VILNIAUS UNIVERSITETAS

MATEMATIKOS IR INFORMATIKOS FAKULTETAS
MATEMATINĖS INFORMATIKOS KATEDRA

Mindaugas Laganeckas

Pagrindinių studijų programa Bioinformatika – 4 kursas

PRAKTIKOS ATASKAITA

Praktika atlikta: Biotechnologijos institutas
(organizacijos pavadinimas)

Organizacijos praktikos vadovas: vyr. m. d. Mindaugas Margelevičius
(pareigos, vardas, pavardė)

Organizacijos praktikos vadovo įvertinimas: _____
(įvertinimas, parašas)

Universiteto praktikos vadovas: prof. Gediminas Stepanauskas
(mokslo laipsnis, vardas, pavardė)

_____
(parašas)

Ataskaitos įteikimo data _____
Registracijos Nr. _____
Įvertinimas _____
(data, įvertinimas, parašas)

Vilnius, 2009

Praktikos vietos aprašymas

Praktika buvo atliekama Biotechnologijos instituto Bioinformatikos laboratorijoje. Biotechnologijos institutas yra biudžetinė mokslinių tyrimų institucija, siekianti užtikrinti valstybės pažangą sparčiai besivystančiose gyvybės mokslų ir biotechnologijų srityse, plėtoti tarptautinio lygio molekulinės biotechnologijos tyrimus, skatinti tarpdisciplininius tyrimus bei mokslo ir verslo bendradarbiavimą.

Bioinformatikos laboratorija yra viena iš šešių institute veikiančių laboratorijų. Joje daugiausiai dirbama su baltymų amino rūgščių sekomis ir jų erdvinėmis struktūromis. Pagrindinės mokslinių tyrimų kryptys yra dvi: naujų metodų kūrimas ir jau esamų taikymas konkrečioms biologinėms problemoms spręsti.

Laboratorijoje darbo sąlygos buvo puikios: suteikta individuali darbo vieta, galimybė daryti trumpas pertraukėles darbo metu, lankstus darbo dienos grafikas (Biotechnologijos institutas įsikūręs Vilniaus miesto pakraštyje, todėl nebuvo reikalaujama pradėti/baigti darbą tiksliai nustatytą valandą). Visais iškilusiais klausimais mielai konsultavo tiek darbo, tiek laboratorijos vadovas bei kiti bendradarbiai.

Įvadas

Praktikos darbo tema - „Interaktyvus internetinis serveris baltymų homologijos paieškoms ir erdvinės struktūros modeliavimui“. Bendras darbo tikslas – sukurti per interneto naršyklę prieinamą ir draugišką vartotojui baltymų tarpusavio giminystės ryšių (homologijos) paieškos ir erdvinės struktūros modeliavimo serverį.

Šiuo metu Bioinformatikos laboratorijoje baigiamas testuoti naujai sukurtas jautrus homologijos paieškų metodas. Šis metodas remiasi baltymų sekų profilių lyginimu ir jų panašumo statistinio reikšmingumo vertinimu. Natūralus tolesnis žingsnis yra padaryti šį metodą prieinamą kiek galima platesniam Lietuvos ir kitų šalių mokslininkų ratui, tokiu būdu prisidedant prie aktualių biomedicinos problemų sprendimo. Efektyviausias būdas tai padaryti yra sukurti internetinį serverį, prieinamą per interneto naršyklę ir pasižymintį paprasto naudojimo vartotojo sąsaja. Tokiu atveju naujuoju homologijos paieškų metodu galėtų lengvai naudotis ne tik bioinformatikai, bet ir įvairiausius biologinius tyrimus atliekantys mokslininkai be specialaus kompiuterinio išsilavinimo.

Internetiniam serveriui buvo keliami šie reikalavimai:

1) Lankstus duomenų įvedimas bei parametrų nustatymas.
2) Galimybė pasirinkti norimas duomenų bazes, kuriose būtų vykdomos paieškos.
3) Suprantamas ir informatyvus rezultatų pateikimas.
4) Galimybė panaudoti paieškos rezultatus erdvinės struktūros modelių sudarymui.

Per tris praktikos mėnesius pavyko įgyvendinti pirmus du reikalavimų punktus iš keturių.

