in nima implementirane metode *SANTA*, zato smo v sklopu diplome to dodali in testirali delovanje novih funkcionalnosti.

1.2 Cilji in struktura diplomskega dela

Cilj diplomskega dela je razviti gradnike za dodatek *Orange Bioinformatics*, ki implementirajo sledeče funkcionalnosti:

- pridobivanje podatkov iz podatkovne zbirke *BioGRID*,
- pridobivanje podatkov iz podatkovne zbirke *STRING*,
- izračun funkcijske obogatenosti skupine genov, kot je predlagano v metodi *SANTA*.

V diplomi najprej predstavimo zbirki, ki smo ju uporabljali za testiranje. Na kratko predstavimo podatke, ki jih zbirki vsebujeta in opišemo podrob-

nosti implementacije. Podrobneje predstavimo metodo *SANTA*. Podrobneje opišemo delovanje funkcij *Knet* in *Knode*, ki sta definirani znotraj metode *SANTA*. Sledi predstavitev grafičnega vmesnika in delovanje razvitih gradnikov. Delovanje gradnikov prikažemo na različnih umetnih in realnih primerih. Zaključimo s sklepnimi ugotovitvami in predlogi za izboljšavo.

Poglavje 2

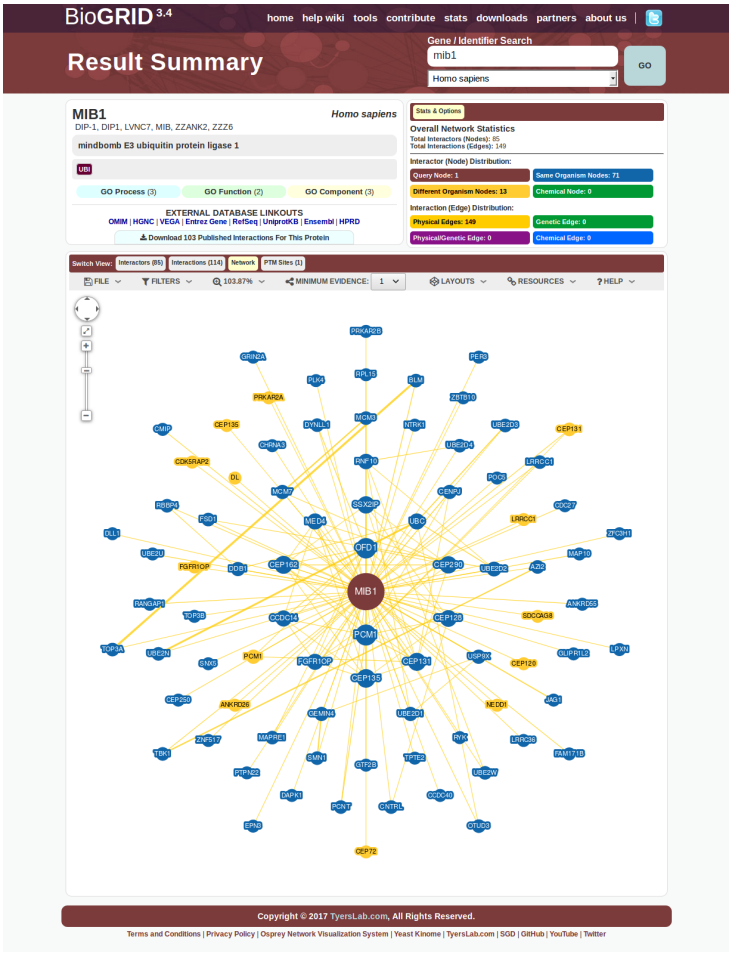
Podatki

Molekularni biologi in genetiki se dandanes ukvarjajo z razumevanjem zapisa DNA, ki določa lastnosti in raznolikost živih bitij. Zapis DNA vsebuje zapise genov, ki preko RNA definirajo proteine, ki vodijo različne procese v celici. Točen namen posameznega gena in proteina biologi ugotovljajo z eksperimentiranjem in primerjavo dobljenih rezultatov. Eksperimentalni rezultati se shrajujo v obsežnih, prosto dostopnih zbirkah, ki so praviloma dostopne preko spleta. V tem poglavju si bomo podrobneje ogledali zbirki *BioGRID* in *STRING*, ki sta pomembni za naše delo, saj hranita omrežje interakcij (ang. interaction) med geni.

2.1 Zbirka *BioGRID*

BioGRID ("The Biological General Repository for Interaction Datasets") je prosto dostopna podatkovna baza, katere glavni namen je beleženje in arhiviranje proteinskih, genetskih in kemičnih interakcij v organizmih. Septembra 2016 je baza vsebovala 1.072.173 genetskih in proteinskih interakcij, izmerjenih v 66 različnih organizmih [2]. Poleg podatkov, ki so dostopni na `https://thebiogrid.org`, lahko na tej spletni strani najdemo tudi aplikacijo (slika 2.1), ki implementira del funkcionalnosti, ki jo bomo pokrili v diplomskem delu. Omogoča namreč iskanje določenega gena in interaktivni

prikaz omrežja sosednih genov.



Slika 2.1: Spletna aplikacije *BioGRID* za vizualizacijo omrežja izbranega gena.

