Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

Analýza génovej expresie kvasinky Candida parapsilosis pri raste na rôznych substrátoch

Bakalárska práca

2019
Kristína Váryová

Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

Analýza génovej expresie kvasinky Candida parapsilosis pri raste na rôznych substrátoch

Bakalárska práca

Študijný program:	Bioinformatika
Študijný odbor:	Informatika a Biológia
Školiace pracovisko:	Katedra informatiky
Školiteľ:	Kristína Váryová

Bratislava, 2019
Kristína Váryová

Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

ZADANIE ZÁVEREČNEJ PRÁCE

Meno a priezvisko študenta:	Kristína Váryová
Študijný program:	bioinformatika (Medziodborové štúdium, bakalársky I. st., denná forma)
Študijné odbory:	informatika biológia
Typ záverečnej práce:	bakalárska
Jazyk záverečnej práce:	anglický
Sekundárny jazyk:	slovenský

Názov: Gene expression analysis in the yeast Candida parapsilosis growing on different substrates
Analýza génovej expresie kvasinky Candida parapsilosis pri raste na rôznych substrátoch

Anotácia: Cieľom práce je analyzovať transkriptomické dáta získané z kvasinky Candida parapsilosis pri raste na rôznych substrátoch (glukóza, 4-hydroxybenzoát a 3-hydroxybenzoát) a gény s výrazne zmenenou expresiou charakterizovať pomocou informácií o ich funkciách z dostupných databáz.

Vedúci:	doc. Mgr. Bronislava Brejová, PhD.
Katedra:	FMFI.KI - Katedra informatiky
Vedúci katedry:	prof. RNDr. Martin Škoviera, PhD.
Dátum zadania:	22.10.2018

Dátum schválenia: 22.10.2018 doc. Mgr. Bronislava Brejová, PhD.
garant študijného programu

.....
študent

.....
vedúci práce

Poďakovanie: V prvom rade by som chcela poďakovať mojej školiteľke doc. Mgr. Bronislave Brejovej, PhD. za pevné nervy, ktoré pri mne musela mať a jej veľkú ochotu mi radiť, opravovať chyby a stále ma tlačiť do písania.

Ďalej moje poďakovanie patrí môjmu priateľovi Karolovi Hrubjákovi, ktorý pri mne stál celý ten čas a veril vo mňa.

Rada by som ešte poďakovala mojej rodine, ktorá mi robila morálnu oporu po celý čas písania.

Veľké ďakujem patrí aj mojim pseudo spolužiakom, ktorý sa trápili rovnako ako ja a tým mi dodávali silu pokračovať, a ktorí boli mojim svedomím v písaní. Bez ich pomoci a priateľstva by som sa na škole veľa natrápila. Ďakujem Marcel, Emo, Juro, Feri a Angelika.

A na záver by som chcela poďakovať mojim Dominikám, jediným spolužiačkám. Bez nich by bolo štúdijum nudné a niektoré prednášky na nevydržanie. Vzájomná podpora počas štúdia mi veľmi pomohla.

Abstrakt

VÁRYOVÁ, Kristína: Analýza génovej expresie kvasinky Candida parapsilosis pri raste na rôznych substrátoch. [Bakalárska práca]. Univerzita Komenského v Bratislave. Fakulta matematiky, fyziky a informatiky; Katedra informatiky. doc. Mgr. Bronislava Brejová, PhD. Stupeň odbornej kvalifikácie: bakalár. Bratislava : FMFI UK, 2019. 36s.

Cieľom práce bolo analyzovať transkriptomické dáta z kvasinky *Candida parapsilosis* pri raste na glukóze, 3–hydroxybenzéne a 4–hydroxybenzéne. Skúmali sme, ktoré GO kategórie boli obohatené na diferencovane exprimované gény. Pomocou Fischerovho presného testu sme zistili, ktoré GO kategórie to sú.

Kľúčové slová: *Candida parapsilosis*, databáza Gene Ontology, databáza Uniprot, RNA-seq, Fischerov presný test, Mannov-Whitneyho U test, génová expresia

Abstract

VÁRYOVÁ, Kristína: Gene expression analysis in the yeast Candida parapsilosis growing on different substrates [Bachelor's thesis]. Comenius University in Bratislava, Faculty of Mathematics, Physics and Informatics; Department of Computer Science. doc. Mgr. Bronislava Brejová, PhD. Professional qualification level: bachelor I. deg. Bratislava : FMFI UK, 2019. 36 pages.

The goal of the thesis is to analyze transcriptomic data obtained from the yeast Candida parapsilosis growing on glucose, 4-hydroxybenzoate and 3-hydroxybenzoate substrates. We examined which GO categories were enriched by differentiated genes. We found out enriched GO categories using Fisher exact test.

Keywords: *Candida parapsilosis*, Gene Ontology database, Uniprot database, RNA-seq, Fischer exact test, Mann-Whitney U test, gene expression

Obsah

Zoznam obrázkov

Zoznam tabuliek

Úvod

Kvasinky sú jednobunkové organizmy a vďaka tomu sa môžu zdať jednoduché. Avšak opak je pravdou. Niektoré sú veľmi užitočné pre ľudí, napríklad používajú sa pri výrobe antibiotík. Iné sú však pre ľudský organizmus škodlivé až patogénne (choroboplodné). Jednou z týchto patogénnych kvasiniek je *Candida parapsilosis*. Patrí medzi jedny z najvýznamnjších patogénov z rodu *Candida*. Tím vedcov [28] zistil, že kvasinka *C. parapsilosis* dokáže prežiť na dvoch neobvyklých médiach, na 3–hydroxybenzoáte (3–OH) a 4–hydrohybenzoáte (4–OH). Je to veľmi zaujímavý objav, pretože objavitelia predpokladajú, že táto vlastnosť jej môže pomáhať pri rezistencii na antimykotiká [15].

Cieľom našej práce je analýza dát o úrovni génovej epresie buniek kvasinky *Candida parapsilosis* z publikácie od Zemana a kol. [28]. Kvasinka rástla na dvoch rôznych médiach - 3–hydroxybenzoáte, 4–hydroxybenzoáte a glukóze. Našim cieľom je charakterizovať gény s výrazne zmenenou expresiou, na čo použijeme rôzne databázy.

V prvej kapitole popíšeme základné pojmy súvisiace s génovou expresiou a jej reguláciou a zároveň opisujeme vlastnosti kvasinky *C. parapsilosis*. Ďalej v práci opisujeme zdroje dát, ktoré budeme používať (kapitola 2). Na analýzu rastu kvasinky potrebujeme pospájať dáta z rôznych databáz. Dôležitá súčasť práce je spracovanie dát a vysvetlenie štatistických testov (tretia kapitola) a výsledky (štvrtá kapitola).

Kapitola 1

Úvod do problematiky

V prvej kapitole zavádzame hlavné pojmy dôležité pre pochopenie práce. Jedným z našich cieľov je skúmať expresiu génov, na ktorú potrebujeme poznať poznatky o centrálnej dogme molekulárnej biológie. Priblížíme si proces transkripcie a translácie a spôsoby ich regulácie.

Ďalej v kapitole spomenieme kvasinku *Candida parapsilosis* a jej základné vlastnosti. Priblížime si nové zistenia o jej raste, ktoré nám posúžili ako základ, na ktorom staviame našu prácu.

1.1 Génová expresia

Génová expresia je proces syntézy proteínov v bunke. Predstavuje vyjadrenie genetickej informácie vo forme finálneho produktu, väčšinou proteínu prípadne funkčnej molekuly RNA.

Gén je postupnosť niekoľkých nukleotidov (stovky až tisícky), ktoré kódujú proteíny.

1.1.1 Centrálna dogma molekulárnej biológie

Centrálna dogma hovorí, že genetická informácia sa kopíruje z DNA do DNA, pri prenose z generácie na generáciu a tiež z DNA na proteíny cez génovú expresiu. Tento preklad na proteíny je realizovaný v dvoch krokoch: transkripciou a transláciou.

Transkiprica je prepis DNA do RNA. Nasleduje translácia, čo je prepis RNA do proteínov. Pred transláciou tiež prebiehajú rôzne modifikácie RNA.

1.1.2 Transkripcia

Transkripcia tvorí prvý krok v génovej expresii, t.j. vytvorenie mRNA. Potrebných je na to niekoľko enzýmov, ako napríklad RNA—polymeráza. DNA vlákno sa transkribuje do mRNA v smere $5'$–$3'$[21].

Transkripcia je rozdelená do troch krokov: iniciácia, elongácia a terminácia. V iniciácii nastáva rozvinutie dvojvláknovej DNA a tým vznikne templátové vlákno. V elongácii sa začne transkribovať. Postupne ako vznikajú pre–mRNA vlákna, pridá sa na ich 5′–konce 7-metylguanozinova čiapočka. Následne v poslednom kroku, terminácii sa na novo vzniknuté vlákno pre-mRNA pridá na 3′–koniec poly(A) chvost. Prekurzorovou pre-mRNA označujeme primárny transkript, ktorý ešte nie je spracovaný [21].

Po tomto kroku, nasleduje ďalšia veľmi dôležitá modifikácia – zostrih (splicing). V tejto časti sa vystrihujú nekódujúce časti pre–mRNA nazývané intróny a kódujúce časti, exóny, sa navzájom spájajú [21].

1.1.3 Translácia

Translácia je druhý veľmi dôležitý krok génovej expresie. Počas translácie, mRNA je prečítaná a preložená do amino kyselín, ktoré sa spájajú do proteínov.

Transkribovaná mRNA je čítaná v tripletoch čo znamená, že jednu aminokyselinu kódujú tri bázy z mRNA. Tomuto tripletu hovoríme kodón.

Translácia prebieha v ribozóme. Začína v mRNA na špeciálnom mieste, nazývanom štart sekvencia alebo štart kodón. Ribozóm ju rozozná a vďaka tomu môže začať translatovať. Proces sa ukončí stop sekvenciou (stop kodónom).

1.1.4 Alternatívny zostrih RNA

Ako sme už spomínali v tejto kapitole, eukaryotické gény obsahujú intróny, ktoré musia byť vystrihnuté. Následne sa musia exóny správne pospájať a vytvoria mRNA.

Intróny môžu byť vyňaté jednotlivo alebo spoločne. Ak sú dva po sebe nasledujúce intróny odstránené spoločne a majú exón medzi sebou, bude vyňatý aj exón. Tým pádom výsledná mRNA bude mať rozdielnu sekvenciu a vystrihnutá časť sa nebude podieľať ani na tvorbe proteínu. Vďaka alternatívnemu zostrihu môže jeden gén kódovať viacero proteínov.

1.2 Regulácia expresie

Nie všetky gény sú transkribované neustále. Namiesto toho, bunka kontroluje expresiu každého génu. Transkribované sú len tie gény, ktorých produkty sú potrebné v daný moment [21].

Eukaryotická bunka sa skladá z viacerých organel. Tieto kompartmenty sú od seba oddelené membránami. Rozdelenie eukaryotickej bunky na organely sťažuje génovú expresiu, pretože jednotlivé kroky sú od seba fyzicky oddelené. Transkripcia sa vykonáva v jadre, zatiaľ čo naviazanie ribozómov na mRNA prebieha v cytoplazme.

Regulácia môže prebiehať ako v jadre, tak aj v cytoplazme a to na rôznych úrovniach – DNA alebo RNA alebo na úrovni polypeptidov. Riadiť transkripciu je pre bunku zložité, lebo tento krok génovej expresie sa vykonáva v jadre a bunka potrebuje dostať signály do jadra z povrchu bunky. Je preto potrebný veľmi komplexný systém odovzdávania signálov. Väčšinou sa z povrchu bunky musí signál dostať cez cytoplazmu a jadrovú membránu až ku chromozómom. Následne môže nastať ovplyvňovanie transkripcie [21].

Regulácia génovej expresie môže byť ovplyvňovaná pozitívnymi aj negatívnymi regulačnými mechanizmami. Pri pozitívnych regulačných mechanizmoch je produkt génu potrebný k zapnutiu expresie iného génu a pri negatívnych regulačných mechanizmoch je produkt génu potrebný k vypnutiu expresie iného génu. Tieto produkty sa nazývajú aktivátory, ak aktivujú expresiu (pomáhajú promotoru sa naviazať) a represory, ak zabraňujú génovej expresie.