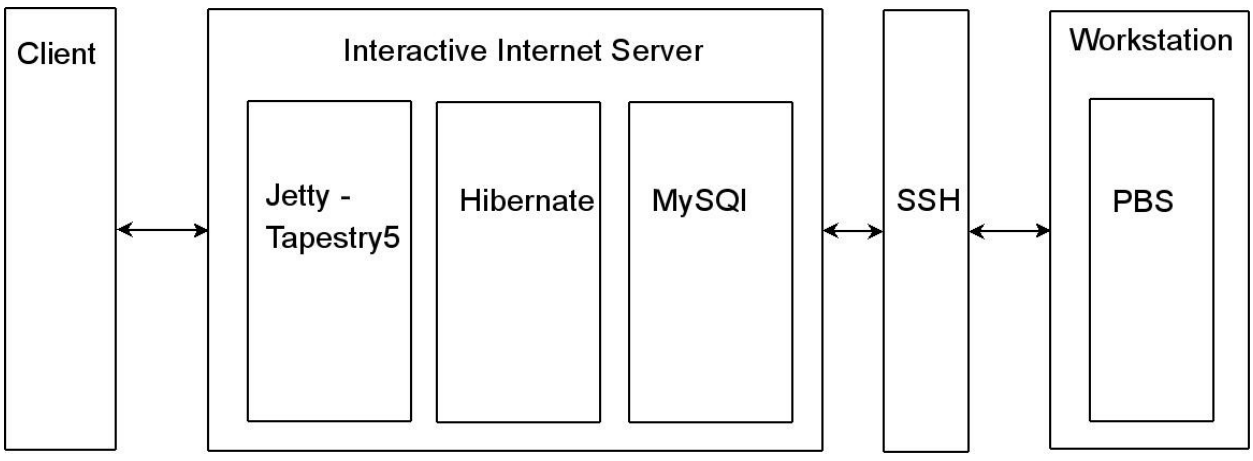
Teorinis–metodinis skyrius

Siekiama sukurti naują internetinį baltymų homologijų paieškos ir struktūros modeliavimo serverį, kurio pagrindas – Bioinformatikos laboratorijoje sukurtas sekų profilių paieškos ir lyginimo metodas. Svarbiausi numatomi serverio bruožai būtų šie:

- Lankstus duomenų įvedimas bei parametrų nustatymas. Paieškai galima būtų įvesti arba tik tiriamo baltymo seką, arba jau sudarytą daugybinį palyginį. Individualios sekos įvedimo atveju, daugybinis palyginys būtų sudaromas naudojant PSI-BLAST. Palyginio įvedimo atveju vartotojas galėtų pasirinkti, ar jį naudoti tiesiogiai, ar taip pat praturtinti vykdant PSI-BLAST paieškas. Po to visais atvejais daugybiniai palyginiai būtų automatiškai transformuojami į sekų „profilius“ su kuriais ir būtų vykdomos paieškos. Tiek palyginio sudarymui, tiek pačiai paieškai vartotojui būtų suteikta galimybė keisti pagrindinius parametrus.
- Galimybė pasirinkti norimas duomenų bazes, kuriose būtų vykdomos paieškos. Tarp jų būtų svarbiausios baltymų struktūrų, baltymų šeimų ir konservatyvių baltymų domenų pagrindu sudarytos profilių duomenų bazės. Iš baltymų struktūrų duomenų bazių būtų įtraukta PDB (http://www.pdb.org) ir SCOP (http://scop.mrc-lmb.cam.ac.uk/scop/). Baltymų šeimų duomenų bazę atsovautų PFAM (http://pfam.sanger.ac.uk/), o konservatyvių domenų - CDD duomenų bazė (http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml). Galimas ir kitų duomenų bazių įtraukimas.
- Suprantamas ir informatyvus rezultatų pateikimas. Rezultatus numatoma pateikti stiliumi, panašiu į NCBI PSI-BLAST serverio rezultatų išvedimo formą. Kitaip tariant, apibendrintus rezultatus numatoma pateikti grafiškai, o detales - teksto forma. Tekstą sudarytų surastų profilių sąrašas su statistinio reikšmingumo įverčiais bei atitinkami surastų profilių ir užklausos sekos palyginiai. Kiekvienas iš pateiktų palyginių priklausomai nuo to, kurioje profilių duomenų bazėje vykdoma paieška, būtų papildomai anotuotas nuorodomis į tėvinę duomenų bazę bei kitus išorinius informacijos šaltinius. Pavyzdžiui, jei paieška vyko profilių duomenų bazėje, sudarytoje sekoms su žinomomis erdvinėmis struktūromis, tai kiekvienas palyginys turėtų nuorodas į atitinkamus PDB ir SCOP įrašus, kurie leistų iš karto susipažinti su surasto baltymo struktūra ir jos vieta klasifikacijos sistemoje. Be to, kiekvieno surasto baltymo anotacija būtų panaudota bibliografijos paieškai vienoje iš atvirai prieinamų mokslinės literatūros duomenų bazių, tokių, kaip, pavyzdžiui, PubMed (http://www.pubmed.gov/).