2.1.1 Priprava podatkov

Celotna zbirka, ki jo lahko prenesemo iz spleta, je shranjena v eni tabeli. Tabela vsebuje 24 stolpcev oziroma atributov. Vsaka vrstica predstavlja eno interakcijo med dvema genoma. Za vsako interakcijo (vrstico v tabeli) so hkrati definirani tudi vsi atributi posameznega gena v paru. Zato smo se odločili, da iz teh podatkov ustvarimo dve novi tabeli: *proteins* (tabela 2.1)

in *links* (tabela 2.2)

stolpec	vrsta	opis
biogrid_id_interactor	text	id proteina/gena v *BioGRID* zbirki
entrez_gene_interactor	text	id proteina/gena v Entrez-Gene zbirki
systematic_name_interactor	text	sistematično ime proteina/gena
official_symbol_interactor	text	splošno ime proteina/gena
synonyms_interactor	text	druga imena za protein/gen
organism_interactor	text	id organizma v NCBI Taxonomy zbirki

Tabela 2.1: Opis tabele *proteins* v zbirki *BioGRID*.

stolpec	vrsta	opis
biogrid_interaction_id	niz	id interakcije
biogrid_id_interactor_a	niz	id prvega v interakciji
biogrid_id_interactor_b	niz	id drugega v interakciji
experimental_system	niz	koda dokaza za interakcijo
experimental_system_type	niz	koda vrste dokaza za interakcijo
author	niz	priimek prvega avtorja objave
pubmed_id	niz	id objave, v kateri je bila pokazana interakcija
throughput	niz	prepustnost
score	realno število	ocena zaupanja
modification	niz	translacijske spremembe
phenotypes	niz	fenotip
qualifications	niz	dodatne informacije o interakciji
tags	niz	dodatne oznake interakcije
source_database	niz	vir baze, v kateri je zabeležena interakcija

Tabela 2.2: Opis tabele *links* v zbirki *BioGRID*.

2.1.2 Podrobnosti implementacije

V okviru dodatka *Orange Bioinformatics* je bil za obravnavanje spletnih zbirk razvit sistem *serverfiles*. Ta prenese predhodno obdelavo podatkov na strežnik, končni uporabnik si lokalno prenese manjše, že obdelane podatke. Koda, ki se je poganjala na strežniku, je bila za *BioGRID* že napisana. Prav tako je bila napisana koda, ki je iz lokalne zbirke vračala podatke, vendar ni omogočala filtriranja. Funkcionalnost smo v sklopu novonastalih metod dodali. Novo nastale metode, so:

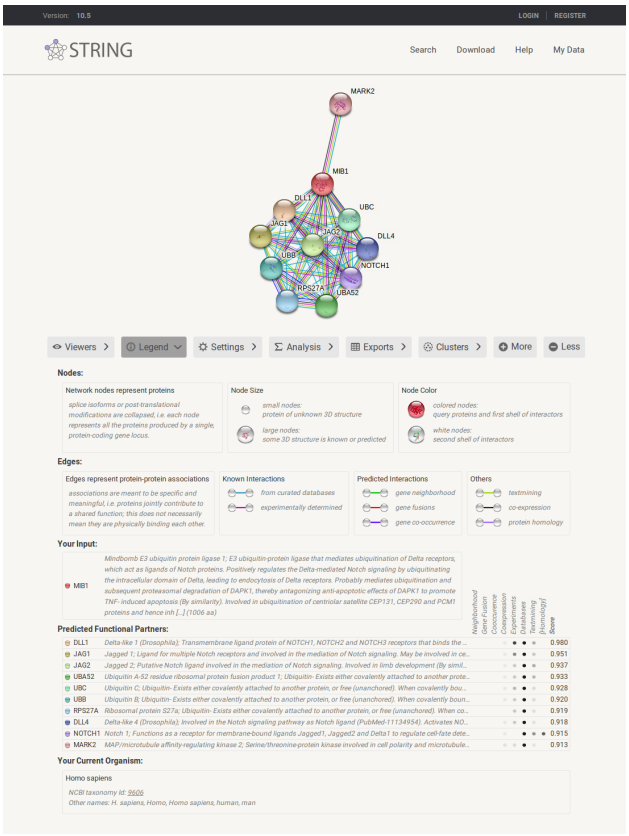
- *number_of_nodes* - za trenutni filter vrne število proteinov oziroma genov,
- *number_of_edges* - za trenutni filter vrne število interakcij,
- *attribute_unique_value* - za podani atribut vrne njegove vrednosti,
- *proteins_table* - vrne *Orange Table* objekt s filtriranimi proteini oziroma geni,
- *links_table* - vrne *Orange Table* objekt s filtriranimi interakcijami,
- *extract_network* - vrne *Orange Network Graph* objekt, zgrajen iz filtriranih podatkov.

Filtriranje omogoča izbiro atributa, po katerem filtriramo in izberemo vrednosti, ki so za izbran atribut sprejemljive. Za predstavitev omrežja smo po standardih *Orange* uporabili že obstoječo rešitev, in sicer dodatek *Orange Network*, natančneje razred *Graph*. Tako smo zadostili konsistentnosti znotraj *Orange* in lahko na dobljenem omrežju uporabljamo tudi ostale funkcionalnosti dodatka *Orange Network*.