Aktivátory a represory regulujú génovú expresiu tak, že sa naviažu na miesta, ktoré susedia s väzobnými miestami pre začiatok transkripcie. Ich naviazanie je regulované malými efektorovými molekulami nazývané aj induktory alebo korepresory. Induktory sa naviažu na represory, a tým im bránia naviazať sa na molekulu DNA. Korepresory sa viažu na represor, a tým mu pomáhajú naviazať sa.

Na molekuly mRNA sa vedia naviazať malé interferujúce RNA (siRNA) alebo mikroRNA (miRNA), ktoré spôsobujú deaktiváciu molekuly RNA. Vznikajú z dvojreťazových RNA.

1.3 Candida parapsilosis

Candida parapsilosis je ľudská patogénna kvasinka. Bola objavená v štáte strednej Ameriky, v Puerto Ricu. Objavil ju pán Ahford v stolici u pacienta s hnačkou v roku 1928. [26]. V roku 1940 sa zistilo, že spôsobuje aj ochorenie endokarditídu čo je srdcové ochorenie a *C. parapsilosis* klasifikovali ako patogénnu. [26].

Vzhľad kvasinky *C. parapsilosis* je podmienený médiom, na ktorom rastie [26]. Napríklad, keď rastie na živnej pôde CHROMagar, je oválneho tvaru alebo guľatého až cylindrického a vytvára ružové kolónie. Ak rastie na živnej pôde *Columbia Horse Blood Agar SAB*, tak má hladkú štruktúru bielej až krémovej farby. Často vytvára biofilmy, ktorú sú veľmi odolné až rezistentné na rôzne druhy antifungálnych liekov. Biofilm je zoskupenie mikroorganizmov, ktoré vylučuje proteíny, sacharidy a podobné látky, čím si vytvára ochrannú štruktúru [22].

Ochorenie spôsobené kvasinkou *C. parapislosis* väčšinou vyžaduje hospitalizáciu. Veľa novorodeneckých ochorení a v niektorých prípadoch až úmrtí je spojených s chorobami spôsobenými *C. parapsilosis*. Ďalšia veľmi vážna choroba, ktorú spôsobuje, je

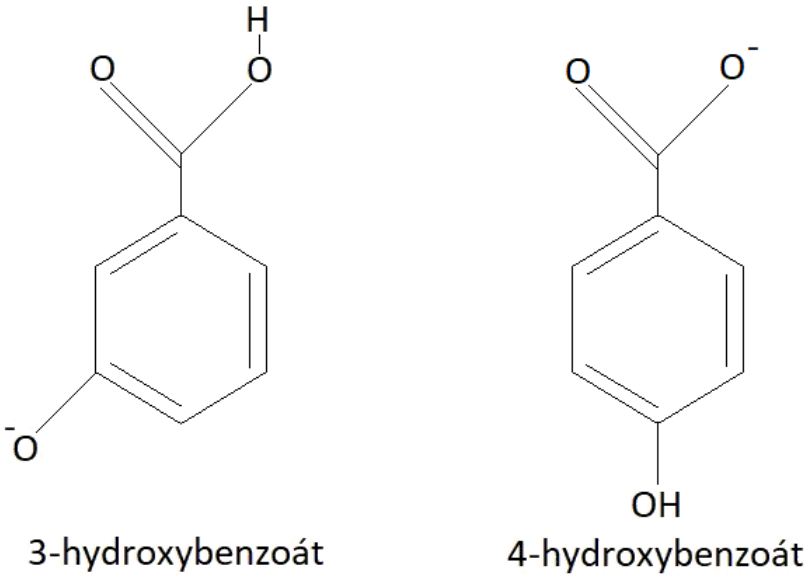
infekcia obehového systému, ktorá bola objavená len nedávno v roku 1980. [26]

1.4 Rast *C. parapislosis* na 3–OH a 4–OH médiach

Kvasinka *Candida parapislosis* vie využiť široké množstvo hydroxyderivátov benzénu a kyseliny benzoovej ako jediný zdroj uhlíka. Tieto zlúčeniny vie metabolizovať cez 3–oxoadipát (3–OAP) alebo cez glutatión (GP) [28], viď. obrázok 1.2.

Zeman a spolupracovníci skúmali rast kvasinky *C. parapsilosis* na médiách obsahujúcich 4–hydroxybenzoát (4–OH) a 3–hydroxybenzoát (3–OH) [28].

Monohydroxybenzoát, 3–hydroxybenzoát, je konjugovanou bázou kyseliny 3–hydroxybenzoovej. Zohráva úlohu v metabolizme u baktérii aj rastlín [1]. Zlúčenina 4–hydroxybenzoát je konjugovanou bázou kyseliny 4–hydroxybenzoovej. Je súčasťou metabolizmu rastlín a *Saccharomyces cerevisiae* [2]. Chemická štruktúra zlúčenín je vidieť na obrázku 1.1.



Obr. 1.1: Chemická štruktúra 3–OH a 4–OH zlúčenín

Zeman a kol. [28] zistili, že gény kódujúce kľúčové enzýmy pre GP a 3–OAP metabolické dráhy sú lokalizované v dvoch génových klastroch. Zhluk génov (klaster) je oblasť genómu obsahujúci niekoľko génov kódujúcich proteíny alebo polypeptidy, ktoré sa nachádzajú blízko seba v genóme a majú podobnú funkciu. Gény v zhlukoch pre GP a 3–OAP metabolické dráhy majú nízku úroveň génovej expresie, keď rastie *C. parapsilosis* na glukóze. Avšak keď sú bunky kultivované na hydroxyaromatických zlúčeninách ako napríklad 4–OH, hydrochinóne, rezorcinole (v 3–OAP), 3–OH alebo na gentisate (v GP), tak expresia týchto génov je vyššia.

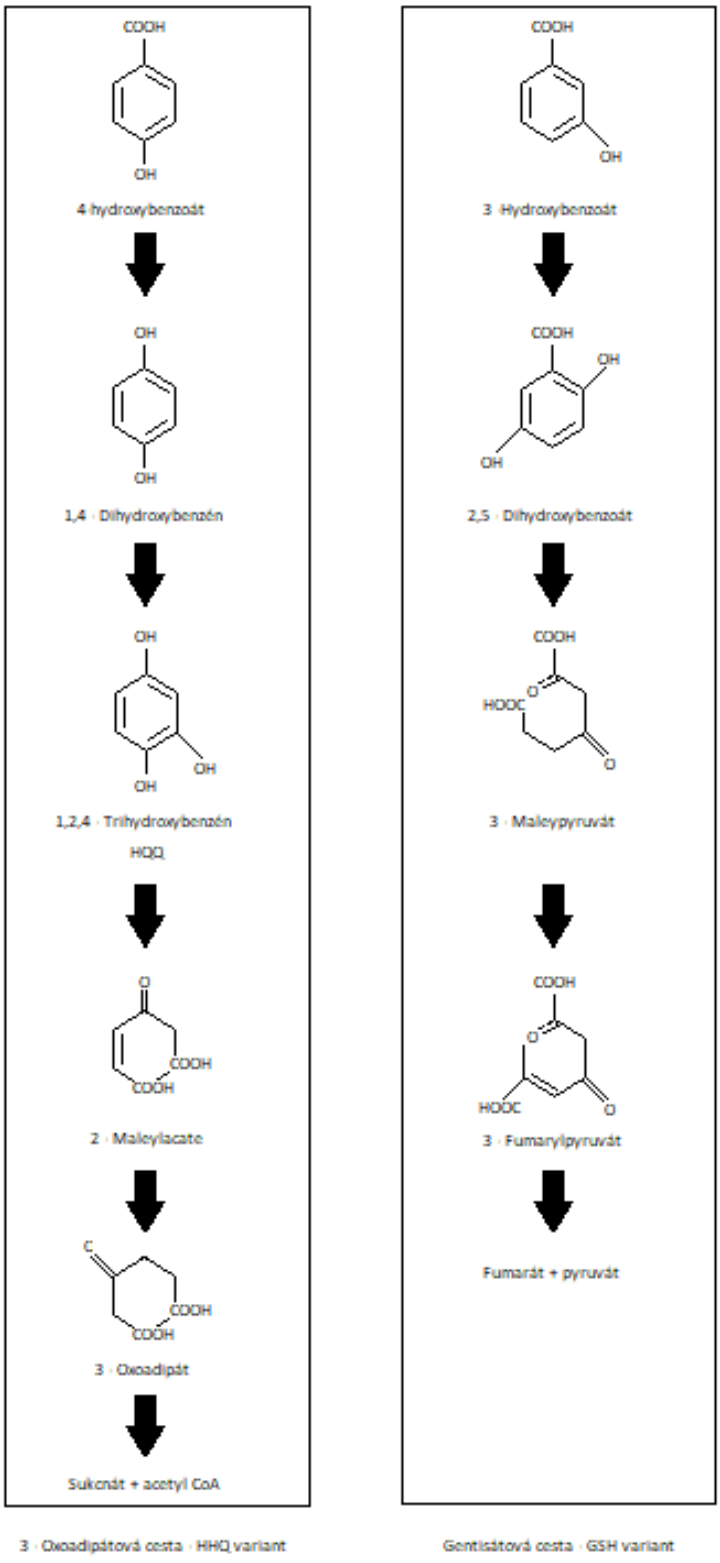
V našej práci budeme vychádzať z práce od Zemana a kol. [28], z ktorej využijeme namerané hodnoty génovej expresie kvasinky *C. parapsilosis* rastúcej na 4–OH a 3–OH

médiách a glukóze.

1.5 Meranie génovej expresie

Génová expresia sa meria rôznymi spôsobmi, my popíšeme iba jeden z nich. RNA-seq je protokol určený na sekvenovanie mRNA extrahovaných z bunkiek, ktorý generuje milióny krátkych sekvencií v jednom behu. Tieto fragmenty sa používajú na meranie úrovne génovej expresie a na identifikovanie nových zostrihových variánt génov [25].

Najskôr sa populácia RNA konvertuje na knižnicu fragmentov cDNA s adaptérmi pripojenými na jeden alebo oba konce. Každá molekula, s amplifikáciou alebo bez nej, sa potom sekvenuje vysoko výkonným spôsobom, aby sa získali krátke sekvencie z jedného konca (sekvenovanie na jednom konci) alebo z oboch koncov (párové sekvenovanie). Po sekvenovaní sú výsledky buď zarovnané s referenčným genómom alebo referenčnými transkriptmi, alebo zostavené de novo bez genómovej sekvencie, aby sa vytvorila transkripčná mapa, ktorá pozostáva z transkriptov štruktúry úrovne expresie pre každý gén v genóme [27].



Obr. 1.2: 3–OAP a GP metabolické dráhy podľa [15]

Kapitola 2

Zdroje dát

V tejto kapitole popíšeme zdroje, z ktorých sme čerpali dáta. Jeden z nich je databáza uniprot, ktorá obsahuje informácie o proteínoch a ďalší zdroj je databáza Gene Ontology, kde sú popísané rôzne metabolické dráhy. V závere kapitoly spomínamé zdroj nameraných dát o génovej expresii.

2.1 UniProt

UniProt je databáza proteínových sekvencií a ich anotácii. Dáta nachádzajúce sa v UniProte sú dostupné zdarma pre verejnosť a využívané tisíckami vedcov každý deň [12].