- Galimybė panaudoti paieškos rezultatus erdvinės struktūros modelių sudarymui. Tais atvejais, kai profilių duomenų bazės sudarytos PDB ar SCOP baltymų sekų pagrindu, galima panaudoti rezultatus tiriamo baltymo erdvinės struktūros modeliui (-iams) sudaryti. Šiuo atveju surastos giminingos struktūros būtų panaudotos kaip šablonai tiriamos sekos modeliui sudaryti. Savaime suprantama, būtų sudaryta galimybė vartotojui peržiūrėti gautus modelius viena iš populiarių struktūrų vizualizavimo programų, o taip pat išsisaugoti modelio atomų koordinates lokaliai. Toks papildomas rezultatų apdorojimas leistų net menkai apie modeliavimą išmanančiam vartotojui labai paprastai pereiti nuo sekų prie trimačių struktūrų informacijos.

Serveriui vieni iš svarbiausių keliamų reikalavimų buvo jo saugumas bei pernešamumas, todėl jo kūrimui pasirinkta Java programavimo kalba, naudojant Java EE 5 technologiją. Serverio privatiems duomenims, tokiems kaip skaičiavimų ar pašto serverio ir pan., saugoti naudojama nemokama MySQL duomenų bazių valdymo sistema [5]. Norint pasiimti duomenis iš duomenų bazės Java aplinkoje naudojama Hibernate technologija, atvaizduojanti duomenų bazės esybes į Java kalbos objektus. Tokiu būdu galima atsiriboti nuo konkrečios duomenų bazių valdymo sistemos ir reikalui esant greitai pereiti prie kitos [4]. Interneto serverio variklis, „besiklausantis“ vartotojų užklausų, paremtas Jetty technologija [3]. Internetinių puslapių turinio dinaminiam generavimui pasirinkta Tapestry 5 technologija, pasižyminti tuo, kad verslo logika (Java klasės) atskiriamos nuo produkto išvaizdos (html failų) tokiu būdu padarydama projekto abi dalis nepriklausomas viena nuo kitos [2]. Tam, kad būtų užtikrintas saugus internetinio bei skaičiavimų serverio bendravimas, naudojamos SSH priemonės [6]. Taip pat sistemoje buvo suprojektuotas ir priimtas lygiagrečių skaičiavimų ir užduočių valdymo sprendimas PBS (angl. Portable Batch System) [7], skirtas paspartinti profilių sudarymo skaičiavimus, profilių duomenų bazių sudarymą bei atnaujinimą (žr. Pav. 1):