2.2 Zbirka *STRING*

Cilj zbirke *STRING* je zbiranje podatkov o interakciji med izraženimi proteini za lažje razumevanje delovanja celičnih funkcij. Zbirka vsebuje direktne

povezave, kot tudi indirektne povezave, če so biološko pomembne. Avgusta 2017 zbirka vsebuje 1.380.838.440 interakcij, 9.643.763 različnih proteinov, 2.031 različnih organizmov [1]. Posebnost zbirke so interakcije med proteini, ki so pridobljene in vrednotene z različnimi načini in ne samo preko eksperimentov. Podobno, kot *BioGRID* spletna zbirka `https://string-db.org` omogoča iskanje posameznih genov in interaktiven prikaz (slika 2.2) omrežja sosedov.



Slika 2.2: Spletna aplikacija *STRING* za vizualizacijo omrežja izbranega gena.

2.2.1 Priprava podatkov

Celotna zbirka vsebuje 39 različnih tabel. Za ustrezno poganjanje metode *SANTA*, ne potrebujemo vseh tabel. Poleg tega smo zaradi velike količine podatkov za vsak organizem ustvarili svojo bazo. V končno bazo smo zajeli tabele *links* (tabela 2.3), *evidence* (tabela 2.4) in *aliases* (tabela 2.5).

stolpec	vrsta	opis
protein_id1	niz	id proteina
protein_id2	niz	id proteina
score	celo število	združena ocena

Tabela 2.3: Opis tabele *links* v zbirki *STRING*.

stolpec	vrsta	opis
protein_id1	niz	id proteina
protein_id2	niz	id proteina
neighborhood	celo število	ocena nastopanja v podobnem genomskem kontekstu
fusion	celo število	ocena združitve v istem genomu
cooccurence	celo število	ocena pojavitve v podobni metabolični poti
coexpression	celo število	ocena skupnega izražanje
experimental	celo število	ocena skupnega nastopanje pri eksperimentih
database	celo število	ocena skupnega nastopanje v bazi
textmining	celo število	ocena skupnega poimenovanje v besedilih

Tabela 2.4: Opis tabele *evidence* v zbirki *STRING*.

stolpec	vrsta	opis
protein_id1	niz	id proteina
alias	niz	oznaka za protein
source	niz	vir oznake

Tabela 2.5: Opis tabele *aliases* v zbirki *STRING*.

2.2.2 Podrobnosti implementacije

Kakor pri zbirki *BioGRID* je bila za zbirko *STRING* že napisana koda, ki se poganja na stežniku in koda za pridobivanje podatkov iz lokalne zbirke. Prav tako ni imela implementirane filtracije. To smo dodali skupaj z ostalimi novimi funkcijami. Novonastale funkcije, so:

- *number_of_nodes* - za trenutni filter vrne število proteinov,
- *number_of_edges* - za trenutni filter vrne število interakcij,
- *proteins_table* - vrne *Orange Table* objekt s filtriranimi proteini,
- *links_table* - vrne *Orange Table* objekt s filtriranimi interakcijami,
- *extract_network* - vrne *Orange Network Graph* objekt, zgrajen iz filtriranih podatkov.

Implementacija filtra se razlikuje z *BioGRID*. *STRING* vsako svojo interakcijo oceni glede na način, kako jo je pridobil (ocene v tabeli 2.4). Zato smo filtriranje implementirali tako, da končni uporabnik izbere način, ki ga zanima in določi spodnjo mejo ocen.

Posebnost nastopi pri implementaciji *proteins_table* saj za vsak protein obstaja veliko alternativnih imen oziroma sinonimov. Tako se pri gradnji tabele najprej iz *aliases* zberejo vsi standardi poimenovanja. Vsak nato postane en stolpec v končni tabeli. Končno se tabela zapolni tako, da za vsak protein podamo njegova imena v različnih standardih poimenovanja v primeren stolpec.

2.3 Periodično posodabljanje podatkov na vmesnem strežniku

Zbirke bioloških podatkov, ki jih uporabljamo pri analizah, so velikokrat obsežne. Končni uporabnik bi tako moral na svoj računalnik prenesti datoteke velikosti nekaj deset GB. Kot rešitev tega problema se je v *Orange* dodal sistem *serverfiles* (shema na sliki 2.3). Njegov glavni namen je predobdelava obsežnih podatkov, pridobljenih iz spletnih portalov. Podatke s pomočjo *serverupdate* skript preoblikujemo v manjše podmnožice in jih nato shranimo v podatkovno zbirko *Sqlite*. Datoteke se nato naloži na lasten strežnik ftp ali http, od koder so dostopne preko spleta. Za dostop do zbirk v programu *Orange*, je v *Orange Bioinformatics* razvit gradnik *Database Update*. Ta omogoča brskanje po *serverfiles* strežniku in prenašanje datotek na lokalni računalnik.

Slika 2.3: Shema *serverfiles*.

Obstoječi kodi *serverupdate* za posodabljanje zbirk *BioGRID* in *STRING* smo dodali ukaze *sql* za združitev tabel in indekse, ki omogočajo hitrejše filtriranje in pridobivanje podatkov.

Poglavje 3

Analiza funkcijske obogatenosti na podlagi omrežja

Metoda *SANTA* (ang. *Spatial Analysis of Network Associations*) je prva, ki za funkcijsko označevanje genskih omrežij uporablja prostorsko analizo omrežja [3]. Metoda dobi na vhod omrežje in podmnožico označenih genov, za katero uporabnik želi izračunati funkcijsko obogatenost. Struktura omrežja odraža trenutno znanje o funkcijski povezanosti genov. Metoda *SANTA* vključuje dve funkciji. Funkcija *Knet* poda indeks razpršenosti označenih genov znotraj omrežja. Funkcija *Knode* poda indeks za označevanje ostalih genov, ki so morebiti soudeleženi v procesih, ki jih opravljajo od uporabnika podani, označeni geni.