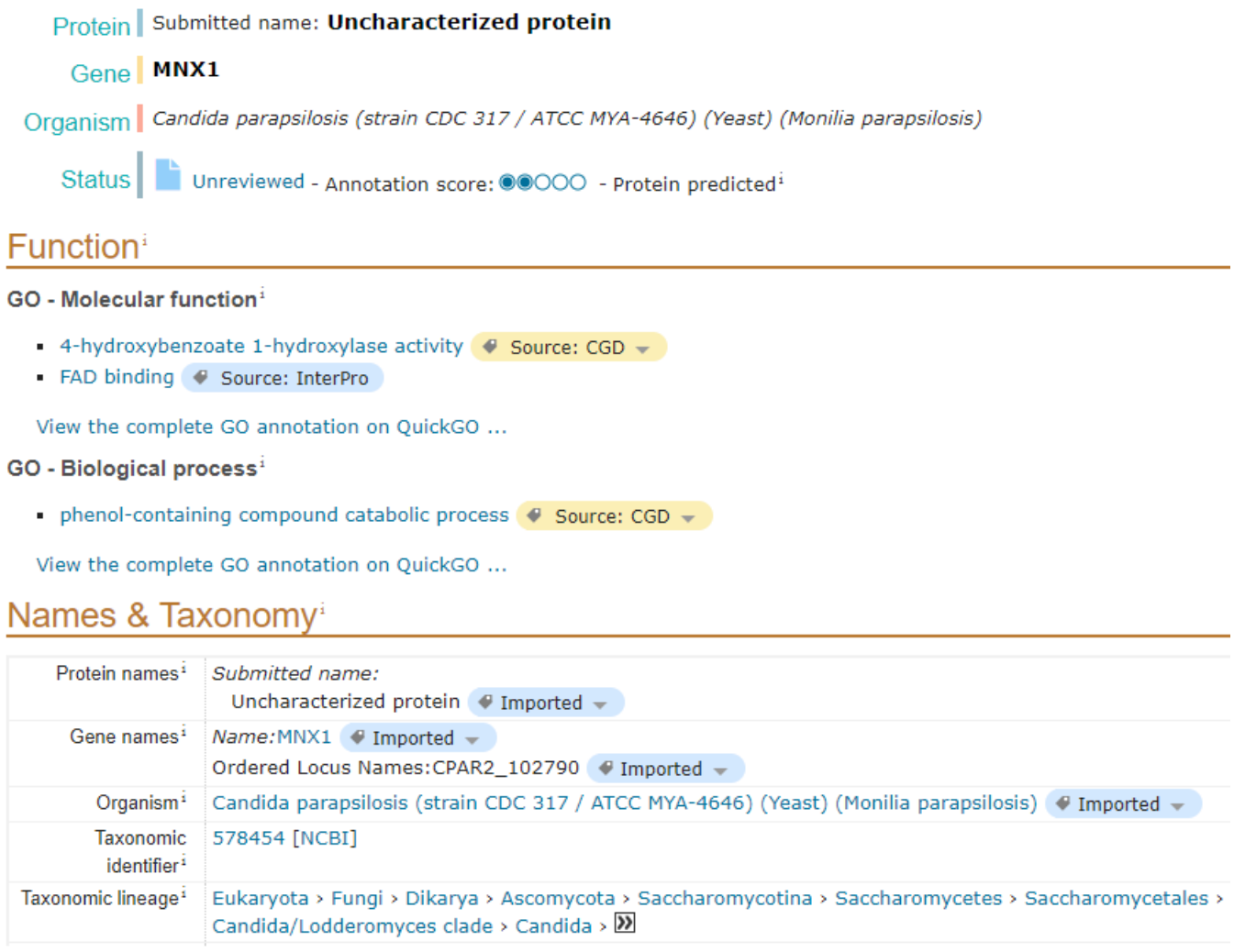
Uniprot sa skladá z viacerých častí. Jedna z nich je *UniProtKB/Swiss-Prot* a obsahuje manuálne anotované dáta. Postupne ako sú nové proteíny experimentálne charakterizované, pribúdajú nové sekvencie do tejto časti. Zvyšné sekvencie sa nachádzajú v časti *UniProtKB/TrEMBL*. Záznamy v tejto časti nie sú manuálne anotované ale dopĺňané automaticky generovanými anotáciami.

Ďalšia časť databázy, nazývaná aj archív Uniprotu je *UniParc*. Je to neredundantná databáza všetkých proteínových sekvencií, získaných z niekoľkých verejných zdrojov. [16]. Obsahuje iba proteínové sekvencie, ostatné informácie o proteíne sa dajú získať použitím krížových odkazov na zdrojovú databázu [10].

Posledná z väčších častí Uniprotu je *UniRef*. Poskytuje zoskupené sady sekvencií z *UniProtKB* a z *UniParc* databázy na získanie úplneho pokrytia sekvencie pri niekoľkých rozlíšeniach. Jej hlavnou výhodou je, že zakrýva nadbytočné sekvencie. V databáze *UniRef*100 sú identické sekvencie a ich fragmenty spojené do jedného klastra. Databázy *UniRef*90 a *UniRef*50 sú tvorené zoskupením sekvencií z *UniRef*100, ktoré mali sekvenčnú identitu 90% a 50% [23].

Databáza ponúka veľké množstvo informácii. Dá sa z nich vyčítať napríklad názov génu, ktorý kóduje daný proteín, kde v bunke sa vyskytuje daný proteín, jeho štruktúra

a aj sekvencia aminokyselín. Navyše obsahuje krížové odkazy na iné databázy, ktoré popisujú funkciu a rolu proteínu v biologických dráhach. Na obrázku 2.1 môžete vidieť príklad - proteín, ktorému zatiaľ v databáze Uniprot neurčili meno, ale kóduje ho gén MNX1, ktorý Holešova a kol. [15] našli v genóme kvasinky *C. parapsilosis* a o ktorom zistili, že kóduje flavoproteín monooxygenázu, ktorá katalyzuje prvý krok v 3–OAP metabolickej dráhe čo je 4-hydroxybenzoát 1-hydroxylázová aktivita 1.2.



Obr. 2.1: Príklad záznamu pre proteín z databázy Uniprot

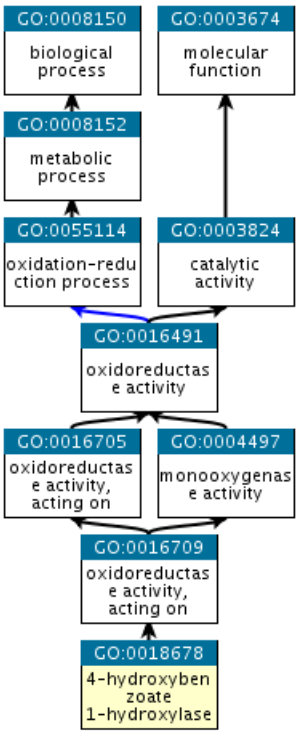
2.2 Gene Ontology

Databáza Gene Ontology obsahuje informácie, ktoré pokrývajú niekoľko domén molekulárnej a bunkovej biológie, vo forme anotácií génov, génových produktov a sekvencií. Ďalej definuje kategórie génových funkcií (GO kategórie) čo sú niečo ako spoločný jazyk pre anotácie. GO kategórie majú medzi sebou špecifikované hierarchické vzťahy ako je vidieť na obrázku 2.2, ktorý zobrazuje graf GO kategórie GO:0018678 zobrazujúc

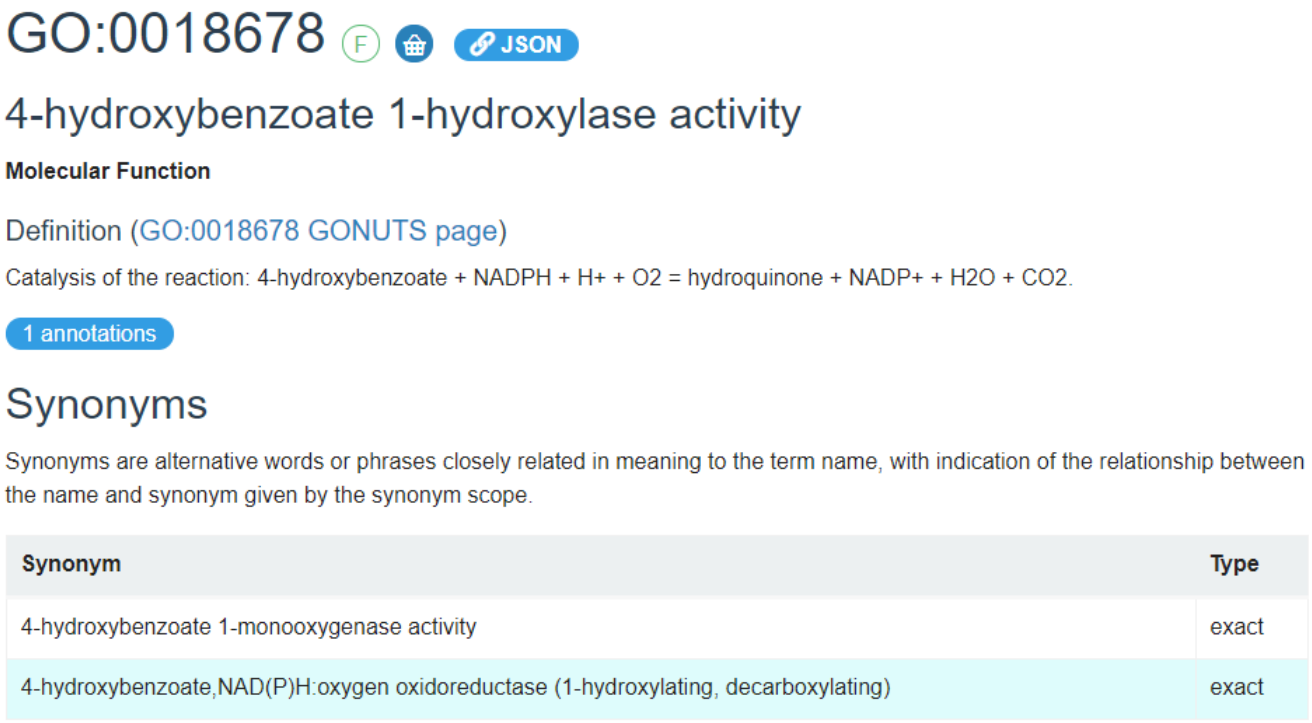
nadkategórie. Databáza ďalej obsahuje základné pojmy ako názov a definícia kategórie (viď pbrázok 2.3). Vzťahy medzi GO kategóriami sú definované pod týmito názvami: is a, part of, regulates, negatively regulates a positively regulates.

Projekt vznikol v roku 1998 ako spolupráca medzi tromi databázami modelových organizmov a to FlyBase, ktorá bola zameraná na *Drosophilu*, ďalej sa projektu zúčastnila *Saccharomyces* Genome Database a Mouse Genome Informatics projekt. Od tých čias sa GO rozrástlo o informácie o rastlinných, zvieracích a mikrobiálnych genómoch [14].

Štruktúra GO kategórií sa neustále vyvíja, mení spolu s novými vedeckými objavmi a predstavuje najaktuálnejšie biologické poznatky.



Obr. 2.2: Štruktúra GO kategórií


GO:0018678 F JSON

4-hydroxybenzoate 1-hydroxylase activity

Molecular Function

Definition (GO:0018678 GONUTS page)

Catalysis of the reaction: 4-hydroxybenzoate + NADPH + H+ + O2 = hydroquinone + NADP+ + H2O + CO2.

1 annotations

Synonyms

Synonyms are alternative words or phrases closely related in meaning to the term name, with indication of the relationship between the name and synonym given by the synonym scope.

Synonym	Type
4-hydroxybenzoate 1-monooxygenase activity	exact
4-hydroxybenzoate,NAD(P)H:oxygen oxidoreductase (1-hydroxylating, decarboxylating)	exact

Obr. 2.3: Príklad záznamu pre GO kategóriu v databáze Gene Ontology - na obrázku je vidieť GO id, názov príslušnej kategórie ako aj jej definíciu a synonymá.

2.2.1 Ontológie

Pre GO projekt existuju tri nezávislé ontológie: [14].

Ontológia „Molekulárna funkcia“ popisuje rôzne aktivity na molekulárnej úrovni, napríklad katalytické aktivity. GO molekulárne funkcie popisujú aktivity a nešpecifikujú kde, kedy a prečo sa akcia vykonáva. Taktiež sa nespomínajú entity, ktoré akciu spôsobujú. Príklad termínu molekulárnej funkcie je kinázová aktivita - kináza je enzým, ktorý prenáša fosfátové skupiny na molekuly [21] Špecifickejší termín je napríklad 6-fosfofruktokinázová aktivita, čo je podtyp kinázovej aktivity [14].

V ontológii „Biologický proces“ sú popísané biologické ciele dosiahnuté molekulárnymi funkciami. Napríklad, proces ako bunková smrť môže mať podtyp apoptóza ale aj subproces apoptopická kondenzácia chromozómov [14].

