Jihočeská univerzita v Českých Budějovicích

Přírodovědecká fakulta

Ústav aplikované informatiky

Instalace a úprava bioinformatického serveru Galaxy

Bakalářská práce

Jaroslav Steinhaisl

Vedoucí bakalářské práce: Mgr. Jiří Pech, Ph. D.

Externí školitel: Ing. Petr Novák, Ph. D.

České Budějovice 2012

Bibliografické údaje

Steinhaisl, J., 2012: Instalace a úprava bioinformatického serveru Galaxy. [Installation and Administration bioinformatic Galaxy server. Bc. Thesis, in Czech.] – 35 p., Faculty of Science, University of South Bohemia, České Budějovice, Czech Republic.

Anotace

Práce se zabývá instalací a administrací Galaxy serveru s následnou konfigurací a implementací programů vyvinutých na Ústavu molekulární biologie rostlin Biologického centra Akademie věd České republiky, v. v. i.. Konfigurace obsahuje upravení výchozího nastavení serveru pro správnou spolupráci Galaxy serveru s ProFTPD serverem, PostgreSQL databázového serveru a Portable Batch System. Každý bioinformatický program byl zaveden pomocí konfigurace XML souboru, který je použit pro definování vstupů a výstupů programu. Kombinace XML souboru s bioinformatickým programem byla následně testována na datech poskytnutých Laboratoří Molekulární Cytogenetiky.

Abstract

This work describes the installation and administration of Galaxy server and following configuration and implementation of bioinformatic programs developed by Institute of Plants Molecular Biology of Biology Centre of the Academy of Sciences of the Czech Republic, v. v. i.. Configuration required modification of default server settings to ensure correct co-operation of Galaxy server and ProFTPD server, PostgreSQL database server and the Portable Batch System. Each bioinformatic program was implemented through XML tool configuration file which is used to define user web interface, types of the input and output data and additional options. Each combination of XML configuration file and bioinformatic program was subsequently tested on data files provided by Laboratory of Molecular Cytogenetics.

Prohlášení

Prohlašuji, že svoji bakalářskou práci jsem vypracoval samostatně pouze s použitím pramenů a literatury uvedených v seznamu citované literatury.

Prohlašuji, že v souladu s § 47b zákona č. 111/1998 Sb. v platném znění souhlasím se zveřejněním své bakalářské práce, a to v nezkrácené podobě elektronickou cestou ve veřejně přístupné části databáze STAG provozované Jihočeskou univerzitou v Českých Budějovicích na jejích internetových stránkách, a to se zachováním mého autorského práva k odevzdanému textu této kvalifikační práce. Souhlasím dále s tím, aby toutéž elektronickou cestou byly v souladu s uvedeným ustanovením zákona č. 111/1998 Sb. zveřejněny posudky školitele a oponentů práce i záznam o průběhu a výsledku obhajoby kvalifikační práce. Rovněž souhlasím s porovnáním textu mé kvalifikační práce s databází kvalifikačních prací Theses.cz provozovanou Národním registrem vysokoškolských kvalifikačních prací a systémem na odhalování plagiátů.

V Dobré Vodě u Českých Budějovic, 20. dubna 2012

.....

Jaroslav Steinhaisl

Poděkování

Rád bych poděkoval Mgr. Jiřímu Pechovi, Ph. D. za rady a vedení mé bakalářské práce. Dále Ing. Petru Novákovi, Ph. D. za objasnění bioinformatických věcí se kterými jsem se setkal a také za podněty k řešení problému během bakalářské práce a objasnění práce jež vykonávají programy vytvořené na ÚMBR, které byly zavedeny do Galaxy.

Obsah

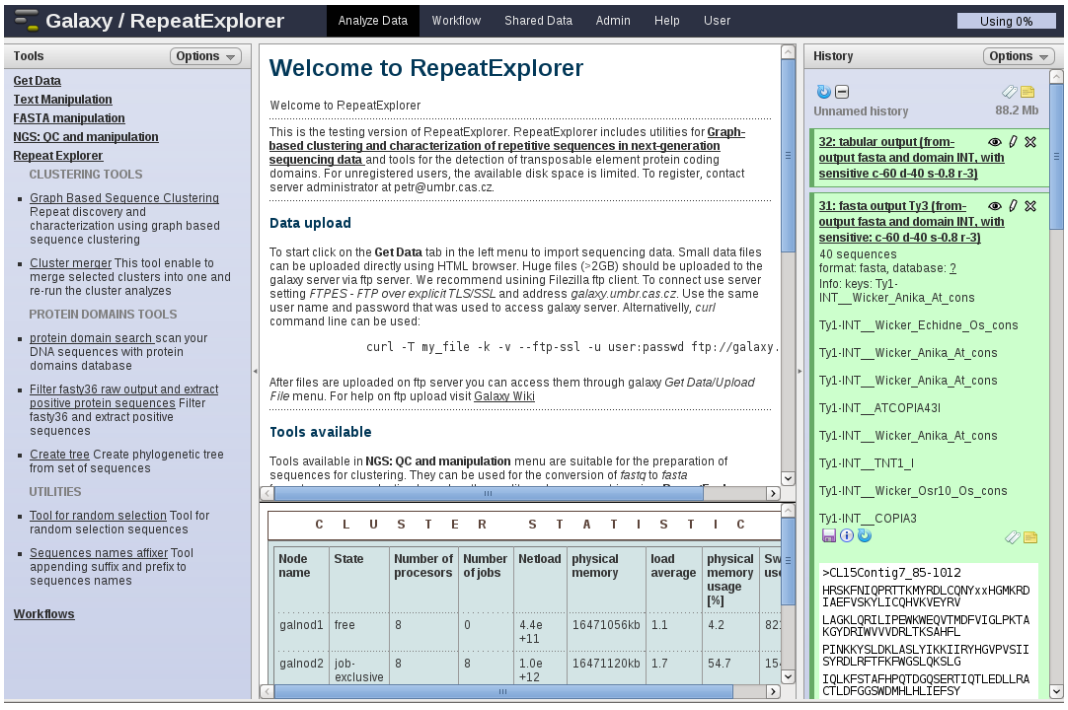
1 Úvod a cíle práce

1.1 Úvodem

Rozvoj biologie s sebou přináší čím dál větší zapojování informatických nástrojů pro různé biologické výpočty. To má za následek rozvoj nových projektů jež jsou vytvářeny s myšlenkou vytvoření programů umožňujících výpočty potřebné pro získání různých biologických informací a zprostředkování těchto programů uživatelům,jimž jsou tyto programy nápomocny. Jedním z takovýchto projektů je i open source projekt Galaxy.

Hlavní náplní práce je zprovoznění Galaxy, což je otevřená on-line platforma pro dostupnost, reprodukovatelnost a transparentnost výpočetního biomedicínského výzkumu[1]. Galaxy je pojat jako webový server, který obsahuje programy pro různé biomedicínské výpočty a také je možné do něj zavést vlastní výpočetní programy. Galaxy server vychází z toho, že jeho funkce jsou prezentovány pomocí webového rozhraní. Toto rozhraní je rozděleno na tři části a lištu, jak je vidět na obrázku 1. Levý sloupec „tools“ obsahuje seznam všech dostupných programů, které mohou být použity uživateli pro výpočty. Střední oblast slouží pro zobrazení stránek programů, jež slouží jako hlavní rozhraní předávající parametry od uživatele programu. Pravý sloupec rozhraní „history“ obsahuje seznam všech vstupů a výstupů již vykonaných programů. Každý výstup je reprezentován jedním blokem, který obsahuje název výstupního souboru, možnost změnit atributy, stručný výpis programu, možnost přidání popisku, uložení výstupního souboru, prohlédnutí výstupního souboru. Je-li obsah výstupního souboru podporovaného typu, pak je také možný rychlý náhled výstupního souboru zobrazující prvních pět řádků. Lišta obsahuje odkazy „User“ – kde je možnost správy uživatelova profilu, „Admin“ – pro správu administrátorských nástrojů, „Help“ – poskytující nápovědu o použití Galaxy a nástrojů obsažených. Odkaz „Shared data“ – slouží pro přehled sdílených

stránek, workflows, historií, „Workflow“ – umožňuje vytvářet vlastní workflow, grafické nastavení propojení jednotlivých programů pro dosažení automatizovaného výpočtu. Každý program má vstupy a výstupy, pomocí workflow lze programy pospojovat a definovat, co má být na vstupu, jaký výstup má být přiveden na jaký vstup. Uživateli takto definovaný proces ušetří práci, kterou by představovalo čekání na výstup jednoho programu, jež je vstupem pro program druhý a podobně. Odkaz „Analyze data“ odkazuje na hlavní rozhraní webové stránky, tedy zobrazení sloupce nástrojů, stránky a sloupce s historií. Pomocí webového rozhraní Galaxy serveru jsou prezentovány biomedicínské programy, což přináší možnost použití těchto programů pro uživatele, kteří nemají zkušenosti s programováním, které jsou vyžadovány při použití samostatných programů, jež se ovládají převážně pomocí příkazové řádky. Další