3.1 Funkcija *Knet*

Funkcija *Knet* temelji na metodah iz področja prostorske statistike. Izpeljana je iz funkcije K Ripley (ang. Ripley's K-function) [5], ki analizira razpršenost točk. S funkcijo lahko preverimo, ali so točke v prostoru razpršene naključno ali tvorijo vzorec. Uporabnost funkcije K v bioinformatiki je pokazala, n. pr., pri prepoznavanju gostiteljskih faktorjev pri okužbi z virusi na podlagi opazovanja razpršenosti okuženih celic na slikah poskusov siRNA [6].

K^{net} se od funkcije K razlikuje v tem, da namesto naključnih točk v prostoru uporablja vse točke v omrežju. Razdalja med točkama v omrežju je definirana kot najkrajša pot med točkama. Funkcija K^{net} je definirana z enačbo [3]:

$$K^{net}(s) = \frac{2}{(\bar{p}n)^2} \sum_i p_i \sum_j (p_j - \bar{p}) \mathbf{I}(d^g(i,j) \leq s))$$

pri čemer je n število točk, $\mathbf{I}(d^g(i,j) \leq s))$ je indikatorska funkcija, ki vrne 1 kadar je razdalja $d^g(i,j)$ manjša ali enaka s ter 0 kadar je večja. Poleg tega definiramo še povprečno utež genov kot $\bar{p} = \frac{1}{n}\sum_{i=1}^{n} p_i$.

V takšni obliki je dobljen indeks odvisen od razdalje s. Kadar so označene točke razporejene bližje druga drugi, nam K^{net} poda visoke vrednosti že pri nizkem s. Končni rezultat je graf, ki je definiran z vrednostjo $s = 0$ do maksimalne razdalje v grafu (premer grafa). Za lažje vrednotenje dobljenih rezultatov, metoda predvidi še dva koraka. Najprej moramo dodati lažjo predstavitev grafa v numerični obliki. Za to uporabimo vrednost AUK (ang. *area under the funkction K curve*), ki predstavlja površino pod grafom in jo izračunamo s trapezno metodo. Drugi korak je izračun statistične značilnosti (vrednost p) dobljenega rezultata. Uteži obravnavanih genov naključno premešamo med gene v omrežju in za novo nastale podatke izračunamo K^{net} ter njegov AUK. Postopek ponovimo N_{perm}-krat in z uporabo testa Z ugotovimo statistično značilnost dobljenega rezultata.

3.2 Funkcija *Knode*

K^{node} nam omogoča funkcijsko označevanje genov, ki niso direktno nastopili v podani skupini genov, a so vseeno z njimi tesno povezani. Definiran je tako, da za vsak neoznačen gen izračunamo indeks pomembnosti z uporabo enačbe [3]:

$$K_i^{node}(s) = \frac{2}{(\bar{p}n)^2} \sum_j (p_j - \bar{p}) \mathbf{I}(d^g(i,j) \leq s))$$

Podobno kot K^{net}, tudi funkcija K^{node} vrača rezultate za vrednosti $s = 0$ do maksimalne razdalje v grafu. Vrednost AUK izračunamo z uporabo trapezne metode na dobljenem grafu vrednosti K^{node} za različne vredosti s. Vrednost AUK končno služi kot indeks pomembnosti. Večja vrednost nakazuje večjo povezanost posameznega gena s podano skupino genov.

3.3 Podrobnosti implementacije metode *SANTA*

Za delovanje metode *SANTA* smo definirali istoimenski razred. Ob inicializaciji razreda, le-ta zahteva objekt tipa *Orange Network Graph*, s katerim je predstavljeno omrežje, in slovar, v katerem so podane uteži vozlišč v omrežju. Napisane metode zahtevajo le omenjene osnovne podatke o omrežju in vozliščih. Računanje *Knet* in *Knode* tako deluje na vseh vrstah omrežij in ne samo na genskih oziroma proteinskih podatkih. Javne metode razreda, so:

- *k_net* - vrne seznam izračunanih *Knet* vrednosti ter vrednost AUK,
- *auk_p_value* - za podano število permutacij izračuna vrednost p,
- *k_node* - vrne seznam neoznačenih vozlišč z izračunano vrednostjo *Knode*,
- *k_node_table* - iz podanega seznam iz *k_node* metode ustvari in vrne objekt tipa *Orange Table*,
- *k_net_table* - iz podanega seznam iz *k_net* metode ustvari in vrne objekt tipa *Orange Table*.