Ontológia „Zložky bunky“ obsahuje informácie o miestach, ktoré sú na úrovni subcelulárnych štruktúr a o makromolekulových komplexoch. Príklad termínov v tejto ontológii predstavujú vnútorná jadrová membrána ako štruktúra bunky a ubichinónový ligázový komplex.

2.2.2 Anotácie

GO anotácia vyjadruje funkciu konkrétneho génu a musí k nej byť priradený konkrétny zdroj, čo môže byť literatúra alebo iná databáza. Spája génový produkt s kategóriou. Príkladom anotácie je gén *MNX*1, ktorý je anotovaný GO termínom *GO:*0018678 4-*hydroxybenzoate* 1-*hydroxylase activity*, pričom zdroj tejto anotácie je *CGD, CAL*0000155825. O správnosť dát sa stará konzorcium odborných biokurátorov po celom svete, ktorí čítajú vedecké práce, identifikujú gény a prideľujú čo najpresnejšie GO termíny [24].

Štruktúra Gene Ontology databázy, ontológie a anotácie umožňujú presnejšie zodpovedanie rôznych dotazov, ako napríklad: „Aké sú všetky funkcie pre ľudský gén ABCA1?“ Alebo „Aké všetky gény sú zapojené do procesu opravy chybného párovania DNA?“ Vďaka schopnosti podporovať dotazy je GO databáza jeden z najzákladnejších nástrojov v biomedicínskom výskume [24].

2.3 Výsledky merania génovej expresie

V našej práci pracujeme s tabuľkou S5 z článku [28], v ktorej sa nachádzajú spracované dáta z RNA-seq. Ukážku dát v tabuľke môžeme vidieť na obrázku 2.4.

Tabuľka obsahuje deväť stĺpcov a sú v nej údaje o diferenciálnej expresii v bunkách kultivovaných v kontrolných podmienkach – v médiu SD (základné médium) a v dvoch alternatívnych hydroxybenzoátových substrátoch – S3OH a S4OH. Mitochondriálne transkripty a tRNA boli zo sekvenovania vylúčené.

Prvý zo skupiny týchto štyroch stĺpcov obsahuje normalizovanú priemernú expresiu v základnej podmienke – SD.

V druhom stĺpci sa nachádzajú normalizované expresie v konkrétnom testovanom stave. V treťom stĺpci je logaritmus podielu týchto hodnôt. Ak je hodnota kladná, znamená to, že expresia génu stúpala a ak záporná tak naopak.

Posledný stĺpec obsahuje p-hodnoty pre test, či ide o štatisticky významné zmeny expresie (bez korekcie viacnásobného testovania).

RNA vzorky boli sekvenované Illuminou GAIIx použitím párových čítaní dĺžky 50 nukleotidov. Čítania boli zarovnané s referenčným genómom *C.parapsilosis* použitím programu tophat2 v2.0.6.

V prvom stĺpci tabuľky sa nachádza CPAR názov génu. Ďalšie štyri stĺpce sú venované údajom získaným z rastu kvasinky *C. parapsilosis* na 3–hydroxybenzoáte a posledné štyri z rastu na 4–hydroxybenzoáte.

	A	B	C	D	E	F	G	H	I
1	**Table S5 Expression data obtained by RNA-seq analysis.**								
2									
3	**Systematic name**	**SDS3OH**	**S3OH basemear**	**S3OH $\log_2$(FC)**	**S3OH P-value**	**SDS4OH**	**S4OH basemear**	**S4OH $\log_2$(FC)**	**S4OH P-value**
4	CPAR2_100020	43,20	22,3793	-0,948827	0,34095	40,86	22,077	-0,888093	0,43045
5	CPAR2_100030	223,22	126,767	-0,816304	0,46135	211,13	149,929	-0,493849	0,7138
6	CPAR2_100040	724,13	225,716	-1,68174	0,22725	684,90	182,769	-1,90587	0,24065
7	CPAR2_100050	130,00	169,675	0,384312	0,73035	122,95	128,58	0,0645631	0,96105
8	CPAR2_100060	377,39	122,742	-1,62041	0,1687	356,94	89,8329	-1,99037	0,14545
9	CPAR2_100070	88,28	52,6108	-0,746714	0,4586	83,50	49,2264	-0,762288	0,48075
10	CPAR2_100080	42,13	7,59214	-2,47222	0,0302	39,85	8,47071	-2,23388	0,06095
11	CPAR2_100090	299,79	41,4992	-2,85277	0,0468	283,54	75,217	-1,91444	0,24265
12	CPAR2_100100	42,66	156,428	1,87447	0,0718	40,35	254,54	2,65721	0,02925
13	CPAR2_100110	151,09	144,471	-0,0645938	0,9535	142,90	115,049	-0,312775	0,81225
14	CPAR2_100120	15,08	55,3229	1,8756	0,0701	14,26	54,1199	1,92424	0,08925
15	CPAR2_100130	230,92	342,388	0,568232	0,67955	218,41	311,456	0,511981	0,74805
16	CPAR2_100140	117,24	133,009	0,182053	0,86075	110,89	94,8201	-0,225847	0,85175
17	CPAR2_100150	20,15	20,1527	0,000466816	0,9988	19,05	25,6115	0,42664	0,6983

Obr. 2.4: Tabuľka tableS5 - Ukážka RNA-seq údajov z tabuľky S5 [28]

Kapitola 3

Spracovanie dát

V tejto kapitole popíšeme prečo a ako sme postupovali. Na začiatku je celkový prehľad nášho prístupu. V ďalších podkapitolách sa nachádzajú detailnejšie informácie o spracovaní dát spolu s ukážkami výstupov. Na záver kapitoly popisujeme štatistické testy použité v našej práci.

3.1 Celkový postup

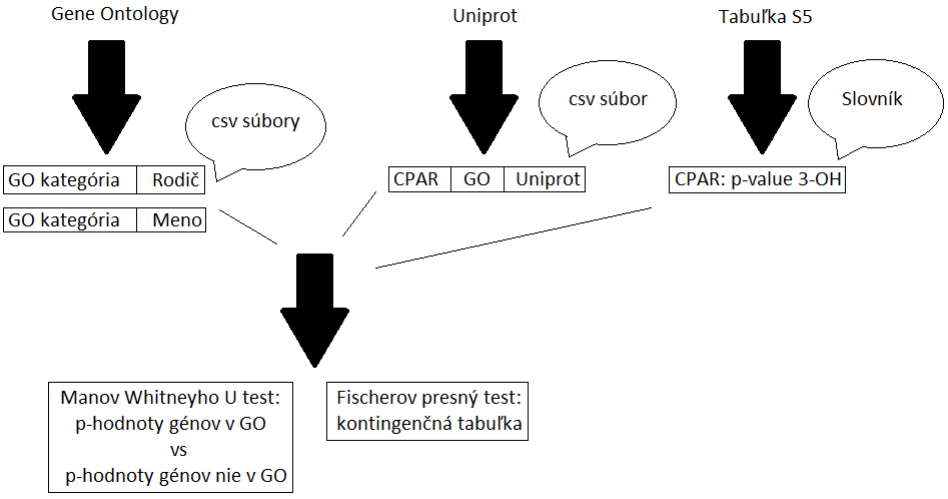
Na úvod uvedieme cieľ, ku ktorému sa postupne dopracujeme. Budeme robiť dva štatistické testy, ktoré nám pre každú GO kategóriu vrátia p-hodnotu. Jeden z nich, Mannov-Whitneyho U test pracuje s dvoma množinami p-hodnôt. Prvá sa skladá z p-hodnôt génov, ktoré sú v určitej GO kategórií a druhá množnina z p-hodnôt, ktoré nie sú v danej GO kategórii. Tento test vypočíta nové p-hodnoty pre každú GO kategóriu. Druhý test, Fischerov presný test pracuje s kontingenčnou tabuľkou. Pre ňu potrebujeme počty génov, ktoré sú v GO kategórií a počty génov, ktoré nie sú v GO kategórii a zároveň, počty génov s menšou p-hodnotou ako 0.01 a s väčšou a rovnou ako 0.01.

Na obrázku 3.1 je schéma našej práce. Z databázy Gene Ontology sme vyrobili dva súbory vo formáte csv. Prvý obsahoval dva stĺpce - GO kategóriu a k nej príslušných rodičov. Druhý súbor obsahoval taktiež dva stĺpce - GO kategóriu a jej názov. Z dát z databázy Uniprot sme vyrobili jeden súbor vo formáte csv s tromi stĺpcami. Pre gén s uniprot id sú tam jeho GO kategórie, a CPAR označenie. Nakoniec sme spracovali tabuľku S5, z ktorej sme vytvorili dva slovníky pre obidva hydroxybenzoáty, v ktorých ako kľúč bol CPAR názov génu a ako hodnota p-hodnota diferenčnej expresie génov.

V tomto kroku si vysvetlíme, ako sme využili spracované súbory pre testy. Začneme Manovým Whitneyhovým testom. Ako sme spomínali, pracuje s množinami p-hodnôt. Tie máme z tabuľky S5. Lenže potrebujeme také p-hodnoty génov, ktoré patria do GO kategórie a všetkých jej potomkov. Na zistenie všetkých potomkov nám slúži súbor go_parent a rekurzívna funkcia, ktorá vyhľadáva všetkých potomkov. Keď už máme

pre GO kategóriu všetkých potomkov a pre gény p-hodnoty, potrebujeme zistiť, ktoré gény patria do potrebných GO kategórii. Na to slúži súbor cpar_go_uniprot vytvorený z databázy Uniprot, ktorý obsahuje pre GO kategóriu zoznam jej CPAR génov.

Pre Fischerov presný test potrebujeme vedieť počty génov v GO kategórii a mimo GO kategórie a zároveň počty génov s p-hodnotou menšou ako 0.01 a väčšou a rovnou ako 0.01. K požadovaným údajom sa vieme dopracovať rovnako ako pri Mannovom-Whitneyovom U teste len nebudeme vytvárať množiny génov ale ich iba počítať.



Obr. 3.1: Schéma našej práce

3.2 Spracovanie dát z databázy Uniprot

Na stránke databázy Uniprot - www.uniprot.org sme vyhľadali dáta o kvasinke *Candida parapsilosis*. Po otvorení stránky sme zvolili vo vyhľadávači možnosť „Proteomes“, napísali **Candida parapsilosis** a stlačili tlačidlo na stiahnutie. Zoznam génov sme si stiahli vo formáte XML. Súbor obsahuje jedinečný identifikátor pre gén, uniprot názov génu, meno proteínu, CPAR názov génu a GO kategórie, kvôli ktorým sme si vybrali túto databázu. Na obrázku 3.2 môžete vidieť ukážku dát.

Na spracovanie formátu XML do prehľadnejšej formy, do csv súboru sme naprogramovali skript v jazyku Python a použili knižnicu lxml a jej triedu etree [3]. XML je hierarchický formát údajov a najlepší spôsob ako ho prezentovať je strom. Knižnica lxml má na tento účel dve triedy - ElementTree predstavuje celý XML dokument ako strom a Element predstavuje jeden uzol v tomto strome. Interakcie s celým dokumentom (čítanie a zápis do/zo súborov) sa zvyčajne vykonávajú na úrovni ElementTree. Interakcie s jedným prvkom XML a jeho čiastkovými prvkami sa vykonávajú na úrovni prvku [3].

Najskôr sme si napísali niekoľko funkcii. Prvá z nich getName nám vráti názov

elementu. Tento krok je veľmi dôležitý lebo ako je vidieť z obrázku, informácie sú uložené v elementoch a vložené v elemente <name> </name>. Funkcia getUniprotId nám vráti identifikátor génu v Uniprot databáze.