Obrázek 1: Úvodní internetová stránka Galaxy s přihlášeným uživatelem.

předností Galaxy je možnost reprodukovatelnosti výsledků. Reprodukovatelnost je zajištěna ukládáním výsledků v nezměněném formátu, a tedy se tento

výsledek dá použít znovu nebo pro jiný výpočet. Další napomáhající věcí jsou metadata, která jsou obsažena u každého výstupu a nesou stručné informace o datovém souboru, použitých nástrojích a parametrech, jež byly zadány pro dosažení takového výsledku. Pomocí těchto dat lze v Galaxy spustit opakované vykonání programu, které bude mít ty stejné parametry, případně chce-li uživatel zjistit, jaký bude výstupní soubor po změně jediného parametru, je mu umožněno předvyplněné parametry změnit[2].

Galaxy může být rozšířen nejen přidáním vlastních programů, ale také využitím služeb typu FTP[1], PostgreSQL[2], Portable Batch System[3] pro ještě větší zefektivnění a usnadnění uživatelům využívat služeb Galaxy.

1.2 Cíle práce

Získání a instalace Galaxy, následné seznámení se s fungováním a konfigurací Galaxy.

Konfigurace Galaxy pro využívání služby Portable Batch System.

Instalace PostgreSQL a konfigurace, která obsahuje vytvoření nového uživatele spolu s novou databází, jež bude následně využívána Galaxy.

Instalace ProFTPD[4][3] spojená s konfigurací, která splňuje požadavky pro bezpečnost šifrované komunikace mezi serverem a klientem.

Úprava a vytvoření programů tak, aby bylo možno zavést bioinformatické nástroje vyvinuté na Ústavu molekulární biologie rostlin do již zprovozněné Galaxy.

Vytvoření dokumentace k serveru.

[1] protokol pro přenos souborů mezi počítači s využitím počítačové sítě

[2] objektově-relační databázový systém

[3] software využívaný na clusteru pro plánování úloh, tedy rozdělení výpočetních úloh na výpočetní zdroje

[4] vysoce konfigurovatelný software pro FTP server s GPL licencí

2 Metodika řešení

Operačním systémem byl zvolen Debian GNU/Linux na základě poznatků, že je pro tento systém mnoho dostupných programů sloužících k bioinformatickým výpočtům, s čímž také souvisí jeho hojné nasazení na Ústavu molekulární biologie rostlin (dále ÚMBR). Pro získání Galaxy bylo použito verzovacího systému Mercurial, pomocí něhož je Galaxy server distribuován, čímž je zajištěna aktuálnost získaných souborů oproti stažení souborů v archívu z internetových stránek Galaxy. Mercurial je rychlý snadno použitelný verzovací nástroj pro softwarové vývojáře[4], který umožňuje udržovat stále aktuální verze souborů, jež jsou pomocí jeho služeb distribuovány. Další funkcí Mercurial je možnost vidět starší verze a změny provedené mezi verzemi. Shodou okolností je na ÚMBR používán verzovací systém Mercurial, takže jeho použití nepředstavovalo žádný problém. Databázový server zajišťuje PostgreSQL. Galaxy podporuje jak PostgreSQL, tak MySQL. Volba PostgreSQL byla provedena na základě tvrzení tvůrců Galaxy, že jejich abstraktní databázová vrstva zajišťována programem SQLAlchemy lépe pracuje s databází PostgreSQL[10]. Funkci FTP serveru obstarává ProFTPD. Galaxy přímo nevyžaduje žádný konkrétní server, ale tvůrci Galaxy využívají služeb ProFTPD, které poskytují veškeré potřebné služby pro správnou komunikaci mezi ProFTPD, PostgreSQL a Galaxy. Po zjištění těchto věcí a malé zkušenosti s FTP servery, bylo rozhodnuto o využití služeb ProFTPD, jež jsou využívány tvůrci Galaxy, což naznačuje funkčnost společně se stabilitou toho spojení.

3 Instalace a konfigurace

3.1 Instalace a seznámení se s Galaxy

Instalace bude provedena na Linuxovém operačním systému Debian GNU/-Linux, jehož verze je v době instalace 6.0.3, který je instalován na cluster zařízení, to bude použito pro účely Galaxy. Dále bylo třeba vytvoření uživatelského účtu na daném systému, který nemá oprávnění správce, aby byla zajištěna bezpečnost. Je vhodné pojmenovat uživatelský účet jménem Galaxy, snadněji se pak určuje, jaké procesy jsou vykonávány pod daným účtem, či zda nedošlo k náhodnému napadení Galaxy serveru a tím spuštění nechtěného programu.

Získání Galaxy bylo provedeno pomocí repozitáře Mercurial, příkazem jež zadáme do terminálu či konzole (záleží na jazykové verzi distribuce)
`hg clone https://bitbucket.org/galaxy/galaxy-dist/`. Ten provede stažení nejnovější verze Galaxy serveru a uloží všechny soubory Galaxy do adresáře „galaxy-dist“, jež se vytvoří v domovském adresáři. Instalace je provedena prvotním spuštěním souboru `run.sh`.

Struktura adresáře `galaxy-dist` je následující.

adresáře		
contrib	cron	database
display_applications	external_service_types	lib
locale	scripts	static
templates	text	test-data
tool-data	tools	

soubory		
create_db.sh	extract_dataset_parts.sh	manage_db.sh
run_functional_test.sh	run.sh	run_unit_tests.sh
tool_conf.xml	universe_wsgi.ini	

`database` – obsahuje složky:

`files` – ve výchozím nastavení určena pro ukládání výsledných souborů programů spuštěných pod Galaxy. Označení souborů a složek je zajištěno Galaxy, kdy název obsahuje číslo, které bylo přiděleno jako pořadové číslo.

`job_working_directory` – zde jsou ve výchozím nastavení vykonávány programy spuštěné pod Galaxy, složky zde vytvářené jsou označeny číslem procesu a mají charakter dočasných složek, které jsou po vykonání programu vymazány.

`pbs` – složka bude použita pro zápis standardních výstupů a souboru obsahujícího chyby při použití clusteru.

`tmp` – slouží pro dočasné ukládání souborů.

`tool-data` – zde se nachází soubory, které jsou používány programy Galaxy, konkrétně při zadávání definovaných parametrů.

`tools` – obsahuje složky, ve kterých se nachází jednotlivé programy, jež můžeme pomocí Galaxy použít.

Další složky v domovském adresáři Galaxy.

`static` – obsahuje soubory spojené s fungováním html stránek Galaxy serveru.

`scripts` – zde jsou skripty, které kontrolují podprogramy využívané Galaxy a také skripty pro mazání souborů vytvořených uživateli.

`test, test-data` – složky obsahující testovací soubory pro podčásti Galaxy a soubory použité při tvorbě testovacích částí programů, tedy při definici testovací části v XML souboru daného programu.

Soubory nacházející se v domovském adresáři Galaxy.

`datatypes_conf.xml` – obsahuje XML[5] strukturu, ve které jsou definovány jednotlivé datové typy. Tato struktura je využita pro rozlišování datových typů pomocí Galaxy, tím dokáže Galaxy rozeznat soubory a uchovávat jejich správnou strukturu.

[5] rozšiřitelný značkovací jazyk

`tool_conf.xml` – pomocí XML definice jsou určeny programy, jež jsou po spuštění Galaxy dostupné a viditelné. Přidání a odstranění programu je řízeno pomocí párového tagu `<section>` více viz. Implementování vlastních programů na straně (16).
`universe_wsgi.ini` – slouží jako hlavní konfigurační soubor pro Galaxy. Zde se nalézají sekce pro nastavení HTTP serveru, filtrů, databáze, složek, správa Galaxy, nastavení pro PBS.
`run.sh` – ústřední program, který zajišťuje spuštění Galaxy. Obsahuje tyto parametry:

`--daemon` – slouží pro spuštění Galaxy pomocí démona na pozadí systému.