Poglavje 4

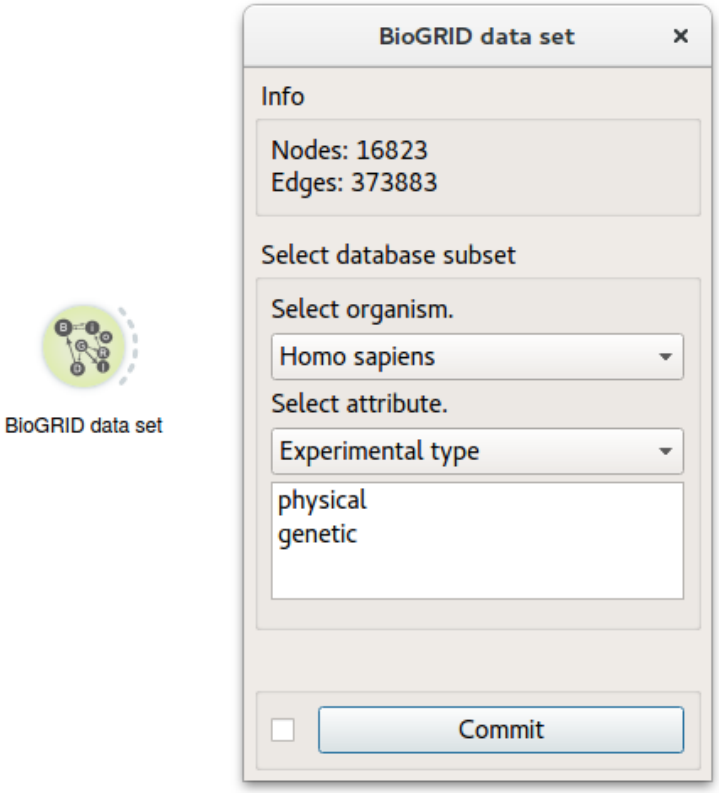
Predstavitev razvitih gradnikov orodja *Orange*

Dodatek *Orange Bioinformatics* smo obogatili s tremi novimi gradniki, ki smo jih razvili v okviru diplomskega dela:

- *BioGRID database*,
- *STRING database*,
- *SANTA*.

4.1 Gradnik *BioGRID database*

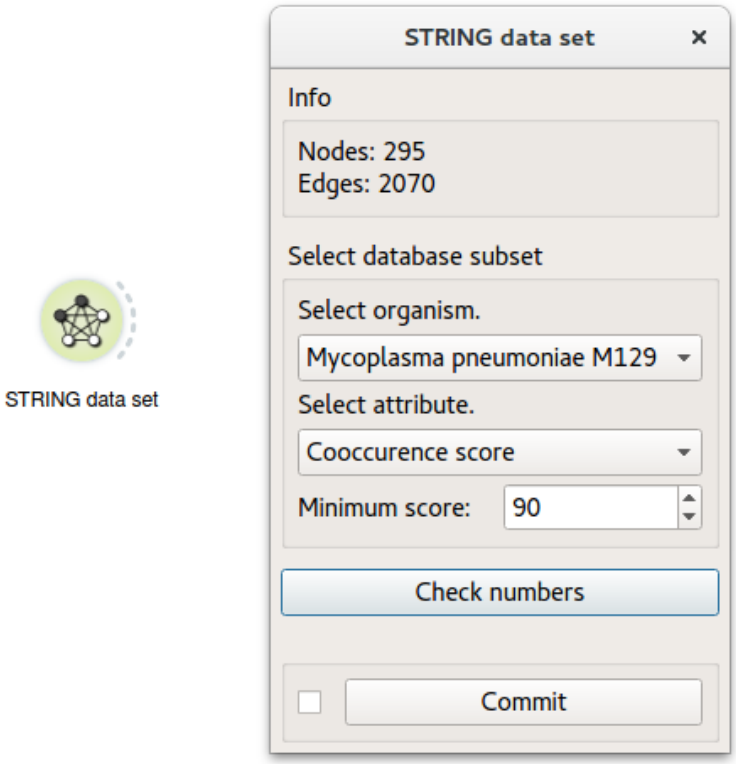
Gradnik *BioGRID database* (slika 4.1) od uporabnika zahteva, da izbere željen organizem. Nato mu ponudi dodatno filtriranje po atributih in njihovih vrednostih. Za izhod poda dva *Orange Table* in *Orange Network Graph*. V prvi tabeli se nahajajo podatki o filtriranih proteinih oziroma genih. V drugi tabeli so podatki o interakcijah. Dobljeno omrežje sestavljajo podatki iz obeh tabel.



Slika 4.1: Izgled ikone in nastavitev gradnika *BioGRID database*.

4.2 Gradnik *STRING database*

Gradnik *STRING database* (slika 4.2) od uporabnika zahteva, da izbere željen organizem. Nato mu ponudi dodatno filtriranje po *STRING* metodah dokaza za interakcijo in njihovih vrednosti. Zaradi velike količine podatkov, gradnik omogoča poizvedovanje o številu vozlišč in povezav v končnem rezultatu. Izhod je enak kot pri *BioGRID database* gradniku. V prvi tabeli se nahajajo podatki o filtriranih proteinih. V drugi tabeli so podatki o interakcijah. Dobljeno omrežje sestavljajo podatki iz obeh tabel.