Identifikátor typu CPAR2 sa nachádza v elemente <gene> </gene>. Jedno uniprot id môže mať viacero CPAR id ako uvádza aj príklad na obrázku. Na zistenie všetkých CPAR id slúži funkcia getCparName, ktorá prechádza cez každý záznam v elemente <gene> </gene>.

```
...
<entry dataset="TrEMBL" created="2012-01-25" modified="2019-04-10" version="34">
<accession>G8BFP5</accession>
<name>G8BFP5_CANPC</name>
<protein>
<recommendedName>
<fullName evidence="1">Histone H3</fullName>
</recommendedName>
</protein>
<gene>
<name type="ordered locus" evidence="9 10">CPAR2_107010</name>
<name type="ordered locus" evidence="8 11">CPAR2_203220</name>
</gene>
...
<dbReference type="CGD" id="CAL0000154377">
<property type="gene designation" value="CPAR2_107010"/>
</dbReference>
<dbReference type="CGD" id="CAL0000148119">
<property type="gene designation" value="CPAR2_203220"/>
</dbReference>
<dbReference type="Proteomes" id="UP000005221">
<property type="component" value="Chromosome 1"/>
</dbReference>
<dbReference type="Proteomes" id="UP000005221">
<property type="component" value="Chromosome 2"/>
</dbReference>
<dbReference type="GO" id="GO:0000786">
<property type="term" value="C:nucleosome"/>
<property type="evidence" value="ECO:0000501"/>
<property type="project" value="UniProtKB-KW"/>
</dbReference>
<dbReference type="GO" id="GO:0005634">
<property type="term" value="C:nucleus"/>
<property type="evidence" value="ECO:0000501"/>
<property type="project" value="UniProtKB-SubCell"/>
</dbReference>
...
```

Obr. 3.2: Ukážka xml súboru, kde je vidieť aj uniprot id, CPAR názov a GO kategórie

Ako posledné sme potrebovali získať GO kategórie. Znova, jeden gén môže obsahovať viacero GO anotácií. Funkcia getGO prejde všetky elementy typu <dbReference> a vráti každý záznam v elemente, ktorý má type="GO".

Po zadefinovaní funkcií, sme mohli prejsť k spracovaniu dát. Na XML súbor sme použili knižnicovú funkciu etree.parse, ktorá nám vytvorila ElementTree z našich dát. Na takto spracované dáta sme aplikovali knižnicovú funkciu getroot, ktorá nám vrátila

zoznam koreňových elementov pre náš strom. Potom sme už iba prechádzali cez každý koreňový element a aplikovali na jeho záznamy horeuvedené funkcie. Dáta, ktoré sme takto získavali, sme uložili do matice so stĺpcami 'uniprot', 'CPAR', 'GO'. Znamená to, že pre každý uniprot názov génu sme si pamätali jeho CPAR názvy a GO kategórie, ktoré mu patria.

Ako posledný krok sme maticu uložili do csv formátu. Použili sme na to knižnicu pandas a modul DataFrame. DataFrame je dvojrozmerná dátová štruktúra, kde sú dáta uložené v tabuľkovej forme v riadkoch a stĺpcoch [8]. Z matice sme vytvorili DataFrame a naň zavolali funkciu to_csv, ktorá nám vytvorila tabuľkový súbor v csv formáte - cpar_go_uniprot.csv. Ukážku výstupu môžete vidieť na obrázku 3.2. Na otvorenie csv formátu sme použili tabuľkový editor LibreOffice.

Obr. 3.3: Spracované dáta z databázy Uniprot

	A	B	C
1	**CPAR**	**GO**	**uniprot**
2	CPAR2_807400	GO:0102131,GO:0102132,GO:0004316,GO:0004315,GO:0004321,	FAS2_CANPC
3	CPAR2_205060	GO:0000235,GO:0005935,GO:0000307,GO:0010494,GO:0005783,	G8BG79_CANPC
4	CPAR2_400020	GO:0005737,GO:0003855,GO:0003856,GO:0003866,GO:0005524,	G8BHT6_CANPC
5		GO:0047011,GO:0004033,GO:0042180	CPRC2_CANPA
6	CPAR2_701810	GO:0032153,GO:0005935,GO:0000778,GO:0001400,GO:0005847,	G8BKA4_CANPC
7	CPAR2_302650	GO:0005835,GO:0008659,GO:0047451,GO:0004313,GO:0004314,	G8B9H7_CANPC
8	CPAR2_109990	GO:0005935,GO:1990023,GO:0097431,GO:0044732,GO:0005634,	G8B632_CANPC
9	CPAR2_106590	GO:0000262,GO:0042645,GO:0005634,GO:0003690,GO:0003697,	MG101_CANPC
10	CPAR2_804190	GO:0036266,GO:0005829,GO:0000837,GO:0000839,GO:0005634,	G8B9Z7_CANPC
11	CPAR2_206900	GO:0030176,GO:0009055,GO:0020037,GO:0046872,GO:0004768,	G8BCQ2_CANPC
12	CPAR2_703030	GO:0005945,GO:0005739,GO:0003872,GO:0005524,GO:0046872,	G8BKM4_CANPC
13		GO:0047011,GO:0042180	CPRC1_CANPA
14	CPAR2_211840	GO:0005743,GO:0005758,GO:0005634,GO:0005524,GO:0016909,	G8BE43_CANPC
15	CPAR2_207220	GO:0005737,GO:0030429,GO:0030170,GO:0034354,GO:0043420,	G8BCT4_CANPC
16	CPAR2_201880	GO:0071162,GO:0005737,GO:0042555,GO:0097373,GO:0005656,	G8BFB2_CANPC
17	CPAR2_300490	GO:0022626,GO:0015934,GO:0042788,GO:0015935,GO:0005524,	G8B8W1_CANPC
18	CPAR2_204230	GO:0005739,GO:0005730,GO:0005654,GO:0008409,GO:0017108,	G8BFZ6_CANPC
19	CPAR2_203530	GO:0005737,GO:0005524,GO:0004781,GO:0019344,GO:0070814,	G8BFS6_CANPC
20	CPAR2_303020	GO:0005737,GO:0005634,GO:0046872,GO:0035529,GO:0047429,	G8B9L2_CANPC
21	CPAR2_701690	GO:0031966,GO:0043596,GO:0005524,GO:0043141,GO:0051880,	G8BK92_CANPC
22	CPAR2_201350	GO:0005737,GO:0005634,GO:0005524,GO:0046872,GO:0004747,	G8BF34_CANPC
23	CPAR2_803640	GO:0005737,GO:0004019,GO:0005525,GO:0000287,GO:0061483,	G8BCE9_CANPC
24		GO:0000790,GO:0003700,GO:0046872,GO:0043565,GO:0036164,	BCR1_CANAL
25	CPAR2_602110	GO:0005829,GO:0005758,GO:0004017,GO:0005524,GO:0006172,	G8B5A0_CANPC
26		GO:0046658,GO:0009986,GO:0005933,GO:0005576,GO:0009277,	CSA1_CANAL

Celkový počet génov stiahnutých z databázy Uniprot je 6034. Z toho 258 uniprot génov nemá CPAR označenie a 1479 nemá žiadnu GO kategóriu. Najviac GO kategórii s počtom 65 má gén G8BG79_CANPC, ktorému zodpovedá CPAR označenie CPAR2_205060. Funkcia génu ešte nebola manuálne charakterizovaná.

3.3 Spracovanie dát z databázy Gene Ontology

Zo stránky http://geneontology.org/ sme stiahli súbor go-basic.obo. Na výber je viacero typov súborov a tento sme vybrali, lebo najlepšie zodpovedal našim požiadavkam a to

pre GO kategóriu zistiť jej popis a rodičov. Tento typ súboru je základná verzia GO kategórii, ktorá na definíciu rodiča používa "is_a"[5].

Obr. 3.4: Ukážka formátu obo

```
[Term]
id: GO:0000001
name: mitochondrion inheritance
namespace: biological_process
def: "The distribution of mitochondria, including the mitochondrial genome, into daughter
cells after mitosis or meiosis, mediated by interactions between mitochondria and the
cytoskeleton." [GOC:mcc, PMID:10873824, PMID:11389764]
synonym: "mitochondrial inheritance" EXACT []
is_a: GO:0048308 ! organelle inheritance
is_a: GO:0048311 ! mitochondrion distribution

[Term]
id: GO:0000002
name: mitochondrial genome maintenance
namespace: biological_process
def: "The maintenance of the structure and integrity of the mitochondrial genome; includes
replication and segregation of the mitochondrial chromosome." [GOC:ai, GOC:vw]
is_a: GO:0007005 ! mitochondrion organization

[Term]
id: GO:0000003
name: reproduction
namespace: biological_process
alt_id: GO:0019952
alt_id: GO:0050876
def: "The production of new individuals that contain some portion of genetic material
inherited from one or more parent organisms." [GOC:go_curators, GOC:isa_complete, GOC:jl,
ISBN:0198506732]
subset: goslim_agr
subset: goslim_chembl
subset: goslim_flybase_ribbon
subset: goslim_generic
subset: goslim_pir
subset: goslim_plant
synonym: "reproductive physiological process" EXACT []
xref: Wikipedia:Reproduction
is_a: GO:0008150 ! biological_process
```

Formát súboru môžete vidieť na priloženom obrázku 3.4. Nepodarilo sa nám nájsť žiadny parser formátu obo, tak sme nemohli použiť žiadnu knižnicu a všetko sme programovali ručne. Na obrázku je vidieť, ako sú informácie ohľadom jednej GO kategórie uložené. Názov je označený ako "id: ". Ďalej sme potrebovali názov danej GO kategórie, ktorá je v riadku začínajúcom reťazcom "name: ". A ako posledný údaj sme zisťovali rodičov danej GO kategórie. Na to slúžia riadky označené ako "is_a".

Ako je vidieť v ukážke, tak niektoré GO kategórie majú viacero rodičov. Na uloženie informácii sme použili defaultdict, ktorý si ako kľúč pamätá GO a ako hodnotu list rodičov. Rovnako sme postupovali aj pri ukladaní názvov GO kategórii ale tam nám stačil dict. Na obrázku 3.7 a 3.6 môžete vidieť ukážku súborov go_parent a go_name zobrazených v textovom editore.

Ako posledné sme dáta uložili do dvoch súborov vo formáte csv, použitím DataF-

rame z knižnice Pandas. Prvý súbor má ako prvý stĺpec identifikátor GO kategórie, v druhom stĺpci jej názov. Druhý súbor má ako druhý stĺpec zoznam rodičov.

	A	B
1	**GO**	**Parent**
2	GO:0014802	GO:0098827
3	GO:0047914	GO:0016881
4	GO:0002339	GO:0002376
5	GO:0021516	GO:0048856
6	GO:0101011	GO:0052745
7	GO:0033209	GO:0019221
8	GO:0047097	GO:0004497
9	GO:0047097	GO:0016705
10	GO:0042433	GO:0042431
11	GO:0042433	GO:0042436
12	GO:0015645	GO:0016878
13	GO:1990923	GO:0032991
14	GO:0061192	GO:0010639
15	GO:0061192	GO:0032889
16	GO:0050016	GO:0016705
17	GO:0045507	GO:0030368
18	GO:0075178	GO:0075148
19	GO:0075178	GO:0075177
20	GO:0017073	GO:0016811
21	GO:0033803	GO:0008171
22	GO:0033803	GO:0008757
23	GO:2000367	GO:0017158
24	GO:2000367	GO:0060046
25	GO:0052562	GO:0052553
26	GO:0052562	GO:0052561

Obr. 3.5: Ukážka súboru go parent

	A	B
1	**GO**	**Name**
2	GO:0009406	obsolete virulence
3	GO:0061681	Entner-Doudoroff pathway through gluconate to D-glyceraldehyde-3-phosphate
4	GO:0010734	negative regulation of protein glutathionylation
5	GO:0031288	sorocarp morphogenesis
6	GO:0001721	obsolete intermediate filament associated protein
7	GO:0034669	integrin alpha4-beta7 complex
8	GO:0034990	nuclear mitotic cohesin complex
9	GO:0071353	cellular response to interleukin-4
10	GO:1903940	negative regulation of TORC2 signaling
11	GO:0036140	peptidyl-asparagine 3-dioxygenase activity
12	GO:0048835	specification of decreased petal number
13	GO:0103012	ferredoxin-thioredoxin reductase activity
14	GO:0009473	obsolete cytochrome c7 (triheme)
15	GO:0006424	glutamyl-tRNA aminoacylation
16	GO:0099512	supramolecular fiber
17	GO:0099116	tRNA 5'-end processing
18	GO:0048701	embryonic cranial skeleton morphogenesis
19	GO:0070149	mitochondrial glutamyl-tRNA aminoacylation
20	GO:0017106	obsolete activin inhibitor activity
21	GO:1900586	arugosin catabolic process
22	GO:0097586	dolichyl-phosphate-mannose-protein mannosyltransferase Pmt4p homodimer complex
23	GO:0001108	bacterial-type RNA polymerase holo enzyme binding
24	GO:0009007	site-specific DNA-methyltransferase (adenine-specific) activity
25	GO:0010313	phytochrome binding
26	GO:0047932	glucosamine N-acetyltransferase activity

Obr. 3.6: Ukážka súboru go name

Počet všetkých GO kategórii, ktoré sme získali zo súboru je 45003. Z toho 17543 je rodičom inej kategórie. Najviac rodičov má kategória GO:1902180 s počtom 10 a jej názov je katabolický proces látky *Verruculogen*.