`--stop-daemon` – tento parametr slouží k zastavení již běžící instance Galaxy, která byla spuštěna pomocí démona.

`--reload` – slouží pro spuštění pomocí konzole, kdy jsou standardní výstupy psány do konzole a běh Galaxy lze přerušit kombinací kláves CTRL+C.

3.2 Konfigurace Galaxy pro využití PBS

K zajištění správné kooperace mezi Galaxy a clusterem je třeba použít DRMAA egg ('Distributed Resource Management Application API' soubor specifikací API pro řízení programového přístupu ke clusteru a cloud systému[5]), ten je již součástí Galaxy. Je tedy nutné říci, kde se nachází zdrojový manažer DRMAA knihovny, toho je dosaženo přidáním `$DRMAA_LIBRARY_PATH` do proměnné prostředí. Na clusteru je cesta ke knihovně `/usr/lib/libdrmaa.so.1.0`, stačí již tedy přidat proměnnou. Přidání provedeme zápisem příkazu `export DRMAA_LIBRARY_PATH=/usr/lib/libdrmaa.so.1.0` do souboru `bashrc`, který se nachází v domovském adresáři uživatele, pod nímž spouštíme Galaxy.

Jelikož cluster použitý pro Galaxy využívá služeb TORQUE a MAUI (TORQUE je open source správce prostředků poskytující kontrolu nad dávkovými úlohami a distribuovanými výpočetními uzly[6], MAUI je plánovač prostředků clusteru[7].) podléhají následující nastavení tomuto PBS. Další částí nastavení je pbs_python egg, který je potřebný pro komunikaci s TORQUE. Tohoto nastavení docílíme spuštěním příkazu
`LIBTORQUE_DIR=/usr/lib/libtorque.so.2.0.0 python scripts/scramble.py -e pbs_python` z adresáře galaxy-dist.

Nyní je třeba provést patřičné úpravy v konfiguračním souboru Galaxy. V sekci Job Execution souboru `universe_wsgi.ini` nastavíme následující parametry. Pro určení spouštěče úloh na clusteru `start_job_runners = pbs`, na určení fronty PBS serveru, do které bude Galaxy posílat úlohy slouží parametr `default_cluster_job_runner = pbs:///`. Těmito parametry jsme zajistili, že Galaxy bude pro vykonávání úloh používat cluster. Další možností nastavení je povolení obnovení úlohy, pokud je Galaxy restartována a na clusteru byly v té době spuštěny úlohy, jsou tyto úlohy vykonány a po zpětném spuštění Galaxy uloženy, dosáhneme toho parametrem `enable_job_recovery = True`. Pro omezení počtu vykonávaných úloh uživatelem na clusteru slouží `user_job_limit = 3`, kde číslo udává počet povolených úloh pro uživatele. Parametr `job_walltime = 600:00:00` určuje po kolika hodinách má být úloha běžící na clusteru ukončena, pokud nebyla do té doby úspěšně dokončena. Tento parametr je pouze pro nastavení Galaxy, kdy Galaxy kontroluje, zda každá úloha běžící na clusteru nepřekročila tuto dobu. Chceme-li využít `walltime` k nastavení maximálního času, ve kterém může úloha vykonávat výpočet na clusteru, je potřeba toto definovat v parametru `default_cluster_job_runner` přidáním parametru `-l walltime = 600:00:00`, výsledek tedy bude `default_cluster_job_runner = pbs:////-l walltime = 600:00:00/`. Pokud chceme nastavit další omezení na PBS, je třeba jej přidat jako další pa-

rametry do `default_cluster_job_runner`. Galaxy v našem případě obsahuje nastavení, které využívá nastavení maximálního času běhu úlohy na clusteru a omezení využití maximálně 15GB operační paměti jednou úlohou. V takovémto případě vypadá konfigurační parametr takto
`default_cluster_job_runner = pbs:////-l walltime = 600:00:00, mem = 15G/`. Změna složky, v níž jsou vytvářeny dočasné soubory se provede parametrem `job_working_directory = /absolutni/cesta/k_adresari`, kde se zadá absolutní cesta k adresáři, který je již vytvořen. Veškeré nastavení parametrů bylo provedeno pomocí souboru `universe_wsgi.ini` a podle internetové stránky „Running Galaxy Tools on a Cluster“[8].

3.3 Instalace PostgreSQL a konfigurace Galaxy

Galaxy po instalaci využívá integrovaného databázového nástroje SQLite. SQLite je databázový nástroj, který neobsahuje samostatný server, celá databáze je tvořena jedním souborem, do kterého zapisuje přímo a z kterého přímo čte[10]. Z toho vychází omezení, že veškeré databázové procesy jsou prováděny pod procesem Galaxy. Proces, není-li volný, nemůže dělat nic jiného, při větším zatížení roste riziko uzamykání transakčních zámků[10]. Z těchto důvodů je doporučeno využít databázového serveru, konkrétně PostgreSQL databázového serveru.

Instalace PostgreSQL byla provedena z repozitáře distribuce Debian, dostupná verze PostgreSQL v době instalace z repozitáře byla 8.4.9 (provedeno pomocí příkazu `aptitude install postgresql`). Po instalaci je třeba vytvořit novou databázi, jež bude použita pro Galaxy. Také je třeba vytvořit uživatele, který bude moci spravovat databázi. Pravomoc vytvářet uživatele a databáze po instalaci má pouze uživatel **postgres**, je tedy nutné se nejdříve jako tento uživatel přihlásit. Toho docílíme, když jsme přihlášení jako root, příkazem `su - postgres`, nyní máme práva superuživatele v databázovém serveru.

V ytvoření uživatele provedeme příkazem `createuser -P -E galaxy`, kde parametr "P" slouží pro nastavení hesla a parametr "E" zajistí uložení hesla v zašifrované formě. Pomocí příkazu `createdb -o galaxy galaxy_db` se vytvoří databáze se jménem `galaxy_db`, jejímž vlastníkem je uživatel `galaxy`, čehož bylo dosaženo parametrem "-o"[11].

Nastavení Galaxy provedeme pomocí souboru `universe_wsgi.ini`, kde se v sekci `Database` nachází parametr `database_connection`, ten doplníme o příslušnou definici připojení k databázi, tedy
`database_connection = postgres://galaxy:heslo@localhost/galaxy_db`. Z čehož parametr `postgres` slouží pro určení, o jaký databázový server se jedná, část `galaxy:heslo@localhost` určuje jméno uživatele, který přistupuje k databázi, jeho heslo a doménu, která může obsahovat i port, na kterém databázový server běží. Za posledním lomítkem se nachází `galaxy_db`, což označuje název databáze. Další možností využití vlastností PostgreSQL je `database_engine_option_server_side_cursors = True`, díky čemuž jsou veškeré databázové požadavky zpracovávány v databázi a Galaxy se předají pouze požadované řádky, tím se sníží náročnost Galaxy na paměť. Jednoho databázového připojení na vlákno lze docílit pomocí
`database_engine_option_strategy = threadlocal`, čímž se zabrání nadměrné režii[10].

3.4 Instalace ProFTPD a následná konfigurace spolu s Galaxy

Galaxy umožňuje nahrávání souborů pomocí HTTP protokolu, toto řešení je však vhodné pouze na soubory, jejichž velikost se pohybuje v řádech jednotek megabytů. Nahrání větších souborů je většinou ukončeno chybovou hláškou. Po zjištění, že repozitář operačního systému Debian obsahuje zastaralou verzi ProFTPD, bylo rozhodnuto stáhnout zdrojové kódy ze stránek projektu a ná-

sledně provést kompilaci[6]. V době instalace je použita aktuální verze 1.3.4a, jelikož ProFTPD má fungovat na šifrované verzi FTP a přihlášení má být provedeno pomocí uživatelských účtů vytvořených v Galaxy, je třeba zahrnout při kompilaci moduly `mod_sql, mod_tls` a parametr `enable-openssl`.

Po stažení a rozbalení zdrojových souborů dojde ke kompilaci se zahrnutím modulů, které potřebujeme pro provoz. Kompilaci a instalaci provedeme sadou příkazů `configure, make, make install`[13]. Při příkazu `configure` je třeba určit, které moduly chceme, aby byly zahrnuty. Pro instalaci na Galaxy byl použit tento příkaz
`configure --enable-openssl --with-modules=mod_sql:mod_sql_postgres:mod_sql_passwd:mod_tls`[14][15]. Z čehož parametr `enable-openssl` zajišťuje funkčnost se službou OpenSSL[7], `mod_sql` slouží pro komunikaci mezi ProFTPD a databázovým serverem, `mod_sql_postgres` definuje, jakého databázového serveru bude využito a `mod_tls` slouží k využití funkčnosti služby TLS[8].