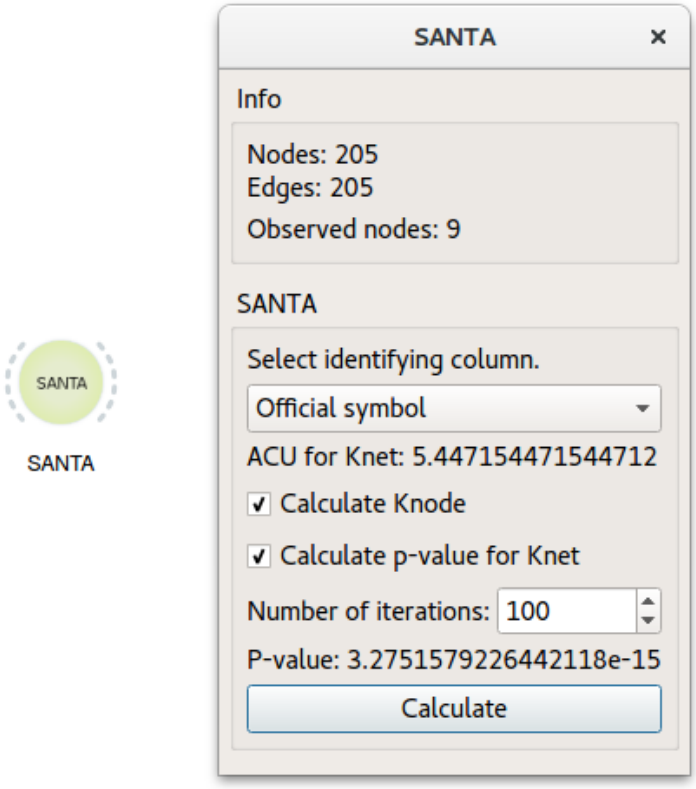


Slika 4.2: Izgled ikone in nastavitev gradnika *STRING database*.

4.3 Gradnik *SANTA*

Gradnik *SANTA* (slika 4.3) kot vhod prejme *Orange Network Graph* in *Orange Table*. V prvem se nahajajo podatki o omrežju, na katerem se bodo izvajale funkcije metode *SANTA*, v drugem podatki o zbranih proteinih oziroma genih. Obenem zahteva, da uporabnik izbere polje v tabeli, kjer se nahajajo oznake za proteine oziroma gene. Uporabnik lahko izbere ali mu gradnik izračuna tudi vrednost p za *Knet* in tabelo rezultatov iz metode *Knode*. Kot izhod, gradnik vrne *Orange Table* z vrednostim za *Knet* za vrednosti $s = 0$ do največje razdalje v omrežju. Ob izbiri vrne tudi dodaten

objekt tipa *Orange Table* s podatki iz metode *Knode*.

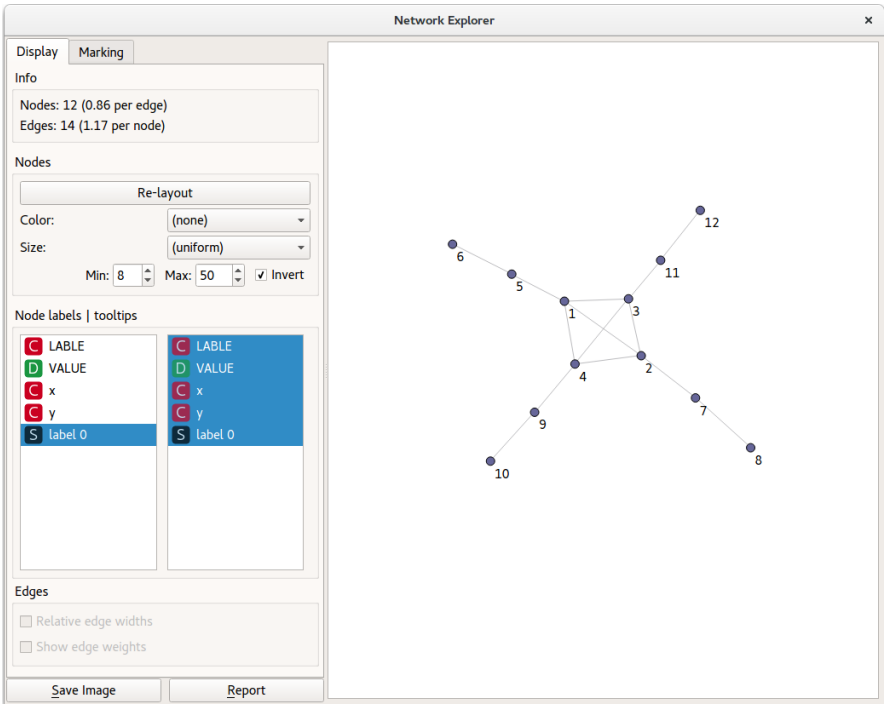


Slika 4.3: Izgled ikone in nastavitev gradnika *SANTA*.

Poglavje 5

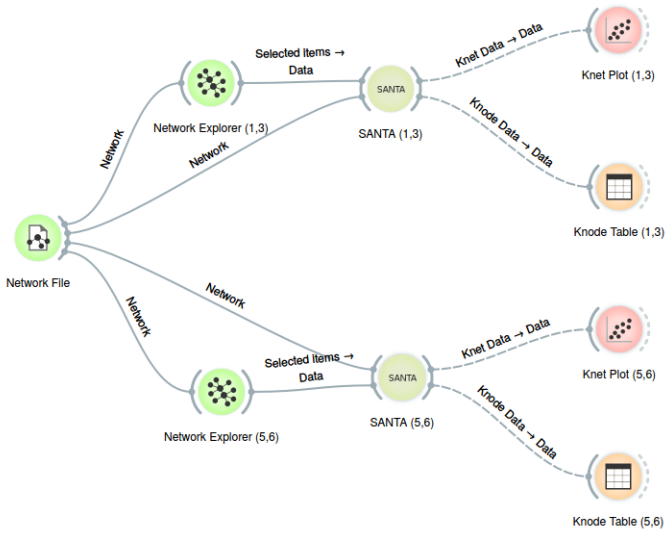
Primeri uporabe

Prvi primer uporabe je enostavno omrežje, ki sta ga uporabila avtorja Cornish in Markowetz v članku, kjer predlagata metodo *SANTA* [3]. Omrežje je prikazano na sliki 5.1.