3.4 Spracovanie dát z tabuľky S5

Na spracovanie dát o expresii z tabuľky S5 sme použili knižnicu xlrd. Táto knižnica je určená na čítanie dát alebo aj ich úpravu vo formátoch programu microsoft Excel [11].

Najskôr sme otvorili dokument funkciou xlrd.open_workbook. Potom sme načítali prvý hárok funkciou sheet_by_index. Následne sme vytvorili dva slovníky, kde ako kľúč slúžilo CPAR označenie génu a hodnota bola p-hodnota pre daný gén na 3–OH pre jeden slovník a 4–OH pre druhý slovník.

Nasledovalo prechádzanie všetkých riadkov tabuľky, kde sme zisťovali či sa reťazec 'CPAR' sa nachádza v nultom stĺpci - to aby sme sa vyhli prázdnym riadkom na začiatku súboru. Ak podmienka bola splnená, zapamätali sme si v slovníkoch hodnoty zo štvrého a ôsmeho stĺpca ak číslujeme od nuly.

Počet génov v tabuľke je 5821. V stĺpci p-hodnôt pre 3-OH má 134 génov p-hodnotu menšiu ako 0.01 a počet génov pre 4-OH s rovnako malou p-hodnotou je 152.

3.4.1 Fisherov presný test

V štúdiach sa často detekujú významné obohatenia alebo ochudobnenia GO kategórií v rámci triedy požadovaných génov, zvyčajne sú to triedy významne diferencovane exprimovaných (DE) génov [20].

Uvažujme napríklad množinu všetkých génov v genóme a zaujímajme sa o gény z určitej GO kategórie. Cieľom je stanoviť či trieda DE génov je obohatená na gény danej GO kategórie vzhľadom na množinu všetkých génov [20].

Nech H_0 označuje nulovú hypotézu, že vlastnosť génu patriaceho do skúmanej GO kategórie je nezávislá od vlastnosti DE, alebo inak povedané, že gény DE sa náhodne vyberajú z celkovej populácie génov [20].

Pred vykonaním testu si zostavíme kontingenčnú tabuľku veľkosti 2x2. Jej prvý stĺpec udáva počet génov v GO kategórií, druhý počet génov, ktoré nie sú v GO kategórií. Prvý riadok sú DE gény a druhý riadok je zvyšok génov zo skupiny skúmaných génov [20]. Príklad kontingenčnej tabuľky pre GO:0018678 je vidieť na obrázku

GO:0018678	V GO	Nie v GO
p-value < 0.01	1	133
p-value > 0.01	0	5687

Obr. 3.7: Kontingenčná tabuľka pre GO:0018678 pre rast na 3–OH médiu

Na výpočet Fischerovho presného testu sme napísali skript v jazyku Python a použili knižnicu stats [4]. Vo funkcii sme zvolili nastavenie alternative='greater', ktoré testuje nulovú hypotézu.

3.4.2 Mannov-Whitneyho U test

Mannov-Whitneyho U test tiež známy ako Wilcoxonov test, testuje rozdiely medzi dvoma skupinami jednej ordinálnej premennej (čo sú premenné merateľné na poradovej škále) bez špecifickej distribúcie [18]. Nech H_0 označuje nulovú hypotézu, ktorá hovorí, že dve skupiny pochádzajú z rovnakého rozdelenia [19].

Mannov-Whitneyho U test je založený na porovnaní každého pozorovania z prvej skupiny s každým pozorovaním z druhej skupiny. Ak obidve skupiny pochádzajú z rovnakej populácie ako predpokladá nulová hypotéza, každý údaj z prvej skupiny má rovnakú šancu byť väčší alebo menší ako údaj z druhej skupiny. Nulová hypotéza je zamietnutá, ak jedna skupina je významne väčšia ako druhá skupina [19].

Výpočet Mannovho-Whitneyho U testu sme programovali v skripte spolu so spracovaním tabuľky S5 v jazyku python. Použili sme knižnicu stats [6] a vo funkcii sme zvolili nastavenie alternative='less', ktoré testuje nulovú hypotézu.

3.5 Korekcia viacnásobného testovania

V niektorých situáciách môžeme vykonať mnoho testov hypotéz. Napríklad Zeman a kol. [28] mali 5821 génov. Ak vykonávame určitý štatistický test pre každý gén, vykonali by sme 5821 samostatných testov hypotéz. Predpokladajme, že každý test sa vykonáva na hladine alfa. Pre jeden test pravdepodobnosť falošného odmietnutia nulovej hypotézy je alfa. Ale šanca aspoň jedného falošného odmietnutia medzi všetkými testami je oveľa vyššia. Toto je problém viacnásobného testovania. Je veľa spôsobov ako riešiť tento problém ale my si povieme iba o dvoch.

Bonferroniho metóda vyžaduje aby pravdepodobnosť falošného odmietnutia hociakej nulovej hypotézy je menšia alebo rovná alfe. Metóda sa snaží, aby bolo nepravdepodobné, že by sme urobili čo i len jedno falošné domietnutie.

Občas je lepšie určiť prah na FDR (false discovery rate), ktorý je definovný ako stredná hodnota počtu falošných zamietnutí vydelený počtom zamietnutí. FDR je teda podiel zamietnutí, ktoré sú nesprávne.

Na vykonanie korekcie viacnásobného testovania sme si napísali skript v jazyku Python a použili knižnicu obsahujúcu tento test - statsmodel [7] a ako metódu sme si zvolili Benjamini a Hochberg, kde keď nám vráti True, znamená to, že platí pre hypotézu, ktorú možno odmietnuť pre danú α.

Kapitola 4

Výsledky

4.1 Štatistické testy

Na testy sme použili GO kategórie, ktoré sa nachádzajú v Uniprot databáze a vybrali sme si tie gény z GO kategórie, ktoré boli v tabuľke S5. Počet génov, ktoré sú v Uniprot databáze pre *C. parapsilosis*, ale nenachádzajú sa v tabuľke TableS5 je 25. Počet génov z tabuľky S5, ktoré nie sú priradené ku GO kategórií a nemohli byť použité je 1504.

4.1.1 Fisherov presný test

Výsledné p-hodnoty vypočítané Fischerovým presným testom pre dáta namerané pri raste kvasinky *C. parapsilosis* na 3–OH médiu sú v tabuľke 4.1 a na 4–OH médiu v tabuľke 4.2. Prvý stĺpec obsahuje identifikátor GO kategórie, druhý, označený ako „X“ obsahuje počet génov danej GO kategórie, ktoré mali p-hodnotu menšiu ako 0.01. Tretí stĺpec označený ako „Y“ obsahuje celkový počet génov GO kategórie, t.j. gény, ktoré boli v prvom stĺpci kontingenčnej tabuľky - všetky gény nachádzajúce sa v GO. Štvrtý stĺpec obsahuje p-hodnoty vypočítané Fischerovým presným testom a piaty stĺpec skrátené názvy GO kategórií. Do tabuľky výsledkov sme dali iba významné GO kategórie - také, ktoré majú p-hodnotu menšiu ako 0.03.

V tabuľke 4.1 vyšla ako najvýznamnejšia GO kategória vyšla oxidoreduktázová aktivita GO:0016491. Enzým oxidoreduktáza katabolizuje intermolekulové oxidačno-redukčné premeny prenosom atómov vodíka alebo elektrónov. Táto kategória patrí do ontológie molekulárna funkcia.

Druhá najvýznamnejšia GO kategória je GO:0016712 s názvom oxidoreduktázová aktivita pôsobiaca na párové donory so začlenením alebo redukciou molekulárneho kyslíka, redukovaného flavínu alebo flavoproteínu ako jedného darcu a začlenením jedného atómu kyslíka. Kategória je tiež zaradená do ontológie molekulárna funkcia. Definícia zo stránky databázy Gene Ontology hovorí, že GO kategória predstavuje katalýzu oxidačno-redukčnej reakcie, pri ktorej sa vodík alebo elektróny prenášajú z reduko-

vaného flavínu alebo flavoproteínu a jedného ďalšieho donora a jeden atóm kyslíka je začlenený do jedného z donorov. Z článku od Holešovej a kol. [15] sa dozvedáme, že gény MNX1 (CPAR2_102790) a MNX2 (CPAR2_704320) kódujú flavoproteín monooxygenázu, ktorá katalyzuje prvý krok v 3–OAP a GP dráhach, presnejšie 4–hydroxybenzoát 1-hydroxylyáza (pre 3–OAP) a aktivita tejto zlúčeniny sa tiež nachádza v našej tabuľke najvýznamnejších GO kategórií - GO:0018678.

Ako tretia najvýznamnejšia GO kategória nám vyšla GO:0019336 s názvom katabolický prenos zlúčeniny obsahujúcej fenol. Podľa definície z Gene Ontology, GO kategória predstavuje chemické reakcie a dráhy, ktoré vedú k rozpadu fenolu, čiže akejkoľvek zlúčeniny obsahujúcej jednu alebo viac hydroxylových skupín priamo pripojených k aromatickému uhlikovému kruhu. 3–hydroxybenzoát a 4–hydroxybenzoát obsahujú aromatické jadro, ku ktorému sú pripojené hydroxylové skupiny. Holešová a kol. [15] zistila, že obidve dráhy môžu prispieť k prežitiu buniek kvasinky v prostredí bohatom na fenolové deriváty.

Zeman a kol. [28] zistili, že bunky prispôsobené hydroxybenzoátom zvýšili expresiu génov SFC1, LEU5, YHM2 a MPC1, ktoré kódujú nosiče sukcináu, fumaráu, koenzýmu A, oxoglutarátu a citrátu a podjednotku pyruvátu. V tabuľke výsledkov 4.1 môžete vidieť, že sa tam nachádza GO kategória GO:0005469, ktorá nesie názov antiportová aktivita sukcinát-fumarátu. Ďalej sa tam nachádza GO:0047617 s názvom acetyl–CoA hydrolázová aktivita. Obidve kategórie sú spojené s proteínmy, ktoré kódujú gény exprimované pri raste buniek na hydroxybenzoátoch.

Ďalej v tomto článku bolo zistené, že niekoľko enzýmov 3–OAP a GP ciest sú spojené s mitochondriou cez proteíny ako nosiče sukcinátu a fumarátu. Dve GO kategórie spojené s membránovým transportom nám vyšli ako významné - GO:0022857 a GO:0055085. Ďalej treba spomenúť GO:0018669 s názvom 3–hydroxybenzoát 6–monooxygenázová aktivita, ktorá katalyzuje reakciu s 3–hydroxybenzoátom.

Medzi zaujímavé výsledky, ktoré treba spomenúť patria aj tieto GO kategórie: GO:0033573 s názvom komplex železa s vysokou afinitou, GO:0005381 transmembránová aktivita transportéra pre ión železa, GO:0005506 viazanie iónu železa, GO:0006783 biosyntetický proces hému. Všetky GO kategórie sú nejakým spôsobom spojené so železom a zatiaľ nevieme príčinu objavu týchto GO vo výsledkoch.

V tabuľke 4.2 nám ako najvýznamnejšia GO kategória vyšla GO:0016712, ktorú sme už popísali vyššie. Druhá najvýznamnejšia GO:0020037 má názov viazanie hému. Ontológiu má priradenú molekulárnu funkciu a podľa jej definície to je selektívna a nekovalentná interakcia s hémom, čo je akákoľvek zlúčenina železa v komplexe s porifyrínovým (tetrapyrolovým) kruhom. Podobné výsledky sú napríklad GO:0005506 s názvom viazanie iónu železa alebo GO:0005381 s názvom transmembránová aktivita transportéru pre ión železa. Tieto výsledky sú veľmi zaujímavé a zatiaľ nevieme prečo sa železo a zlúženiny s ním spojené objavili vo výsledkoch.