Před samotnou konfigurací ProFTPD je potřeba zajistit patřičnou část databáze, pomocí které se budou ověřovat uživatelské účty. Dále vytvořit nového uživatele databáze, který bude mít přístup k uživatelskému jménu, heslu, pro možnost ověření skrz ProFTPD. Nového uživatele vytvoříme použitím již zmíněného příkazu `createuser galaxyftp`. Ten nám vytvoří nového uživatele bez práv vytvoření databáze a práva na vytváření rolí. Je třeba nově vytvořeného uživatele přiřadit do databáze a určit mu roli. Toho docílíme parametrem `ALTER ROLE`, jehož pomocí nastavíme atributy role pro účet. Nejdříve je potřeba přihlásit se do databáze Galaxy, na to lze využít `psql galaxy_db`, nyní

[6]překlad zdrojových kódů pomocí překladače programovacího jazyka do výsledného fungujícího programu

[7]Open Source soubor nástrojů realizujících Secure Sockets Layer, protokol využívající kryptografických knihoven pro šifrovanou komunikaci[12]

[8]Transport Layer Security, protokol využívající kryptografických knihoven pro šifrovanou komunikaci[12]

lze zadat `ALTER ROLE galaxyftp PASSWORD 'heslododatabaze';`[18], tímto má uživatel "galaxyftp" nastavenou roli. Role je v PostgreSQL brána jako koncepce oprávnění pro přístup k databázi[16]. Dále parametr
`GRANT SELECT ON galaxy_user TO galaxyftp;` tento příkaz dává zvláštní oprávnění na objekt databáze, konkrétně oprávnění na příkaz "SELECT" z tabulky "galaxy_user", pro uživatele "galaxyftp"[17].

Nastavení ProFTPD je provedeno pomocí konfiguračního souboru `proftpd.conf`. V tomto souboru byly nastaveny následující parametry.
`Include /usr/local/etc/proftpd-tls.conf` slouží pro načtení dalšího konfiguračního souboru, na nějž odkazuje. `ServerName "Public Galaxy FTP"` nastaví jméno serveru, `ServerType standalone` nastavuje typ serveru, kde "standalone" je režim, ve kterém ProFTPD naslouchá příchozím FTP relacím a sám je vyřizuje pomocí podřízených procesů. Tento režim je vhodný pro vysoký provoz, kde veškeré nastavení se provádí pomocí konfiguračních souborů pro ProFTPD. V režimu "inetd" se očekává, že bude spuštěn pomocí inetd/xinetd, který naslouchá na portu 21, pro požadavky připojení. Toto je vhodné pro nízký provoz, obsluhující malé FTP session. `Port 21` nastaven port, na kterém ProFTPD naslouchá. `User galaxy` a `Group galaxy` slouží pro určení uživatele a skupiny, pod kterou bude server spuštěn.
`PassivePorts 30000 40000` rozsah portů, jež server využije pro komunikaci po již proběhlé autorizaci. `CreateHome on dirmode 700` dovoluje uživateli automaticky vytvořit domovskou složku s danými právy. `AllowStoreRestart on` pokud dojde k přerušení nahrání souboru, je možno po znovu obnovení pokračovat od přerušeného místa. `Umask 037 037` nastavuje nově vytvořeným souborům práva, v tomto případě "umask" zahrnuje plná práva pro vlastníka souboru, právo číst pro skupiny a všichni ostatní nemají žádné právo. `MaxInstances 30` tímto omezujeme počet souběžných připojení v jednom čase,

ochrana proti DoS útoku[9]. `DefaultRoot ~` omezení uživatele, aby mohl být pouze ve své složce.
`ListOptions "-A"` umožňuje výpis adresáře i souboru vyjma "." a "..".

Využití SQL v ProFTPD je dosaženo pomocí parametrů, jež jsou zadané v konfiguračním souboru. `SQLPasswordEngine on` spouští modul `sql_password` umožnující SQL přístup k heslům v databázi.
`SQLPasswordEncoding hex` nastaví kódování, které bude očekáváno při manipulaci s daty uloženými v databázi. `SQLEngine on` zajišťuje spuštění modulu pro využívání knihovny SQL. `SQLBackend postgres` určuje, jaký „backend[10]“ má být využit, pro funkci s PostgreSQL musí být parametr `postgres`, pro funkčnost MySQL databáze slouží parametr `mysql`.
`SQLConnectInfo galaxy@localhost galaxyftp heslododatabaze` zde jsou informace nutné pro připojení do databáze, jako první je uživatel, pod kterým je spuštěna databáze Galaxy spolu se serverem, na kterém běží, následuje uživatel, který existuje v databázi a má přístup k přihlašovacím údajům. Předem již vytvořený uživatel galaxyftp. `SQLAuthTypes SHA1` určuje metodu, pomocí níž jsou hesla šifrována. `SQLAuthenticate users` nastavení ověřovacích funkcí modulu `mod_sql`. `SQLUserInfo custom:/LookupGalaxyUser` slouží pro definování vlastní syntaxe a odkazujeme se na konfigurovatelný příkaz `SELECT SQLNamedQuery`.
`SQLNamedQuery LookupGalaxyUser SELECT "email,password, '1024', '1024','/home/cesta_ke_slozce/ftp/%U','/bin/bash' FROM galaxy_user WHERE email='%U'"` nakonfigurovatelný příkaz SELECT, který má sloupce: "e-mail, password, UID - číslo uživatele, GID - číslo skupiny, cesta k adresáři, kde se ukládají soubory z FTP, kde %U značí proměnnou, jejíž hodnota je uživatelův e-mail, cesta ke shellu FROM z tabulky "galaxy_user" WHERE

[9]DoS - útok provedený vytvořením spousty dotazů, které se směřují na danou službu s cílem zahltit ji požadavky a učinit ji tak nedostupnou

[10]balík, který v sobě nese informace o přístupu a příkazech dané databáze

kde "email=je roven uživatelovo e-mailu". Nastavení cesty, kam FTP ukládá soubory, je také třeba nastavit v souboru `universe_wsgi.ini` parametrem `ftp_upload_dir =/home/cesta_ke_slozce/ftp` čímž zajistíme, aby Galaxy mohl přistoupit do této složky pro nahrání uloženého souboru.

Nastavení souboru `proftdp-tls.conf` pro použití SSL/TLS komunikace. Jelikož se jedná o rozšiřující modul ,je třeba psát veškeré nastavení mezi parametr `<IfModule mod_tls.c>` a uzavírací parametr `</IfModule>`. Použití TLS zajišťuje parametr `TLSEngine on`, parametrem
`TLSLog /var/log/proftpd-tls.log` se určuje, kam se bude zapisovat log. Využití SSL verze 3 a TLS verze 1 zajistí parametr `TLSProtocol SSLv3 TLSv1`. Vynucení použití komunikace pomocí TLS při kontaktování FTP klienty provede `TLSRequired on`. Parametry
`TLSRSACertificateFile /usr/lib/proftpd/openssl-cert/server.crt`
`TLSRSACertificateKeyFile /usr/lib/proftpd/openssl-cert/server.key`
se definuje kde je certifikát serveru a klíč. Ověřování klientů, kteří chtějí použít FTP přes TLS vypne parametr `TLSVerifyClient off`. Vypnutí vyžadování "SSL renegotiation[11]" se provede pomocí `TLSRenegotiate required off`. Pro vygenerování certifikátu byl použit příkaz
`openssl req -new -x509 -days 365 -nodes -out /usr/lib/proftpd/openssl-cert/server.crt -keyout /usr/lib/proftpd/openssl-cert/server.key`.[23]

[11] umožňuje modifikovat šifrovací klíč pro zabezpečenou relaci bez ukončení existujícího spojení[22]

4 Implementování vlastních programů

4.1 Zobrazení programů v Galaxy

Zobrazení programů je zajištěno díky definici pomocí XML souboru `tool_conf.xml`, ten ve své definici nese odkazy na programy, jež se zobrazují na webové stránce Galaxy. Jak je vidět na obrázku 1. XML soubor obsahuje následující strukturu.

```
<?xml version="1.0"?>
 <toolbox>
 <section name="Název sekce" id="oznaceni">
 <label text="Popisek" id="oznaceni_popisku" />
 <tool file="cesta/k_souboru.xml"/>
 </section>
</toolbox>
```

