VILNIAUS UNIVERSITETAS
MATEMATIKOS IR INFORMATIKOS FAKULTETAS
PROGRAMŲ SISTEMŲ KATEDRA

Bioinformatikos

Trečiasis laboratorinis darbas

pirma dalis

Atliko: 4 kurso 5 grupės studentas
Aurelijus Banelis

Vilnius – 2012

Turinys

1. Paleidimas ir nustatymai

Programa parašyta Pitono programavimo kalba. Programa paleidžiama:

```
python U3-Aurelijus-banelis.py
```

Linux (Ubuntu) aplinkoje programa turėtų veikti be papildomo koregavimo. Pagrindiniai nustatymai yra prie:

```
class HpvAnalyser:
    # Configuration
```

Naudojamos bibliotekos ir išorinės programos:

- Biopython
- CD-Hit

2. Programos ir failų struktūra

Pagrindinė programos dalis yra sudėta į klasę `HpvAnalyser`, viešus metodus kviečiant iš programos kodo. Privatūs metodai prasideda „_“. Kiekvienas viešas metodas turi trumpą aprašymą.

Rezultatus programa saugoja tame pačiame kataloge. Failų vardai prasideda „HPV“ ir toliau eina viruso tipo numeris. CD-HIT rezultatai turi priesagą „-cdhit“. Sujungtas failas yra „HPV-all.fa“. Sulygiuotas failas yra „HPV-aligned.fa“.

`U3-Aurelijus-banelis.py` yra autonomiška, t. y. visi HPV*.fa failai gali būti ištinti, nes jie yra sugeneruojami dinamiškai pačios programos. HPV*.fa failai yra pateikti, kaip programos veikimo rezultatų pavyzdys.

3. Projektiniai sprendimai ir algoritmai

3.1. Fragmentų sąrašo sudarymas

3.1.1. Įrankių pasirinkimas

Nors užduotyje buvo pasiūlyta „seką galite naudoti per blast užklausas rinkdami sekas analizei“, patogiau buvo tiesiogiai kreiptis į duomenų bazes, nes:

1. Naudojant Fasta fragmentą Blast užklausose, nei su viena duomenų baze nebuvo gaunami visi HPV tipai
2. Pati užklausa užtrukdavo ilgai ir naudotojui nebuvo įmanoma pranešti apie progresą
3. GenBank formatas leidžia lengvai nustatyti L1 geno regioną ir jį išskirti

3.1.2. Pradinė analizė

Kadangi Blast neduodavo visų tipų, todėl buvo nuspręsta, tiesiog sugeneruoti užklausą pagal reikiamus HPV tipus (realizuota `generateEntrezQuery(self)`). Pagrindinė problema buvo, kad nepavykdavo rasti *Entrez query* parametrų, kurie ir tiktų visiems HPV tipams, ir sumažintų pasikartojimų skaičių. Kita problema buvo, kad duomenų bazė pateikdavo ir labai trumpas segmentų dalis. Mažesnės dažniausiai dalys būdavo didesnių poaibis, todėl tik padidindavo rezultatų kiekį ir jų atsiuntimą. Kaip optimalus parametras buvo pasirinktas palyginio ilgis nuo 1000 iki 8000 (51 tipas įtakojo apatinį rėžį, kitų tipų palyginiai buvo apie 7000 ilgio). Taip pat, kaip paieškos parametras naudojamas geno regiono pavadinimas: „L1“. Dauguma palyginių turėjo ir „L1“ kaip baltymo pavadinimą. Užklausos variantai buvo tikrinami naudojant internetinę paieškos versiją (http://www.ncbi.nlm.nih.gov), tol, kol užklauso tiko visiems HPV tipams. Parametrų derinimo kiekvienam tipui individualiai buvo atsisakyta, nes kitaip pati programa neturėtų jokios pridėtinės vertės.