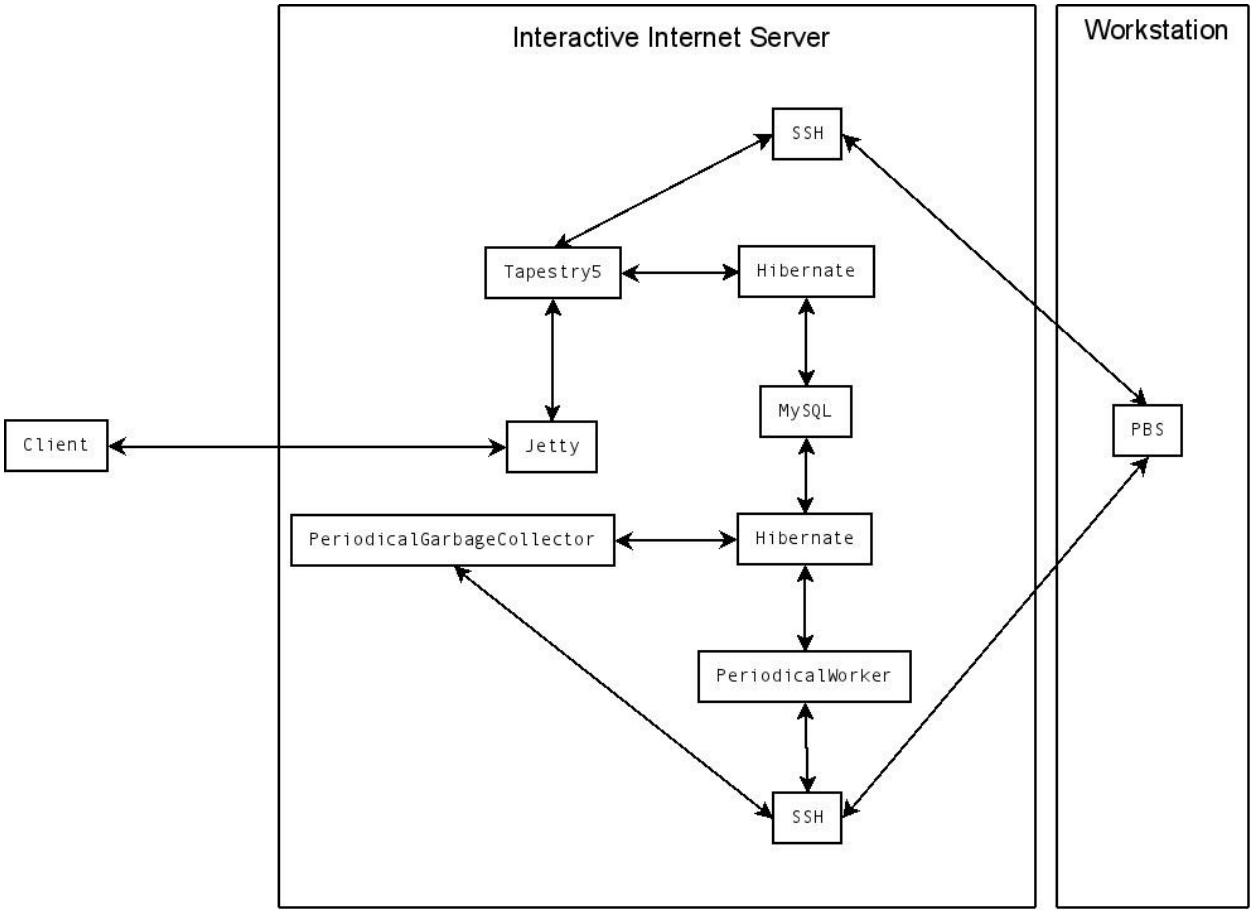


Pav. 1

Tiriamasis–analitinis skyrius

1. Sistemą sudarantys komponentai

Naudojant Teoriniame–metodiniame skyriuje aprašytas pasirinktas technologijas pasirinkta tokia sistemos komponentų architektūra (žr. Pav. 2):