Párový tag `<toolbox>` označuje hlavní oblast, jež se zobrazuje na stránce Galaxy. Uvnitř se nachází párový tag `<section>`, ten slouží pro zobrazování kategorií, otevírací tag obsahuje parametr `name=""`, do nějž se zadává název kategorie, ten se zobrazí v Galaxy a má funkci rozklikávací nabídky. Další parametr je `id=""`, pomocí nějž se identifikuje daná kategorie. Zadaný identifikátor musí být jedinečný, měl by obsahovat malá písmena a číslice. Pro snadnější organizaci programů v jedné kategorii, pokud je v této kategorii více programů, přičemž určité programy vytvářejí specifické skupiny, je možno vložit popis a to pomocí tagu `<label />`, ten obsahuje parametr `text=""`, do nějž se zadá text, který chceme, aby byl zobrazen. Parametr `id=""` slouží pro identifikaci popisku a měl by zase splňovat požadavky na jedinečnost spolu se zápisem malými písmeny, číslicemi nebo podtržítkem. Zobrazení programu je zjištěno pomocí tagu `<tool />` jehož parametr `file=""` slouží pro odkaz na hlavní XML soubor daného programu, do parametru se zadává cesta k onomu

souboru. Ve výchozím nastavení se vychází z toho, že hlavní složka obsahující programy, je složka jménem `tools`, tudíž se zadává cesta v podobě adresářů obsažených v adresáři `tools`. Pro příklad je v adresáři `tools` adresář `umbr_programs`, který obsahuje adresář `seqclust` uvnitř něhož je program `seqclust` společně s hlavním souborem `seqclust.xml`. Výsledný parametr tak bude `<tool file="umbr_programs/seqclust/seqclust.xml">`, tím je dosaženo odkázání na hlavní XML soubor programu, jež v sobě nese název, popis a další údaje. Díky tomuto je na webové stránce Galaxy v oblasti programů zobrazen název programu společně s popisem a zároveň slouží jako odkaz na stránku programu.

4.2 Tvorba XML souboru vlastního programu

Hlavní věcí, kromě vytvoření samotného programu, který se bude starat o potřebné výpočty, je vytvoření XML souboru. Pomocí definic XML souboru Galaxy zajistí zobrazení webové stránky, pomocí které komunikuje s uživatelem a následně programem. Struktura XML souboru je následující.

```
<tool id="seqclust" name="Graph Based Sequence Clustering" versi-
on="1.0.0">
<description>Repeat discovery and characterization using graph
based sequence clustering </description>
<command interpreter="bash">
seqclust-main.sh -s $input -m $mvstup -o $ovstup -v $vystup1 -w
$vystup2 -y $vystup3 -x pdffile.pdf -z RMSKtable -r $__root_di-
r__ -t $output_html -u '$output_html.extra_files_path'
</command>
<inputs>
 <param format="fasta" type="data" name="input" label="input DNA
 sequences" help="sequences from shotgun genomic sequencing" />
```

```
 <param name="ovstup" type="integer" size="6" value="50" label="
 minimal overlap for assembly" help="this parameter affect assem-
 bly but not clustering" />
 <param name="mvstup" type="integer" size="6" value="100" label="
 Number of reads in a cluster to make its directory" />
 <param name="paired" type="boolean" truevalue="true" falsevalue=
 "false" checked="False" label="All sequence reads are paired"
 help="check if you are using pair reads and and input sequences
 contain both read mates and  left mates alternate with their
 right mates"/>
 </inputs>

<outputs>
 <data format="zip" name="vystup1" label="Archive with cluster-
 ing results from ${input.name}" />
 <data format="fasta" name="vystup2" label="Contigs from ${inp-
 ut.name} based on clustering" />
 <data format="txt" name="vystup3" label="Log file from (from
 ${input.name}" />
 <data format="html" name="output_html" label="HTML summary of
 graph based clustering of ${input.name} " />
</outputs>

<help>
**What it does**
All to all sequence comparison of sequence reads is performed
using megablast.
</help>
```

```
</tool>
```

Výše uvedená definice XML je ze souboru „seqclust.xml“ (Příloha C), jež zajišťuje definování parametrů potřebných pro program vyvinutý na ÚMBR, který zpracovává sekvence DNA a jehož výstupem jsou grafy, soubory s hity sekvencí a souboru archívu, v němž jsou všechny soubory z výstupu obsaženy. V XML souboru pro program je několik parametrů Všechny parametry, které je možné zadat jsou zveřejněny na webových stránkách „Galaxy Tool XML File“[19].

Úvodní párový tag `<tool>` obsahuje parametr `id=""`, který nese jedinečné označení programu. Platí zde stejné omezení jako pro ”id” v souboru `tool_conf.xml`, parametr `name=""` slouží pro zadání názvu programu, který je zobrazen v panelu programů na webové stránce Galaxy, a také na webové stránce programu zobrazené v Galaxy. Poslední parametr `version=""` slouží pro zadání čísla verze programu. Párový tag `<description>` je využit pro popis charakterizující program. Tento popis je zobrazen v panelech programů hned za názvem programu.

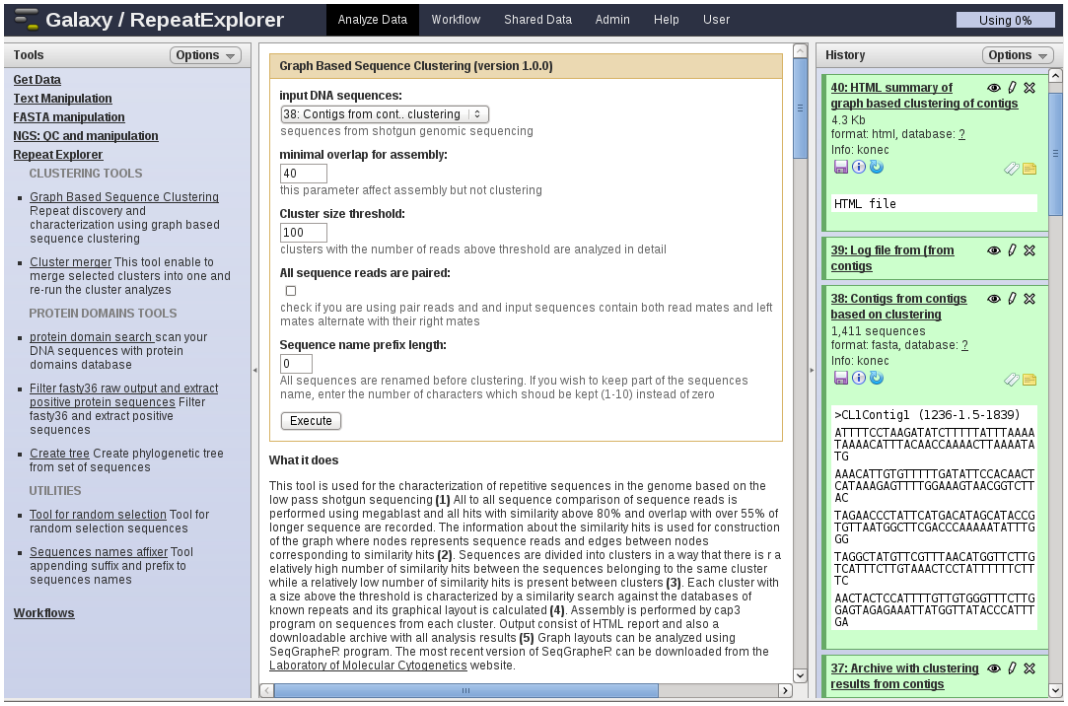
4.2.1 Volání programu v XML souboru

Další párový tag `<command>` slouží pro definování vytvořeného programu, který se stará o výpočty a další operace, tento tag obsahuje parametr `interpreter="bash"`, pomocí něhož se určuje, jakým programovacím jazykem je napsán program, který chceme spustit. Za otevírací tag `<command>` se zadá název programu, který má být spuštěn spolu s parametry. Parametr `__root_dir__` je definován pomocí proměnné v Galaxy a obsahuje cestu k adresáři, ve kterém je Galaxy nainstalován, takzvaný domovský adresář. Další parametr `$output_html.extra_files_path` je definován také pomocí Galaxy a slouží pro vytvoření složky se stejným jménem jako výstupní soubor, kdy je tato složka uložena společně s výstupním souborem v adresáři pro ukládání souborů

a obsahuje soubory, které je možno do ní pomocí programu uložit. V souboru `extractfasta.xml` je použit parametr `$__app__.config.job_working_directory`, pomocí něhož je předána cesta k adresáři `job_working_directory`, ve kterém jsou vytvářeny dočasné složky sloužící pro vykonání programů, jež jsou uživatelem spuštěny a pomocí Galaxy předány clusteru.