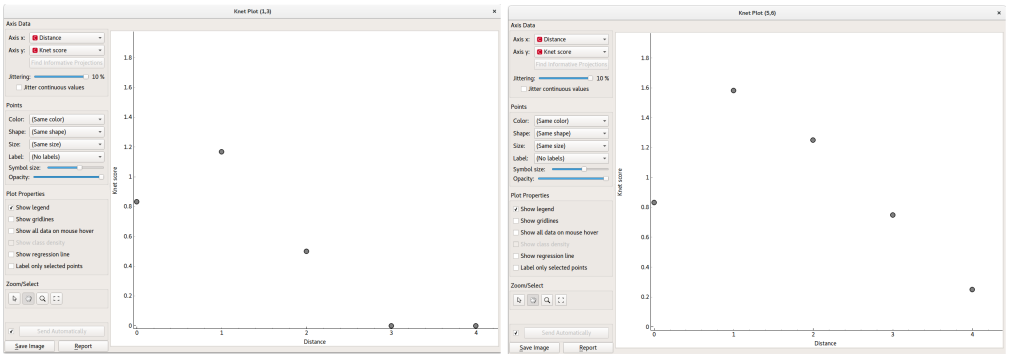
Slika 5.1: Prikaz omrežja v gradniku *Network explorer*.

Za testiranje gradnika *SANTA* smo v vhod podali dva para vozlišč. Najprej smo podali vozlišči 1 in 3, nato še vozlišči 5 in 6. Celoten izgled postavitve in povezave gradnikov lahko vidimo na sliki 5.2.



Slika 5.2: Prikaz sheme gradnikov za testiranje gradnika *SANTA*.

Ko poženemo gradnik *SANTA*, lahko v njemu razberemo rezultat za AUK in vrednost p. Z omrežje, kjer sta izbrani vozlišči 1 in 3 ima AUK vrednost 2,083, vrednost p je 0,347. Pri omrežju, kjer sta izbrani vozlišči 5 in 6 je vrednost AUK enaka 4,125 in vrednost p znaša 0,001. Rezultati se zdijo smiselni, saj se vozlišči 1 in 3 nahajata v centru omrežja in imata več povezav z ostalimi vozlišči. O enakih rezultatih poroča tudi članek [3]. Dobljene grafe *Knet* si lahko podrobneje ogledamo na sliki 5.3.

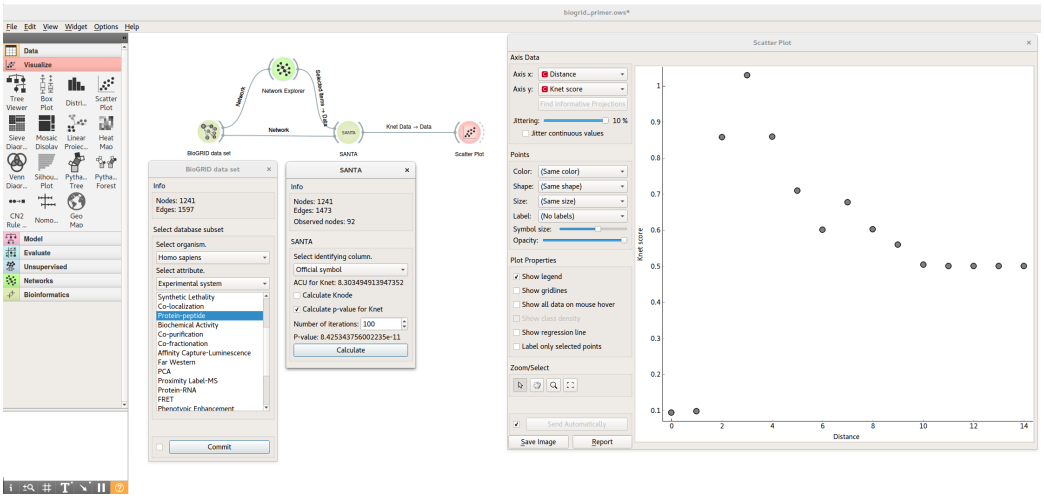


Slika 5.3: Prikaz rezultatov *Knet* v razsevnem diagramu, gradniku *Scatter Plot*.