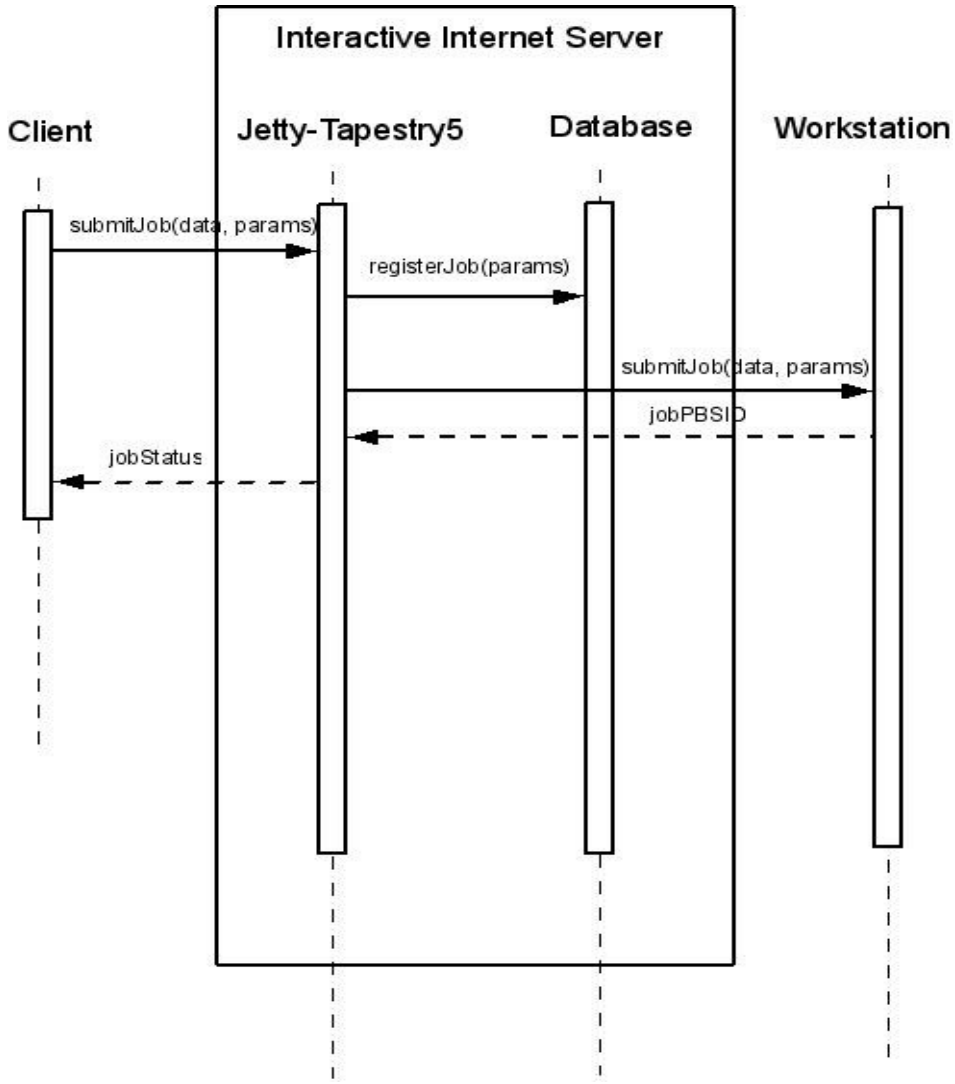
Pav.2

Kliento užklausų „klausosi“ Jetty serveris ir perduoda jas už internetinių puslapių dinaminį generavimą atsakingam Tapestry5. Šis savo ruožtu, priklausomai nuo užklausos turinio per SSH susisiekia su darbo stotimi (Workstation) arba per Hibernate su duomenų baze (MySQL) ir suformuoja užklausos rezultatus, kurie vėlgi per Jetty serverį grąžinami klientui. PeriodicalWorker atsakingas už periodinį vykdomų darbų patikrinimą, ar jie padaryti ir, jei padaryti, atitinkamų rezultatų suformavimą ir įrašymą į duomenų bazę. SSH atsakingas už duomenų perdavimą tarp interaktyvaus serverio bei darbo stoties ir atvirkščiai. PBS – savo ruožtu – už lygiagretų užduočių paskirstymą darbo stotyje. PeriodicalGarbageCollector atsakingas už darbų, kurių galiojimo laikas baigėsi, sunaikinimą. Taip pat ir sunaikinimą su jais susijusių duomenų, parametrų bei rezultatų. Toliau bus detaliau aptarti sistemos veikimo scenarijai.

2. Sistemos veikimo scenarijai

2.1 Darbo „užsakymas“

Klientas „užsako“ darbą pateikdamas užduočiai reikalingus duomenis bei parametrus (žr. Pav. 3):