4.2.2 Vstupní část XML souboru

Pomocí párového tagu `<inputs>` se označují vstupy, které jsou zobrazeny na webové stránce, buďto jako textové pole, vybírací seznam nebo zaškrtávací políčko. Na obrázku 2 je vidět, jak Galaxy zobrazí stránku programu, na níž jsou obsaženy vstupní parametry. Každý vstup je jeden tag `<param />`, tento tag má několik parametrů. Parametr `name=""` slouží pro jedinečné pojmenování, které je použito jako parametr pro vstupní soubor programu. Parametrem `type=""` se určuje datový typ vstupních proměnných, jehož hodnotou může být „text“ – určující, že vstupní data budou ve formě textu získaná z formuláře, „integer“ a „float“ – značí, že vstupní data budou čísla získaná z formuláře, „boolean“ – reprezentuje zaškrtávací políčko a předává logickou hodnotu, „data“ – značí, že uživatel vybere soubor z dostupných souborů, které uložil do Galaxy nebo je získal jako výstup vykonaného programu pod Galaxy. Obsahuje-li parametr `type="data"` hodnotu „data“, přidá se parametr `format="tabular"`, pomocí kterého se nastaví typ souboru. Další parametr `label=""` slouží pro zadání názvu či popisu, pomocí tohoto, je známo o jaký vstup se jedná. Zároveň je tento text na webové stránce zobrazen nad vstupním polem tučným písmem. Je-li potřeba přidat vysvětlující popis ke vstupnímu parametru, provede se to pomocí parametru `help=""`, do něhož se zapíše vysvětlující text, jenž je na webové stránce programu zobrazen pod vstupním polem. Nastavení velikosti vstupního pole je dosaženo pomocí `size=""`. Je-li



Obrázek 2: Galaxy a zobrazení stránky programu.

`type="integer"` nebo `type="float"` dalším parametrem, může být `min=""` určující minimální možnou hodnotu, která smí být zadána, `max=""` naopak určuje maximální hodnotu, jež může být zadána. Zadá-li uživatel jinou hodnotu, která je mimo dovolený rozsah, je na tuto skutečnost upozorněn před vykonáním programu pomocí zčervenání vstupního pole a upozorňujícího textu, že zadaná hodnota není v mezích povoleného rozsahu.

Pro příklad v programu `extractfasta.xml` (Příloha D) je tento tag `<param name="cvstup" type="integer" size="3" value="60" min="0" max="100" label="minimum similarity [%] " />` po zadání špatné hodnoty je uživatel upozorněn, že hodnota může být pouze v rozmezí 0 až 100. Předvyplněné pole je možno nastavit pomocí parametru `value=""`, přičemž hodnota v tomto parametru je následně zobrazena ve vstupním poli na webové stránce, a také předána jako parametr, není-li změněna. Pokud je parametr `type="boolean"`, pak parametr `truevalue=""` nese hodnotu, jež je zadána

a tu předává jako parametr, pokud je vyhodnocení pravdivé, není-li vyhodnocení pravdivé, pak se předá hodnota již obsahuje parametr `falsevalue=""`. Parametr `checked=""` nastavuje výchozí hodnotu zaškrtávacího políčka, hodnota `False` zobrazí nezaškrtnuté políčko a hodnota `True` zobrazí zaškrtnuté políčko. Pokud je parametr `type="select"`, musí být určeny hodnoty, ze kterých se bude vybírat. To se dělá pomocí párového tagu `<option>`. Pro příklad ukázka ze souboru `maketree.xml` (Příloha E), který obsahuje tyto hodnoty.

```
<param name="mvstup" type="select" label="distance method">
 <option value="0" selected="true">uncorrected</option>
 <option value="1">Jukes-Cantor</option>
 <option value="2">Kimura</option>
</param>
```

Parametr `value=""` v tagu `<option >` nese hodnotu, která se následně předá jako hodnota vstupního parametru `$mvstup`. Parametr `selected="true"` určuje, že ta hodnota, která tento parametr obsahuje, bude nastavena jako výchozí hodnota. Text zadaný mezi otevíracím tagem `<option >` a uzavíracím tagem `</option>`, se bude zobrazovat ve výběrové nabídce čímž reprezentuje danou hodnotu. Další možností, jak vytvořit vybírací nabídku, je použití textového souboru, který má podobu tabulek, kde jednotlivé sloupce nesou požadované hodnoty, jako je jméno pro zobrazení a hodnotu, jež bude předána parametrem. V souboru `fastaload.xml` (Příloha F) je této možnosti využito, pro ukázku část kódu, která této možnosti využívá.

```
<param name="domains" type="select" label="Choose protein domain
type">
 <options from_file="protein_domains.txt">
 <column name="name" index="2" />
 <column name="value" index="0" />
 </options>
```

Parametr `from_file="protein_domains.txt"` určuje, že se hodnoty pro výběrové menu budou brát ze souboru `protein_domains.txt`, který se nachází ve složce `tool-data`. To je výchozí složka, do které je potřeba zadat veškeré soubory, které by měly být použity pro služby výběrové nabídky. Tato složka se nachází v domovském adresáři Galaxy. Každá položka ve výběrové nabídce reprezentuje jeden řádek ze zdrojového souboru. Parametr `name="name" index="2"` určuje, že tato hodnota se bude zobrazovat jako hodnota jednotlivých položek výběrového menu a číslo v parametru `index=""` označuje číslo sloupce zdrojového souboru, ve kterém jsou obsaženy názvy. Parametr `name="value" index="0"` určuje hodnotu, jež se bude předávat parametrem, přičemž tato hodnota se nachází v nultém[12] sloupci zdrojového souboru. Zdrojový soubor má následující strukturu. Řádek označen symbolem `#` je vyhodnocen jako komentář.

#Code	Name	Nick
GAG	TE_domains_newest_GAG	GAG
CHDCR	TE_domains_newest_CHDCR	CHDCR
CHDII	TE_domains_newest_CHDII	CHDII
INT	TE_domains_newest_INT	INT
PROT	TE_domains_newest_PROT	PROT
RH	TE_domains_newest_RH	RH
RT	TE_domains_newest_RT	RT

4.2.3 Výstupní část XML souboru

Další částí XML souboru je definování výstupů, tuto úlohu zajišťuje párový tag `<outputs>`, výstupní data se definují pomocí tagu `<data />`. Výstup v Galaxy je pojat tak, že výstup je soubor a co výstup, to samostatný soubor, což ve skutečnosti znamená definovat každý výstup zvlášť. Má-li program tři výstupní

[12] označováno řečí programátora

soubory, například jeden PDF[13] soubor s textem, druhý PDF soubor obsahující grafy a třetí textový soubor obsahující log[14] programu, je nutno definovat ve výstupní části XML souboru tři výstupy. Programy vytvořené na ÚMBR produkují desítky až stovky výstupních souborů, proto byla zvolena možnost všechny tyto soubory vložit do jednoho souboru, jehož datový typ je archív, čímž se zajistí získání jednoho výstupního souboru. Druhou možností je použití parametru `extra_files_path` pro ukládání souborů do složky pomocí extra cesty (zmíněno v podsekci 4.2.1). Kdy se soubory uložené do této složky zobrazí za pomoci jednoho HTML výstupního souboru, ve kterém se vytvoří odkazy tak, aby odkazovaly na soubory ve složce.

Tag `<data />` má parametr `format=""`, který slouží pro určení typu, jež bude mít výstupní soubor. Parametr `name=""` slouží pro jedinečné označení, které je použito jako parametr pro výstupní soubor programu. Parametrem `label=""` se určuje text, který bude zobrazen na webové stránce v části výstupních souborů. Pomocí tohoto textu se identifikují již provedené výstupy z programů. V tomto parametru je možno použít proměnných a vnitřních funkcí Galaxy pro ještě lepší upřesnění, o jaký výstup s jakými parametry se jedná. Pro příklad část kódu ze souboru `extractfasta.xml`.

```
<data format="fasta" name="vystupTy1" label="fasta output Ty1
(from-${input.name}, with sensitive: c-${cvstup} d-${dvstup}
s-${svstup} r-${rvstup})" />
```

Zde proměnná `input.name` slouží pro přidání názvu vstupního souboru, konkrétně názvu vstupního souboru, jež má `name="input"`. Další možností upřesnění je přidání hodnot vstupních proměnných. Přidáním proměnných `${cvstu-`

[13]Portable Document Format - formát vyvinutý firmou Adobe Systems je globálním standardem pro vytváření a prohlížení souborů s libovolným obsahem. Zajišťuje stejné zobrazení na jakémkoliv operačním systému, čímž jej tvoří snadno sdílným s kýmkoliv a kdekoliv.[20]

[14]záznamy programu obsahující výpisy programu o průběhu programu, chybových hlášení, či vlastních příkazů nastavených pro zápis do logu

`p},`

`${dvstup}, ${svstup}, ${rvstup}` docílíme zobrazení vstupních hodnot těchto proměnných. V případě souboru `extractfasta.xml` jde o zobrazení čísel, protože vstupní hodnoty jsou číselného typu.