Tretia najvýznamnejšia GO kategória - GO:0048037 má názov väzba kofaktoru. Časť definície hovorí, že kofaktor môže byť aj organický a v tom prípade to je koenzým. Taktiež GO:0050662 sa nachádza v tabuľke výzmaných GO kategórií a má názov väzba koenzýmu. Podľa Holešovej a kol. [15] je koenzým ako výsledný produkt génov exprimovaných v bunkách, ktoré rástli na hydroxybenzoát derivátoch. Taktiež sa v tabuľke nachádzajú GO:0047617, GO:0022857 a GO:0018678, ktoré sú už popísané vyššie.

4.1.2 Mannov-Whitneyho U test

Výsledné p-hodnoty vypočítané Mannovým-Whitneyho U testom môžete vidieť pre kvasinku *C. parapsilosis*, ktorá rástla na 3–OH médiu v tabuľke 4.3, kde sme vybrali iba tie GO kategórie, ktoré majú p-hodnotu menšiu ako 0.003 a pri raste na 4–OH médiu v tabuľke 4.4, kde hranica p-hodnôt je tvorená 0.0009. Rôzne hranice sme vybrali z dôvodu veľkosti tabuľky.

Najvýznamnejšia kategória pre rast kvasinky na 3–OH médiu nám vyšla GO:0030684 s názvom preribozóm. Je zaradený v ontológii zložka bunky a predstavuje akýkoľvek komplex pre-rRNA a ribozomálnych proteínov vytvorených počas ribozómovej biogenézy. Druhá najvýznamnejšia kategória je GO:0048037 s názvom väzba kofaktorov, ktorá je popísaná v časti 4.1.1. Tretia významná kategória GO:001649 s názvom aktivita oxidoreduktázy je tiež popísaná v časti 4.1.1, kde nám vyšla ako najvýznamjšia pri raste na 3–OH médiu.

Ďalšie kategórie, ktoré máme v tabuľke 4.3 a súvisia s dátami sú GO:0050662 väzba koenzýmu a GO:0016712, ktoré sú tiež popísané v kapitole 4.1.1.

V tabuľke 4.4, pre rast kvasinky *C. parapsilosis* na 4–OH médiu, nám taktiež ako prvá GO kategória vyšla GO:0030684 s názvom preribozóm, ktorá je popísaná vyššie. Ako druhá najvýznamnejšia GO kategória je GO:0016491 s názvom aktivita oxidoreduktázy, ktorú sme už spomínali. Tretia najvýznamnejšia kategória je GO:0005730 s názvom jadierko. Nachádza sa v ontológii zložka bunky. Definícia ho opisuje ako malé, husté teliesko, ktoré sa nachádza v jadre eukaryotických buniek. Je bohaté na RNA a proteíny.

Spomenieme aj GO:0048037 viazanie kofaktoru a GO:0050662 viazanie koenzýmu, ktoré sú popísané vyššie.

4.2 Korekcia viacnásobného testovania

Po vykonaní korekcie viacnásobného testovania pre výsledky Fischerovho presného testu pre 4–OH nám skript vrátil pre všetky p-hodnoty False, čo znamená že zamietol hypotézu pre všetky p-hodnoty. Ďalej vrátil opravené p-hodnoty, ktoré sa všetky rov-

nali 1.0. Pre výsledky Fischerovho presného testu pre 3–OH nám pre prvé štyri GO kategórie vrátil True a opravil p-hodnoty pre prvých osem kategórií (viď 4.5).

GO kategória	p-hodnota
GO:0016712	2.7308245335685544e-08
GO:0020037	1.5151269779622683e-07
GO:0048037	1.5151269779622683e-07
GO:0005506	0.003982845140434503
GO:0016491	0.012131988845689955
GO:0050662	0.014649576699949415
GO:0050660	0.6409080976407078
GO:0050661	0.6879917064838967

Tabuľka 4.5: Nové p-hodnoty pre 8 GO kategórií, ktoré vrátila korekcia viacnásobného testovania

4.3 Overenie a porovnanie výsledkov Fischerovho presného testu

Stránka *Candida* genome database obsahuje nástroj GO term finder www.candidagenome.org/cgi-bin/GO/goTermFinder, ktorá robí podobné testy ako my v našej práci. Zo stránky nevieme zistiť aké testy a nastavenia použili. Tiež nevieme povedať s akými databázami pracuje.

Vybrali sme si kvasinku *Candida parapsilosis*. Následne sme zadali množinu génov, ktoré majú p-hodnoty menšie ako 0.01 pre rast na 3–OH médie s počtom 134 z tabuľky S5. Vybrali sme ontológiu biologický proces a z pozadia (zbytok génov) sme zašrtli iba ORF a stlačili tlačidlo hľadania. Tieto kroky sme zopakovali ešte dva krát ale zakaždým sme si zvoli inú ontológiu - molekulárna funkcia a zložka bunky.

Výsledky môžete vidieť v tabuľke 4.7. Vložili sme do nej iba významné výsledky, také ktorých p-hodnota bola menšia ako 0.01. Porovnaním výsledkov s Fischerovým presným testom v tabuľke 4.1 môžete vidieť rozdiely v p-hodnotách aj stĺpci X. Môže to byť preto lebo použili iné anotácie génov do GO kategórií.

GO kategória oxidoreduktázová aktivita vyšla v oboch testoch medzi najvýznanejšími kategóriami.

GO term	X	Y	p-value	FDR (%)
Iron ion transport	7 54%	25 0.4 %	0.00036	0.00
Transmembrane transport	26 19.4 %	425 7.3 %	0.00108	0.00
Oxidoreductase activity	23 17.2 %	426 7.3%	0.00958	2.00

Tabuľka 4.6: Významné p-hodnoty vypočítané web stránkou z údajov rastu kvasinky *C. parapsilosis* na 3–OH médie

Výsledky pre gény, ktoré mali p-hodnotu menšiu ako 0.01 pri raste na 4–OH médiu sa nachádzajú v tabuľke 4.7. Počet génov v tabuľke S5 s p-hodnotou menšou ako 0.01, ktoré sme použili v teste je 152

Ako najvýznamnejšia GO kategória nám vyšla oxidoreduktázová aktivita pôsobiaca na párové donory so začlenením alebo redukciou molekulárneho kyslíka, redukovaného flavínu alebo flavoproteínu ako jedného darcu a začlenením jedného atómu kyslíka rovnako ako vo Fischerovom presnom teste pre 4–OH 4.2. Ako druhá väzba železa, ktorá sa tiež zhoduje miestom vo Fischerovom testom. Ďalšie spoločné GO kategórie sú oxidoreduktázová aktivita, väzba kofaktoru a väzba iónu železa.

GO term	**X**	**Y**	**p-value**	**FDR (%)**
Oxidoreductase activity, acting on paire donors	8 5.3 %	10 0.2 %	8.64e-10	0.00
Heme binding	13 8.6 %	41 0.7 %	1.62e-09	0.00
Tetrapyrrole binding	13 8.6 %	41 0.7 %	1.62e-09	0.00
Oxidoreductase activity	35 23.0 %	426 7.3 %	5.37e-08	0.00
Cofactor binding	26 17.1 %	251 4.3 %	1.03e-07	0.00
Monooxygenase activity	9 5.9 %	25 0.4 %	1.06e-06	0.00
Oxidoreductase activity, acting on paire donors	9 5.9 %	50 0.9 %	0.00049	0.00
Oxidation-reduction process	27 17.8 %	401 6.9 %	0.00114	0.00
Iron Ion binding	8 5.3 %	52 0.9 %	0.0057	0.00

Tabuľka 4.7: Významné p-hodnoty vypočítané web stránkou z údajov rastu kvasinky *C. parapsilosis* na 4–OH médie

GO	X	Y	p-value	Name
GO:0016491	16	290	0.000932	oxidoreductase activity
GO:0016712	3	10	0.001272	oxidoreductase activity, acting on paired donors, ...
GO:0019336	2	3	0.001554	phenol-containing compound catabolic process
GO:0016021	40	1145	0.002849	integral component of membrane
GO:0042183	2	4	0.003062	formate catabolic process
GO:0033573	2	4	0.003062	high-affinity iron permease complex
GO:0005381	2	4	0.003062	iron ion transmembrane transporter activity
GO:0051287	4	30	0.004624	NAD binding
GO:0071949	3	20	0.010209	FAD binding
GO:0000103	2	7	0.010240	sulfate assimilation
GO:0016709	2	7	0.010240	oxidoreductase activity, acting on paired donors, ...
GO:0047617	2	7	0.010240	acyl-CoA hydrolase activity
GO:0005506	4	38	0.010812	iron ion binding
GO:0022857	9	162	0.012016	transmembrane transporter activity
GO:0055085	9	163	0.012476	transmembrane transport
GO:0004190	3	24	0.016954	aspartic-type endopeptidase activity
GO:0016575	2	9	0.017032	histone deacetylation
GO:0005887	2	10	0.020972	integral component of plasma membrane
GO:0006783	2	10	0.020972	heme biosynthetic process
GO:0018669	1	1	0.023020	3-hydroxybenzoate 6-monooxygenase activity
GO:0004020	1	1	0.023020	adenylylsulfate kinase activity
GO:0052837	1	1	0.023020	thiazole biosynthetic process
GO:0032406	1	1	0.023020	MutLbeta complex binding
GO:0004851	1	1	0.023020	uroporphyrin-III C-methyltransferase activity
GO:0051038	1	1	0.023020	negative regulation of transcription ...
GO:0000117	1	1	0.023020	regulation of transcription involved in G2/M ...
GO:0002134	1	1	0.023020	UTP binding
GO:0004418	1	1	0.023020	hydroxymethylbilane synthase activity
GO:0018678	1	1	0.023020	4-hydroxybenzoate 1-hydroxylase activity
GO:0004089	1	1	0.023020	carbonate dehydratase activity
GO:0007129	1	1	0.023020	synapsis
GO:0004594	1	1	0.023020	pantothenate kinase activity
GO:1990166	1	1	0.023020	protein localization to site of double-strand break
GO:0005469	1	1	0.023020	succinate:fumarate antiporter activity
GO:0002135	1	1	0.023020	CTP binding
GO:0035601	2	11	0.025251	protein deacylation
GO:0046618	2	12	0.029850	drug export

Tabuľka 4.1: VVýznamné p-hodnoty vypočítané Fischerovým presným testom z údajov rastu kvasinky *C. parapsilosis* na 3–OH médie

GO	X	Y	p-value	Name
GO:0016712	8	10	0.000000000008	oxidoreductase activity acting on paired donors...
GO:0020037	11	30	0.000000000096	heme binding
GO:0048037	24	195	0.000000000129	cofactor binding
GO:0005506	8	38	0.000004522106	iron ion binding
GO:0016491	21	290	0.000017218264	oxidoreductase activity
GO:0050662	13	128	0.000024949606	coenzyme binding
GO:0050660	6	47	0.001273447824	flavin adenine dinucleotide binding
GO:0050661	4	20	0.001562286021	NADP binding
GO:0005381	2	4	0.003926542547	iron ion transmembrane transporter activity
GO:0033573	2	4	0.003926542547	high-affinity iron permease complex
GO:0042183	2	4	0.003926542547	formate catabolic process
GO:0051287	4	30	0.007217858827	NAD binding
GO:0016021	42	1145	0.010163593707	integral component of membrane
GO:0047617	2	7	0.013050732155	acyl-CoA hydrolase activity
GO:0072593	2	9	0.021618975399	reactive oxygen species metabolic process
GO:0022857	9	162	0.025340398772	transmembrane transporter activity
GO:0051908	1	1	0.026112351830	double-stranded DNA 5'-3'exodeoxyribonuclease
GO:0000709	1	1	0.026112351830	meiotic joint molecule formation
GO:0004594	1	1	0.026112351830	pantothenate kinase activity
GO:0097344	1	1	0.026112351830	Rix1 complex
GO:0004089	1	1	0.026112351830	carbonate dehydratase activity
GO:0018678	1	1	0.026112351830	4-hydroxybenzoate 1-hydroxylase activity
GO:0005887	2	10	0.026566089383	integral component of plasma membrane