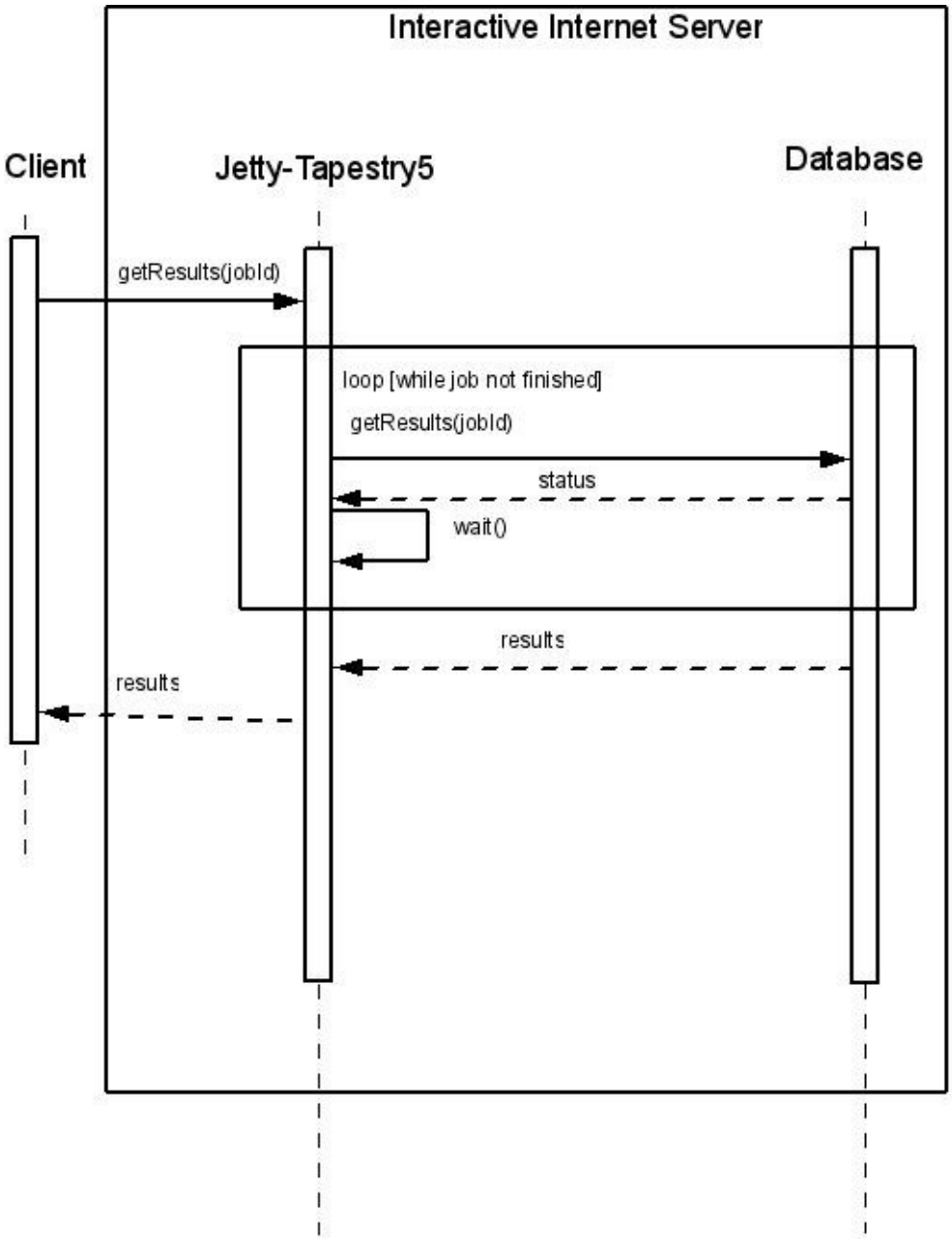


Pav. 3

1. Klientas duomenis ir rezultatus perduoda interaktyviam internetiniam serveriui.
2. Suformuota užduotis užregistruojama duomenų bazėje.
3. Užregistruota užduotis nusiunčiama į skaičiavimų serverį (Workstation).
4. Skaičiavimų serveris grąžina užduoties ID.
5. Klientui perduodama užduoties būsena.

2.2 Rezultatų „laukimas“

Klientas rezultatų gali laukti iš karto, kai „užsako“ darbą arba vėliau, nurodydamas norimos užduoties ID. Abu atvejai analogiški, tik pirmu atveju sistema žino užduoties ID pati ir jai jo perduoti nebereikia (žr. Pav. 4):