4.2.4 Zbylá část XML souboru

4.2.5 Část pro HELP

Jako další součást XML souboru je párový tag `<help>`, který slouží pro zobrazení nápovědy pro uživatele. Tato nápověda může popisovat jak samotný program, tak práci s programem, případně vysvětlit, co způsobí jaké chyby, které se mohou vyskytnout. Nápověda je zobrazena pomocí webového rozhraní programu pod sekcí vstupních parametrů. Definování struktury textu se provede pomocí parametrů z reStructuredText[15]. Pro ukázku přehled některých parametrů.

[15]prostý text využívající jednoduché a intuitivní konstrukce pro vytvoření dokumentu[21]

parametr	význam
.. class:: warningmark	zobrazí žlutý trojúhelníček s vykřičníkem
.. class:: infomark	zobrazí modrý čtverec se zaoblenými rohy a symbolem i
.. image:: ./static/images/obrazek.png	zobrazí obrázek, cesta musí být do adresáře static/images, který je v domovské složce Galaxy
.. _ukazatel: http://galaxy.umbr.cas.cz	slouží pro definování odkazu na internetovou stránku
jakýkoliv text::	vytvoří takzvaný paragraph, kdy text psaný za tímto bude malým písmem a zarovnán doleva
text mezi	tento text bude kurzívou
text mezi	tento text bude tučný
* text	vytvoří seznam, další symboly pro tvorbu seznamu jsou "-", "+"
-----	vytvoří horizontální čárkovanou čáru

Další značky, které je možné použít jsou na stránkách projektu „reStructuredText Markup Specification“[21].

4.2.6 Část pro test

V XML souboru je možno vytvořit testovací funkce, které po spuštění zjistí, zda byl program správně vykonán. Testování se provádí pomocí definování vstupů a definici očekávaných výstupů, při čemž se provede jejich porovnání s výstupem z programu, kterému byl předán definovaný vstup. Struktura má tuto podobu.

```
<tests>
<test>
<param name="input" value="test_soubor_pro_testovani.txt" />
<output name="output" file="test_soubor_pro_testovani_vystup.tx-
t">
<assert_contents>
<has_text text="Hlavicka"/>
<not_has_text text="dneska"/>
</assert_contents>
</output>
</test>
</tests>
```

V tagu `<param />`, který určuje vstupní část testovací části je parametr `name="input"`, ten slouží pro určení vstupního souboru. Jméno musí být shodné se jménem, které je dáno u vstupního parametru vstupní části XML souboru. Parametr `value="soubor.pripona"` slouží pro identifikování testovacího vstupního souboru. Tag `<output>` slouží pro definování výstupního testovacího souboru a jeho parametr `name="output"` musí mít stejný název jako výstup ve výstupní části souboru. Parametr `file="vystup_soubor.pripona"` definuje výstupní soubor testovací části. Vstupní a výstupní soubory testovací části se nacházejí v adresáři `test-data`, jenž je umístěn v domovském adresáři Galaxy. Párový tag `<assert_contents>` slouží pro definování sekvencí kontrol, pomocí kterých je možné detailněji určit testovací parametry.

5 Testování

5.1 Testování Galaxy

Po prvním spuštění Galaxy serveru je možno použít příkaz
`sh run_functional_tests.sh` z domovského adresáře Galaxy. Ten zajistí vykonání interních testů, které jsou distribuovány spolu s Galaxy, jež zajistí otestování všech programů Galaxy, které mají definovány testovací parametry v XML souborech. Výstupním souborem je HTML stránka nacházející se v domovském adresáři, jejíž název bude `run_functional_tests.html`. Tato HTML stránka podává přehled o všech provedených testech, společně s jejich výsledkem. Pomocí tohoto otestování je možné zjistit nefunkčnost některého programu, která se nedá zjistit z logu při spuštění a projevila by se až při použití onoho programu. Pomocí `sh run_unit_tests.sh` lze otestovat pouze Galaxy bez zahrnutí testovacích součástí programů.

Pro prvotní testování programů vložených do Galaxy společně s otestováním, jak Galaxy server komunikuje s PBS, byly použity soubory obsahující sekvence DNA, jež poskytl ÚMBR. Tyto soubory měly přiměřený počet sekvencí pro zajištění, že výpočty budou provedeny v relativně krátkém čase, řádově desítky minut až hodin. Po předložení testovacího souboru programu se sledovalo, zda se program ukončí bez chyby a zda výstupní soubor má požadovaný obsah. Při vykonávání programu bylo pomocí příkazů PBS zjišťováno, zda je výpočet vykonáván na nodech[16] clusteru. Pro ověření, že Galaxy dodržuje nastavené parametry, bylo spuštěno několik úloh pod jedním uživatelem, ze kterých byly vykonávány současně pouze tři, čímž došlo k potvrzení fungujícího omezení počtu spuštěných úloh na uživatele.

Po provedení několika testovacích úloh, které trvaly různě dlouhou dobu, se vyskytla chyba, kdy při vykonání programu byla absence výstupu. Tento

[16] výpočetní uzly clusteru

problém se projevoval u výpočtů, jež měly krátkou dobu výpočtu. Po vyloučení možnosti, že je chyba způsobena programem, bylo zjištěno, že se jedná o chybu, kdy síťová služba zajišťující přenos výpočtů mezi jednotlivými nody, ukládá data. V době po vykonání programu však Galaxy uložená data neviděl. Tato chyba se projevuje při použití NFS[17] mezi jednotlivými nody clusteru. V souboru `universe_wsgi.ini` je zmíněn parametr `retry_job_output_collection = 5`, který určuje počet pokusů při dívání se na PBS server pro výsledná data, kdy mezi jednotlivými pokusy čeká sekundu. Další možností je použití parametru `-noac` při připojování pomocí NFS. Ovlivnění NFS však neslo pomalejší přenos souborů, to bylo posléze pracovníky ÚMBR vyhodnoceno jako neúnosné a upustilo se od tohoto řešení. Jako výsledné řešení se proto do programů použilo přidání funkce čekací smyčky, která je spuštěna po výpočtech programu a má za úkol čekat minutu. Tím je zajištěno, že každý program bude trvat minimálně minutu, čímž by se mělo předejít výše zmíněné chybě.

5.1.1 Testování ProFTPD

Pro otestování, zda ProFTPD využívá nastaveného šifrovaného přenosu, bylo použito testování síťového provozu pomocí programu Wireshark. Nejprve byl ProFTPD server spuštěn bez šifrování. Ve zjištěných paketech, sloužících pro FTP, bylo zjištěno snadno čitelné heslo, které bylo zadáno při přihlašování. Po spuštění ProFTPD serveru s využitím šifrování se v příslušných paketech nacházel pouze šifrovaný obsah.

Při uložení souboru pomocí FTP, byly veškeré soubory neviditelné Galaxy serverem. Bylo zjištěno, že uložené soubory nejsou vlastněny uživatelem „galaxy“, pod kterým byl ProFTPD server spuštěn. Způsobily to parame-

[17]Network File System, síťový souborový systém sloužící pro sdílení souborů v síti, nebo připojení síťových disků.

try `SQLMinUserGID 620 SQLMinUserUID 620`, díky nimž byly soubory ukládány pod jiným UID[18] a GID[19][24]. Řešení se skrývá v nastavení parametrů `SQLDefaultGID 1024 SQLDefaultUID 1024`, kdy čísla byla nastavena tak, aby odpovídala číslům uživatele „galaxy“.

Funkčnost ProFTPD při používání začala vykazovat chyby při nahrávání souborů, kdy soubory se nahrávaly ve smyčce bez ukončení, nebo nešly nahrát vůbec. Po prohlédnutí logu z ProFTPD serveru byla nalezena chyba `client did not reuse SSL session, rejecting data connection`.
Někteří FTP klienti neumí opětovně použít SSL sezení, pro vyřešení tohoto problému se do souboru `proftpd-tls.conf` přidal parametr
`TLSOptions NoSessionReuseRequired`.