Galutinė užklausa atrodo:

```
("Human papillomavirus type 16"[Organism] OR "Human papillomavirus type 18"[Organism] OR "Human papillomavirus type 31"[Organism] OR "Human papillomavirus type 33"[Organism] OR "Human papillomavirus type 35"[Organism] OR "Human papillomavirus type 51"[Organism] OR "Human papillomavirus type 52"[Organism] OR "Human papillomavirus type 6"[Organism] OR "Human papillomavirus type 11"[Organism] OR "Human papillomavirus type 40"[Organism] OR "Human papillomavirus type 42"[Organism] OR "Human papillomavirus type 43"[Organism] OR "Human papillomavirus type 44"[Organism] OR "Human papillomavirus type 57"[Organism] OR "Human papillomavirus type 81"[Organism]) AND L1[Gene Name] AND 1000:8000[Sequence Length]
```

3.1.3. Duomenų bazės pasirinkimas

Duomenų bazės pasirinkimui daugiausia įtakos turėjo įrašų visuose duomenų bazėse paieška[1] . Buvo pirmiausia patikrintos daugiausia rezultatų siūlančios duomenų bazės. *Nucleotide* duomenų bazės buvo pasirinkta, nes ji duoda rezultatus visiems tipams, bei ją naudojant galima gauti užduotyje pateiktą fasta failą[2].

Įrašai iš duomenų bazės paimami naudojant `retrieveData(self)` funkciją.

1 Bendros paieškos nuoroda: http://www.ncbi.nlm.nih.gov/gquery

2 Pavyzdinis Fasta failas: http://www.ncbi.nlm.nih.gov/nuccore/333031?from=5559&to=7154&report=fasta

3.1.4. Duomenų vaizdavimo pasirinkimas

Pagrindinis GenBank formato privalumas – L1 regiono pradžios ir pabaigos nurodymas; pagrindinis trūkumas: rezultatus galima gauti tik tekstinių formatu (kitaip nei xml). Kadangi GenBank turi visą informaciją kaip ir Fasta formatas, todėl buvo nuspręsta pasirašyti GenBank failo konvertavimą į Fasta formą. *Biopython SeqIO* klasė pateikia funkcijas tik darbui su failais, todėl papildomas GenBank duomenų išsaugojimas į failą tik palėtintų programos veikimą.

Konvertavimas realzuotas `_saveToFasta(self, genBankText, id)` funkcija.

3.1.5. Duomenų parsisiuntimas

Duomenys persisiunčiami 2 etapais: užklausos rezultatų Id numerių gavimu (`retrieveIds`) ir pačių sekų GenBank formatu atsisiuntimu (`retrieveData`). Kaip siūloma Biopython dokumentacijoje[3], atsisiuntimas atliekamas naudojant `webenv` ir `queryKey` parametrus, vietoj Id numerių pateikimo tiesiogiai. Duomenų bazės resursų optimizavimui taip pat naudojamas atsisiuntimas dalimis (`batchSize`). Į failų sistemą išsaugomi tik sukonvertuoti į Fasta formatu failai.

Kadangi programa gali būti paleista kelis kartus, programos veikimo pradžioje visų fasta failų turinys yra ištrinamas (`_emptyCachedFiles`). Kadangi vienam HPV tipui gali būti keli palyginiai, todėl veikimo metu fasta failų turinys yra pridedamas (`open(self._fastaName(type), 'a')`).

3.2. CD-HIT

Programoje naudojama vietinė (įdiegta į kompiuterį) CD-HIT versija. Kurta Ubuntu (Linux) aplinkoje. Reikėtų pakoreguoti:

```
cdhit = '/usr/bin/cdhit'
```

pagal naudojamos sistemos Cd-hit programos vietą.

3.3. Mafft

Programoje naudojama vietinė (įdiegta į kompiuterį) Mafft versija. Kūrimo metu, nenurodžius `MAFFT_BINARIES` kintamojo, mafft programa rašydavo, kad ji yra blogai įdiegta. Reikėtų pakoreguoti:

```
mafft_exe = "/usr/bin/mafft --localpair --maxiterate 1000"
mafft_lib = "/usr/lib/mafft/lib/mafft"
```

3 BioPython dokumentacija: http://biopython.org/DIST/docs/tutorial/Tutorial.html

pagal naudojamos sistemos mafft programos vietą. Mafft parametrai yra parinkti, kad būtų kuo didesnis sulyginimo tikslumas.

3.4. Naudotojo sąsaja

Tarpiniams veiksmams/būsenai nurodyti naudojam *logging* biblioteka. Jei nenorite, kad būtų rodomi tarpiniai pranešimai, pakeiskite `level` reikšmę:

```
logging.basicConfig(format='%(levelname)s:%(message)s',
level=logging.WARNING)
```

Paskutinė programos žinutė turėtų būti:

```
INFO:Finished: 134.0 seconds
```

Kurioje nurodomas programos vykdymo laikas. Taip pat turėtų būti sugeneruotas sulygintų sekų failas: **HPV-all.fa**. Taip pat tame pačiame kataloge bus palikti ir tapriniai HPV*.fa failai.