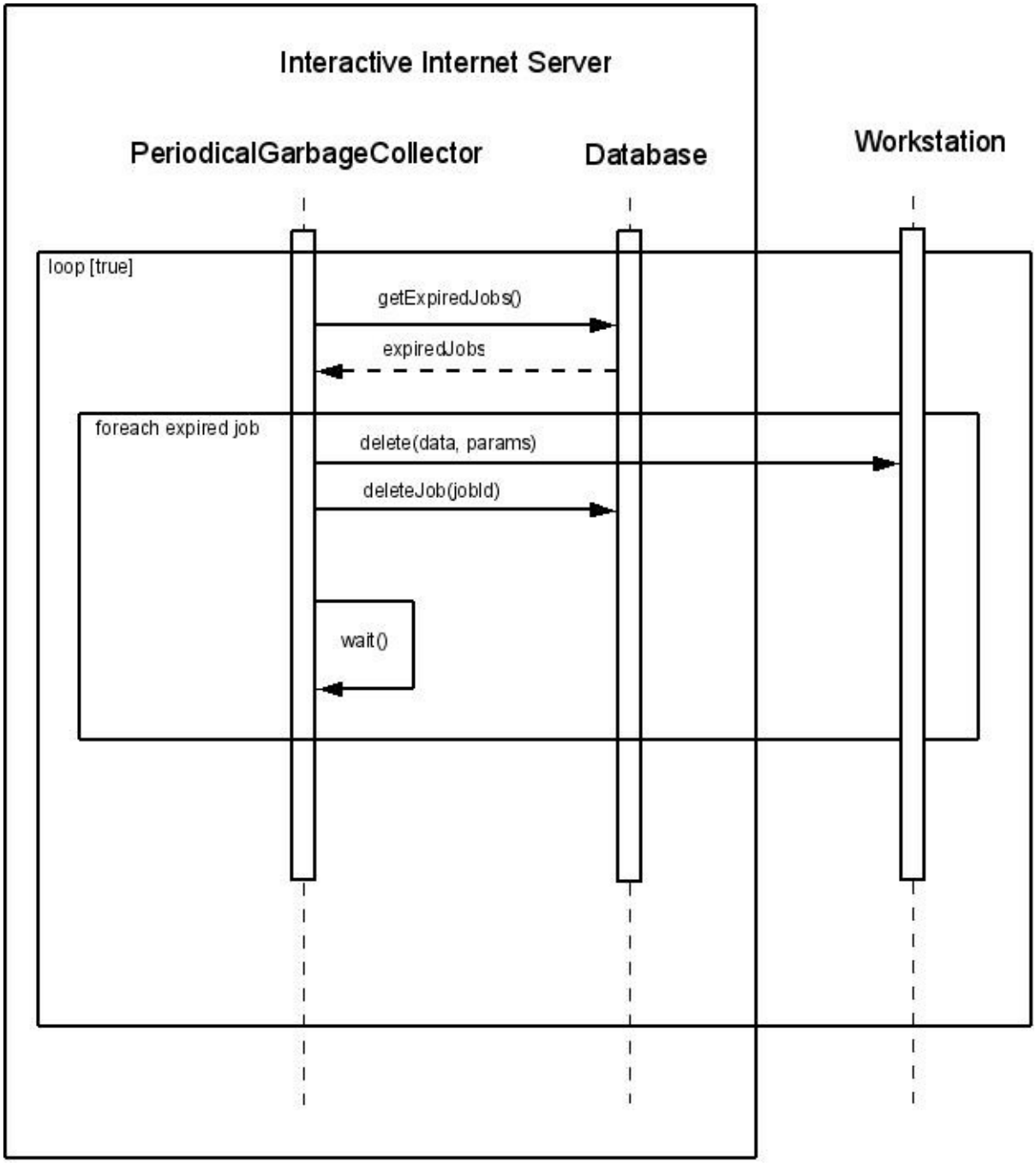


Pav. 4

1. Sistemai perduodamas užduoties ID (arba iš karto po darbo “užsakymo ji jį žino”).
2. Duomenų bazėje tikrinama užduoties būsena tol, kol būsena nelygi “Baigta”.
3. Kaskart gavus neigiamą atsakymą, padaroma pertrauka, per kurią nedaroma nieko.
4. Kai būsena tampa lygi “Baigta”, rezultatai perduodami vartotojui.

2.3 Periodinis užduočių naikinimas

Pasibaigus užduoties galiojimo laikui ji ir visi su ja susiję duomenys turi būti sunaikinti (žr. Pav. 5):