Tabuľka 4.2: Významné p-hodnoty vypočítané Fischerovým presným testom z údajov rastu kvasinky *C. parapsilosis* na 4–OH médie

GO	p-value	Name
GO:0030684	0.00000000029	preribosome
GO:0048037	0.00000017582	cofactor binding
GO:0016491	0.00000182463	oxidoreductase activity
GO:0006364	0.00000186734	rRNA processing
GO:0006412	0.00000295479	translation
GO:0030687	0.00000351495	preribosome, large subunit precursor
GO:0006096	0.00000731330	glycolytic process
GO:0050662	0.00000977108	coenzyme binding
GO:1990904	0.00008986945	ribonucleoprotein complex
GO:0032040	0.00013645467	small-subunit processome
GO:0019843	0.00014723066	rRNA binding
GO:0016712	0.00015909622	oxidoreductase activity, acting on paired donors, ...
GO:0005730	0.00023371481	nucleolus
GO:0005975	0.00027544225	carbohydrate metabolic process
GO:0015934	0.00032365998	large ribosomal subunit
GO:0004190	0.00058596249	aspartic-type endopeptidase activity
GO:0042183	0.00065053832	formate catabolic process
GO:0006267	0.00079189687	pre-replicative complex assembly ...
GO:0051287	0.00097601030	NAD binding
GO:0005762	0.00113298943	mitochondrial large ribosomal subunit
GO:0020037	0.00117507883	heme binding
GO:0009228	0.00127238259	thiamine biosynthetic process
GO:0000470	0.00142481026	maturation of LSU-rRNA
GO:0010181	0.00150863203	FMN binding
GO:0000472	0.00175826815	endonucleolytic cleavage to generate mature 5'-end ...
GO:0000103	0.00178417041	sulfate assimilation
GO:0050660	0.00198692953	flavin adenine dinucleotide binding
GO:0005887	0.00235762044	integral component of plasma membrane
GO:0000027	0.00238410935	ribosomal large subunit assembly
GO:0070814	0.00246012198	hydrogen sulfide biosynthetic process
GO:0000447	0.00262544279	endonucleolytic cleavage in ITS1 ...
GO:0016614	0.00262627831	oxidoreductase activity, acting on CH-OH group of donors
GO:0005656	0.00269183968	nuclear pre-replicative complex
GO:0005840	0.00295416636	ribosome
GO:0015238	0.00298447399	drug transmembrane transporter activity

Tabuľka 4.3: Významné p-hodnoty vypočítané Mannovým-Whitneyho U testom z údajov rastu kvasinky *C. parapsilosis* na 3–OH médie

GO	p-value	Name
GO:0030684	0.0000000000000002	preribosome
GO:0016491	0.0000000000013577	oxidoreductase activity
GO:0005730	0.0000000008347371	nucleolus
GO:0006364	0.0000000016777715	rRNA processing
GO:0048037	0.0000000026963696	cofactor binding
GO:0030687	0.0000000109557119	preribosome, large subunit precursor
GO:0050662	0.0000000249909616	coenzyme binding
GO:0000469	0.0000013456646554	cleavage involved in rRNA processing
GO:0050660	0.0000015733091273	flavin adenine dinucleotide binding
GO:0000447	0.0000017747118327	endonucleolytic cleavage in ITS1 to separate SSU-rRNA ...
GO:0003824	0.0000021135899420	catalytic activity
GO:0016712	0.0000021241551688	oxidoreductase activity, acting on paired donors, ...
GO:0032040	0.0000031350520120	small-subunit processome
GO:0000470	0.0000045465792839	maturation of LSU-rRNA
GO:0000472	0.0000128610367586	endonucleolytic cleavage to generate mature 5'-end ...
GO:0000480	0.0000188473448840	endonucleolytic cleavage in 5'-ETS of tricistronic rRNA ...
GO:0006267	0.0000243999656832	pre-replicative complex assembly ...
GO:0000463	0.0000254145453758	maturation of LSU-rRNA from tricistronic rRNA transcript ...
GO:0016614	0.0000612588847336	oxidoreductase activity, acting on CH-OH group of donors
GO:0006096	0.0000725892544166	glycolytic process
GO:0005656	0.0000994373611316	nuclear pre-replicative complex
GO:0030686	0.0001254881032725	90S preribosome
GO:0020037	0.0001781977232477	heme binding
GO:0042555	0.0002063418514877	MCM complex
GO:0005506	0.0002483050580800	iron ion binding
GO:0042274	0.0004046792129712	ribosomal small subunit biogenesis
GO:0005654	0.0004202888129498	nucleoplasm
GO:0042183	0.0005255947327935	formate catabolic process
GO:0006772	0.0005635807695929	thiamine metabolic process
GO:0017116	0.0007898532222989	single-stranded DNA-dependent ATP-dependent DNA helicase
GO:0016616	0.0008828922085403	oxidoreductase activity acting on the CH-OH ...

Tabuľka 4.4: Významné p-hodnoty vypočítané Mannovým-Whitneyho U testom z údajov rastu kvasinky *C. parapsilosis* na 4–OH médie

Záver

Cieľom práce bolo analyzovať transkriptomické dáta z patogénnej kvasinky *Candida parapsilosis* na raste na 3–hydroxybenzéne a 4–hydroxybenzéne, čo sa nám úspešne podarilo.

V prvej kapitole sme vysvetlili základné pojmy a vysvetlili dáta z článku [28], z ktorého sme vychádzali a na ktorý sme nadvezovali. V ďalšej kapitole sme popísali zdroje dát a ich formáty. Na záver kapitoly sme opísali a vysvetlili dáta z tabuľky S5, na ktorých sme v tretej kapitole robili štatistické testy. Avšak aby sme mohli vykonávať tieto testy, najskôr sme si museli upraviť formu dát z databáz. Po vykonaní testov, sme popísali vysledky v štvrtej kapitole.

Na overenie správnosti výsledkov sme použili online stránku, ktorá robí rovnakú prácu a to pre GO kategórie počíta p-hodnoty. Výsledky nám vyšli približne rovnaké. Odchýlky môžu byť spôsobené tým, že stránka používa iné databázy alebo inú metódu počítania.

Zistili sme, ktoré GO kategórie sú obohatené na diferencovane exprimované gény. Pri raste na 3–OH a 4–OH médiach sa vytvárajú génové produkty z génov zapnutých keď kvasinka rastie na týchto médiach. Zistili sme, že významné GO kategórie súvisia s týmito génovými produktami.

Vo výsledkoch sa nám vyskytli aj GO kategórie, ktoré súvisia so železom. Nevieme povedať ako to môže súvisieť s rastom kvasinky *C. parapsilosis* na 3–OH a 4–OH médiach ale táto informácia sa dá využiť pri pokračovaní na práci, t.j. preskúmať ako súvisí rast kvasinky na 3–OH a 4–OH médiach so železom.

Na práci sa dá pokračovať ďalej aj v skúmaní proteomiky. Dajú sa zanalyzovať dáta z merania proteómov a porovnať s dátami z merania transkriptov.

Na záver ešte spomenieme že existuje aj iný nástroj na prácu podobnej našej - g:Profiler. Je to online nástroj na webe. Avšak my sme ho použiť nemohli lebo je limitovaný druhmi organizmov a naša kvasinka sa tam nevyskytuje. Na grafické zobrazenie výsledkov g:Profiler používa program Cytoscape, ktorý sme sa snažili použiť aj my. Vytvorili sme typy súborov podobné ako vytvára g:Profiler, ale zrejme sme prehliadli nejakú drobnosť a Cytoscape pre naše súbory nefungoval.

Po malých modifikáciach sa naše skripty dajú použiť aj na iné organizmy, kým g:Profiler a Candida genome database majú vopred dané organizmy.

Literatúra

[1] 3-hydroxybenzoate. https://pubchem.ncbi.nlm.nih.gov/compound/3-Hydroxybenzoate.

[2] 4-hydroxybenzoate. https://pubchem.ncbi.nlm.nih.gov/compound/4-hydroxybenzoate.

[3] The elementtree xml api. https://docs.python.org/3/library/xml.etree.elementtree.html.

[4] Fischer exact test dokumentácia. https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.fisher_exact.html.

[5] The gene ontology resource. http://geneontology.org/docs/download-ontology.

[6] Mann whitney u test dokumentácia. https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.mannwhitneyu.html.

[7] Multiple test dokumentácia. https://www.statsmodels.org/stable/generated/statsmodels.stats.multitest.multipletests.html.

[8] pandas.dataframe. https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.html.

[9] Translation: DNA to mRNA to Protein. https://www.nature.com/scitable/topicpage/translation-dna-to-mrna-to-protein-393.

[10] UniParc at the EBI. https://www.ebi.ac.uk/uniparc.

[11] xlrd documentation. https://xlrd.readthedocs.io/en/latest.

[12] Uniprot: a hub for protein information. *Nucleic Acids Research*, 43, 2014.

[13] Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, et al. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25, 2000.

[14] Gene Ontology Consortium. The Gene Ontology (GO) database and informatics resource. *Nucleic acids research*, 32(suppl_1):D258–D261, 2004.

[15] Zuzana Holesova, Michaela Jakubkova, Ivana Zavadiakova, Igor Zeman, Lubomir Tomaska, and Jozef Nosek. Gentisate and 3-oxoadipate pathways in the yeast Candida parapsilosis: identification and functional analysis of the genes coding for 3-hydroxybenzoate 6-hydroxylase and 4-hydroxybenzoate 1-hydroxylase. *Microbiology*, 157(7):2152–2163, 2011.

[16] Rasko Leinonen, Federico Garcia Diezand David Binns, Wolfgang Fleischmann, Rodrigo Lopez, and Rolf Apweiler. UniProt archive. *Bioinformatics*, 20, 2004.

[17] Matthew D Lieberman and William A Cunningham. Type I and Type II error concerns in fMRI research: re-balancing the scale. *Social cognitive and affective neuroscience*, 4(4):423–428, 2009.

[18] Patrick E McKnight and Julius Najab. Mann-Whitney U Test. *The Corsini encyclopedia of psychology*, pages 1–1, 2010.

[19] Nadim Nachar et al. The Mann-Whitney U: A test for assessing whether two independent samples come from the same distribution. *Tutorials in quantitative Methods for Psychology*, 4(1):13–20, 2008.

[20] Isabelle Rivals, Léon Personnaz, Lieng Taing, and Marie-Claude Potier. Enrichment or depletion of a GO category within a class of genes: which test? *Bioinformatics*, 23(4):401–407, 2006.

[21] D. Peter Snustad and Michael J. Simmons. *Genetika*. Masarykova univerzita, 2017.

[22] MUDr. Júlia Stracenská. Sledovanie antimikrobiálnej aktivity materiálov vlhkého hojenia pri ošetrovaní chronických rán. 2013.

[23] Baris E. Suzek, Yuqi Wang, Hongzhan Huang, Peter B. McGarvey, and Cathy H. Wu. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*, 31, 2014.

[24] The Gene Ontology Consortium. The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Research*, 47(D1):D330–D338, 11 2018.

[25] Cole Trapnell, Lior Pachter, and Steven L Salzberg. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, 25(9):1105–1111, 2009.

[26] David Trofa, Attila Gácser, and Joshua D. Nosanchuk. Candida parapsilosis, an emerging fungal pathogen. *Clin Microbiol Rev*, 21(4)(PMC2570155), 2008.

[27] Zhong Wang, Mark Gerstein, and Michael Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews genetics*, 10(1):57, 2009.

[28] Igor Zeman, Martina Neboháčová, Gabriela Gérecová, Kornélia Katonová, Eva Jánošíková, Michaela Jakúbková, Ivana Centárová, Ivana Dunčková, Ľubomír Tomáška, Leszek P. Pryszcz, Toni Gabaldón, and Jozef Nosek. Mitochondrial Carriers Link the Catabolism of Hydroxyaromatic Compounds to the Central Metabolism in Candida parapsilosis. *G3: GENES, GENOMES, GENETICS*, 6(12 4047-4058), 2016.