6 Závěr

V této práci bylo představeno základní seznámení se strukturou domovského adresáře `galaxy-dist` serveru Galaxy. Představeny byly možnosti konfigurace a názorná konfigurace pro využití Galaxy společně s databázovým serverem PostgreSQL. Také byly ukázány základní příkazy, s jejichž pomocí se vytvořila nová databáze a nový uživatel databáze v PostgreSQL. Představena byla také konfigurace, jež umožňovala použití PBS. Další část práce ukázala, jak nakonfigurovat ProFTPD server pro společné fungování s Galaxy.

Druhá část práce zabývající se implementací programů představila možnosti základního nastavení souborů XML tak, aby bylo možno pomocí XML souboru obsluhovat daný program. Tvorba XML byla rozdělena na část vstupní, která zajišťovala definování vstupních parametrů, jež se předávaly programu. Ve výstupní části se definovaly výstupy, které program má a jak je bude Galaxy prezentovat. Poslední částí bylo předvedení možností, jak vytvořit nápo-

[18] číslo uživatele, pod kterým je uživatel identifikován v systému

[19] číslo skupiny ve které je uživatel zařazen a pomocí níž je v systému identifikován

vědu, která slouží pro objasnění, co program vykonává nebo jaké jsou potřeba vstupní soubory.

Cíle práce byly splněny. Přínos práce pro mé vlastní poznání byl velký. Práce mi umožnila náhled do dějů bioinformatických výpočtů, jež pracují se sekvencemi DNA. Dalším přínosem pro mě byla možnost práce s clusterem a FTP serverem. Práce mě také v některých případech naučila trpělivosti postupnému přemýšlení, před zbrklým jednáním a následné kolizi v podobě ztracení části databáze či poničení konfigurace Galaxy.

Literatura

[1] FrontPage - Galaxy-Wiki. *Galaxy-Wiki* [online]. [b. r.], last modifed 3.2.2012 [cit. 2012-02-14]. Dostupné z: http://wiki.g2.bx.psu.edu/FrontPage

[2] Jeremy Goecks, Anton Nekrutenko, James Taylor, The Galaxy Team. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biology 2010 11:R86* [online]. ©2010 [cit. 2012-3-29]. Dostupné z: http://genomebiology.com/2010/11/8/r86

[3] *The ProFTPD Project: Home: ProFTPD: Highly configurable GPL-licensed FTP server software* [online]. The ProFTPD Project, ©1999-2011 [cit. 2012-02-15]. Dostupné z: http://proftpd.org/

[4] *Mercurial - Mercurial: Mercurial* [online]. Mercurial, [b. r.], last modifed 21.1.2012 [cit. 2012-02-15]. Dostupné z: http://mercurial.selenic.com/wiki/

[5] *OGF DRMAA Working Group* [online]. [b. r.][cit. 2012-02-16]. Dostupné z: http://www.drmaa.org/

[6] Kristina Wanous *Resource Manager: Torque* [online]. 2008 [cit. 2012-02-16]. Dostupné z: http://debianclusters.org/index.php/Resource_Manager:_Torque

[7] Kristina Wanous *Scheduler: Maui* [online]. 2008 [cit. 2012-02-16]. Dostupné z: http://debianclusters.org/index.php/Scheduler:_Maui

[8] Admin/Config/Performance/Cluster: Running Galaxy Tools on a Cluster. *Galaxy-Wiki* [online]. [b. r.], last modifed 2012-01-23 [cit. 2012-02-16]. Dostupné z: http://wiki.g2.bx.psu.edu/Admin/Config/Performance/Cluster

[9] *About SQLite* [online]. [b. r.][cit. 2012-03-03]. Dostupné z: http://www.sqlite.org/about.html

[10] Admin/Config/Performance/Production Server: Running Galaxy in a production environment. *Galaxy-Wiki* [online]. [b. r.], last modifed 2012-01-26 [cit. 2012-03-03]. Dostupné z: http://wiki.g2.bx.psu.edu/Admin/Config/Performance/Production%20Server

[11] The PostgreSQL Global Development Group *PostgreSQL 8.4.11 Documentation* [online]. ©1996-2009 [cit. 2012-03-03]. Dostupné z: http://www.postgresql.org/docs/8.4/interactive/index.html

[12] *OpenSSL: The Open Source toolkit for SSL/TLS: Welcome to the OpenSSL Project* [online]. The OpenSSL Project, ©1999-2009 [cit. 2012-03-04]. Dostupné z: http://www.openssl.org/

[13] Proftpd *A User's Guide* [online]. Mark Lowes, ©2001 [cit. 2012-03-04]. Dostupné z: http://proftpd.org/localsite/Userguide/linked/userguide.html

[14] ProFTPD mini-HOWTO - SQL and mod_sql *SQL and mod_sql* [online]. [b. r.], last modifed 2011-03-16 [cit. 2012-03-04]. Dostupné z: http://www.proftpd.org/docs/howto/SQL.html

[15] SQLBackend *SQLBackend – Set the SQL backend module* [online]. [b. r.][cit. 2012-03-04]. Dostupné z: http://www.proftpd.org/docs/directives/linked/config_ref_SQLBackend.html

[16] PostgreSQL: Documentation: Manuals: Database Roles and Privileges *Chapter 18. Database Roles and Privileges* [online]. PostgreSQL Global Development Group ©1996-2012 [cit. 2012-03-05].

Dostupné z: http://www.postgresql.org/docs/8.1/static/user-manag.html

[17] PostgreSQL: Documentation: Manuals: GRANT *GRANT* [online]. PostgreSQL Global Development Group ©1996-2012 [cit. 2012-03-05]. Dostupné z: http://www.postgresql.org/docs/8.4/interactive/sql-grant.html

[18] Admin/Config/Upload via FTP - Galaxy Wiki: Enabling upload to Galaxy via FTP. *Galaxy-Wiki* [online]. [b. r.], last modifed 2012-02-02 [cit. 2012-03-05]. Dostupné z: http://wiki.g2.bx.psu.edu/Admin/Config/Upload%20via%20FTP

[19] Admin/Tools/Tool Config Syntax - Galaxy Wiki: Galaxy Tool XML File. *Galaxy-Wiki* [online]. [b. r.], last modifed 2011-12-19 [cit. 2012-03-14]. Dostupné z: http://wiki.g2.bx.psu.edu/Admin/Tools/Tool%20Config%20Syntax

[20] PDF files | Adobe Portable Document Format - Acrobat: Adobe PDF history. ADOBE SYSTEMS INCORPORATED. *Adobe* [online]. Adobe Systems, ©2012 [cit. 2012-03-15]. Dostupné z: http://www.adobe.com/products/acrobat/adobepdf.html

[21] ReStructuredText Markup Specification. DAVID GOODGER. *Docutils: Documentation Utilities* [online]. 03.01.2012 [cit. 2012-03-15]. Dostupné z: http://docutils.sourceforge.net/docs/ref/rst/restructuredtext.html

[22] Nový nástroj hackerů snadno odstaví výkonné servery i z jediného notebooku. *Security World* [online]. IDG Czech Republic, 26.10.2011 [cit. 2012-03-21]. Dostupné z: http://securityworld.cz/securityworld/novy-nastroj-hackeru-snadno-odstavi-vykonne-servery-i-z-jedineho-notebooku--3842

[23] ProFTPD + MySQL + Quota + šifrování. In: *AbcLinuxu.cz - Linux na stříbrném podnose* [online]. AbcLinuxu.cz, 3.7.2009 [cit. 2012-03-21]. Dostupné z: http://www.abclinuxu.cz/clanky/site/proftpd-mysql-quota-sifrovani#sifrujeme

[24] ProFTPD module mod_sql. *The ProFTPD Project* [online]. The ProFTPD Project, 2000-2012, 26.1.2012 [cit. 2012-03-21]. Dostupné z: http://www.proftpd.org/docs/contrib/mod_sql.html#SQLDefaultGID

Přílohy

Seznam příloh

A Dokumentace vytvořená pro nastavení Galaxy serveru - přiložena na CD

B sampleFasta.xml - přiloženo na CD

C seqclust.xml - přiloženo na CD

D extractfasta.xml - přiloženo na CD

E fastaload.xml - přiloženo na CD

F maketree.xml - přiloženo na CD