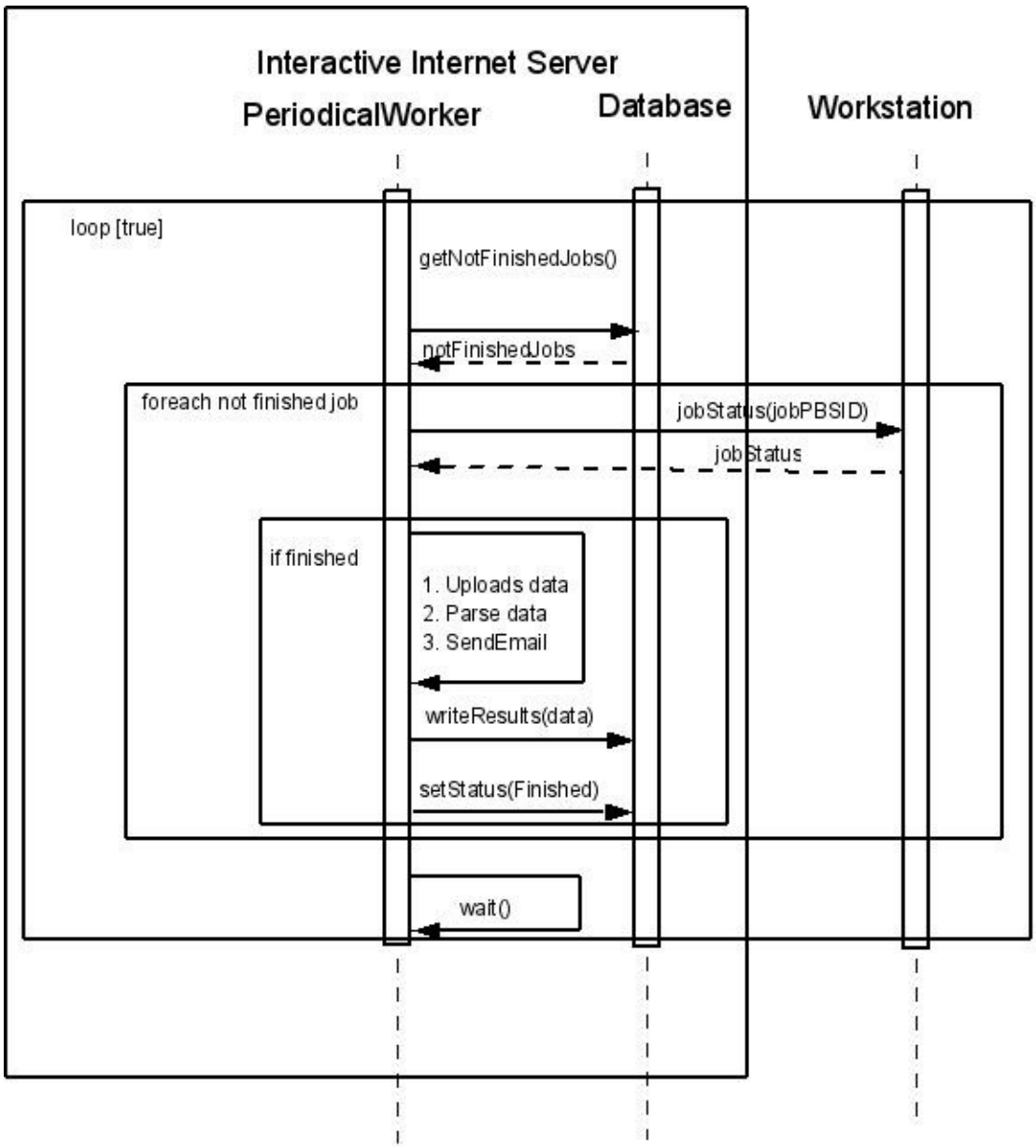
Pav. 5

1. Periodinis darbų naikinimas – nesibaigiantis darbas.
2. Iš duomenų bazės gaunamos visos užduotys, kurių galiojimo laikas baigėsi.
3. Kiekvienos užduoties duomenys bei parametrai sunaikinami darbo stotyje.
4. Kiekvienos užduoties parametrai bei rezultatai sunaikinami duomenų bazėje.

5. Sistema laukia nustatytą laiko tarpą. Po to viskas kartojama iš naujo.

2.4 Periodinis darbų vykdymas

Reikia nuolat tikrinti, ar nebaigtas užduotis skaičiavimų serveris jau įvykdė ir, jei taip, paruošti rezultatus vartotojui (žr. Pav. 6):



Pav. 6

1. Periodinis darbų vykdymas – nesibaigiantis darbas.
2. Iš duomenų bazės gaunami visos užduotys, kurių būsena nėra „Baigta“.
3. Kiekvienai nebaigtai užduočiai patikrinama skaičiavimų serveryje, ar ji jau baigta.
4. Kiekvienos baigtos užduoties parsiunčiamas rezultatų failas ir paruošiamas vartotojui.
5. Jei vartotojas nurodė el. pašto adresą, jam išsiunčiamas laiškas, kad užduotis baigta.
6. Rezultatai surašomi į duomenų bazę.
7. Užduoties būsena nustatoma „Baigta“.

Išvadų ir pasiūlymų skyrius

Tikslai, iškelti praktikos darbe pasiekti ir Biotechnologijos instituto Bioinformatikos laboratorijoje interaktyvaus serverio sistema jau funkcionuoja. Pasirinkta sistemos architektūra pasiteisino: sistema yra saugi, pernešama bei integrali. Sistemos testavimo metu visos pastebėtos klaidos buvo pašalintos. Praktikai iškelti reikalavimai buvo įgyvendinti, tačiau visų sistemai keliamų reikalavimų įgyvendinimas užima daugiau laiko nei numatyta praktikoje, todėl sistemos vystymo darbai bus tęsiami toliau.

Literatūros sąrašas

1. http://tapestry.apache.org/tapestry5/
2. http://docs.codehaus.org/display/JETTY/Jetty+Wiki
3. http://en.wikipedia.org/wiki/Hibernate_(Java)
4. http://lt.wikipedia.org/wiki/MySQL
5. http://en.wikipedia.org/wiki/Secure_Shell
6. http://en.wikipedia.org/wiki/Portable_Batch_System