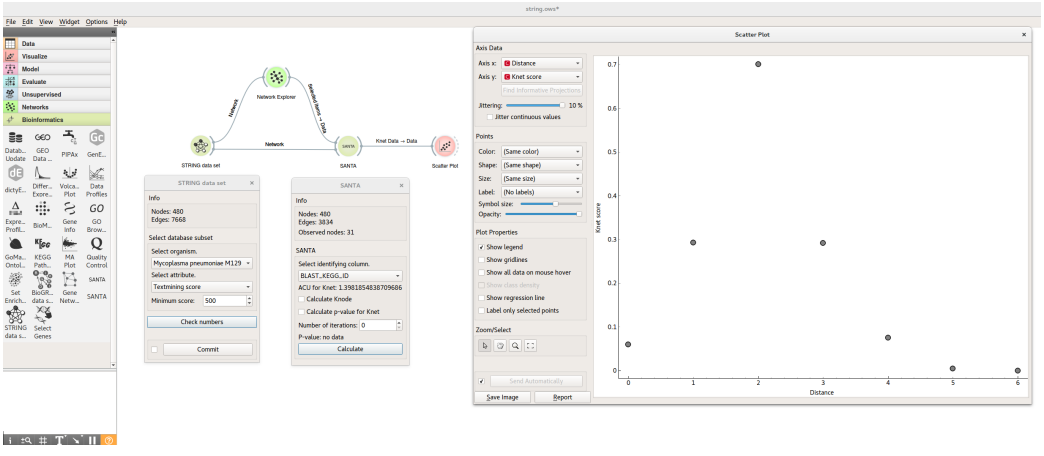
Poleg tabele *Knet*, gradnik *SANTA* vrne tudi rezultate *Knode* v obliki tabele vozlišč z rezultati funkcije. Iz tabele je možno enostavno razbrati vpliv izbranih vozlišč na rezultate. Za izbrani vozlišči 5 in 6 metoda *Knode* vrne pozitivno vrednost le za sosednje vozlišče 1, za vse ostale pa negativno. Za izbrani vozlišči 1 in 3 metoda najbolje oceni vozlišči 5 in 11.

Drugi primer prikazuje delovanje *SANTA* na podatkih iz zbirke *BioGRID*. Slika 5.4 prikazuje delovno površino. V gradniku *BioGRID* najprej izberemo iskan organizem, v našem primeru je to *Homo sapiens*. Dobljeno omrežje zmanjšamo z izbiro vrednosti "*Protein-peptide*" pri atributu "*Experimental system*." Rezultat gradnika *BioGRID* se nato pošlje v gradnik *Network Explorer*, kjer smo izbrali podmnožico vozlišč v središču omrežja. Izbrana podmnožica vozlišč se pošlje v gradnik *SANTA* ter prikaže graf *Knet*.

Tretji primer uporabe je podoben drugemu z izjemo, da je tokrat uporabljena zbirka *STRING*. Slika 5.5 prikazuje delovno površino. Pri gradniku *STRING* uporabnik najprej izbere željen organizem. Nato lahko izbor rezultatov zmanjša z dodatnim filtriranjem po najmanjši vrednosti, izbrane ocene.



Slika 5.4: Prikaz primera uporabe *SANTA* skupaj z *BioGRID*.



Slika 5.5: Prikaz primera uporabe *SANTA* skupaj s *STRING*.

Poglavje 6

Sklepne ugotovitve

V sklopu diplomskega dela smo dodatek k *Orange*, imenovan *Orange Bioinformatics*, obogatili s tremi novimi gradniki. Dva gradnika omogočata dostop do podatkov iz spletnih zbirk bioloških podatkov in znanja ter njihovo pretvorbo v omrežje interakcij med geni. Tretji gradnik omogoča izvajanje metode *SANTA* za izračun funkcijske obogatenosti skupine genov na osnovi omrežja znanih interakcij med geni. Implementirani gradniki so sestavni del dodatka, ki je dostopen na strani github `https://github.com/biolab/orange-bio`.

V nadaljnjem delu bi bilo smiselno uporabiti drugo knjižnico za predstavitev omrežja. Računsko najzahtevnejši del metode *SANTA* je računanje razdalj med vozlišči. Trenuten algoritem je implementiran v Python. Z uporabo knjižice, ki omenjeni izračun implementira s pomočjo programskega jezika C, bi znatno pohitril izvajanje algoritma *SANTA*. V takem primeru bi morali spremeniti celoten dodatek *Orange Network*, da bi lahko zadostili konsistentni uporabi gradnikov.

Literatura

[1] (2017) String: functional protein association networks. Dostopno na: `https://string-db.org/`

[2] Andrew Chatr-aryamontri, Rose Oughtred, Lorrie Boucher, Jennifer Rust, Christie Chang, Nadine K Kolas, Lara O'Donnell, Sara Oster, Chandra Theesfeld, Adnane Sellam, et al. The biogrid interaction database: 2017 update. *Nucleic acids research*, 45(D1):D369–D379, 2017.

[3] Alex J. Cornish and Florian Markowetz. Santa: Quantifying the functional content of molecular networks. *PLOS Computational Biology*, 10(9):1–11, 09 2014.

[4] Janez Demšar, Tomaz Curk, Aleš Erjavec, Črt Gorup, Tomaž Hočevar, Mitar Milutinovič, Martin Možina, Matija Polajnar, Marko Toplak, Anže Starič, et al. Orange: data mining toolbox in python. *Journal of Machine Learning Research*, 14(1):2349–2353, 2013.

[5] Brian D Ripley. *Statistical inference for spatial processes*. Cambridge university press, 1991.

[6] Apichat Suratanee, Ilka Rebhan, Petr Matula, Anil Kumar, Lars Kaderali, Karl Rohr, Ralf Bartenschlager, Roland Eils, in Rainer König. Detecting host factors involved in virus infection by observing the clustering of infected cells and sirna screening images. *Bioinformatics*, 26(18):i653–i658, 2010.