OrthoMCL - instrukcja obsługi

Małgorzata Wańczyk

1 Opis programu

Program **OrthoMcl** identyfikuje grupy ortologów dla dwóch lub więcej organizmów na podstawie ich sekwencji aminokwasowych. Dodatkowo rozróżnia tzw. in-paralogi od ortologów. W programie można wyróżnić następujące fazy:
-filtrowanie sekwencji (filtrowanie);
-porównanie BLAST'em każdej sekwencji z każdą;
-obliczanie procentów dopasowanej długości dla każdego najlepszego dopasowania;
-szukanie potencjalnych ortologów i in-paralogów;
-użycie programu Markov Cluster algorithm (MCL).

Do poprawnego działania programu wymagane są:
- system **UNIX**;
- **mysql-server-5.1** lub nowszy (można zainstalować z lini komend *sudo aptitude install mysql-server-5.1*);
- **mcl** (można zainstalować z lini komend *sudo aptitude install mcl*);
- standardowy **PERL** i biblioteka **DBI**;
- program **BLAST**.

2 Instalacja

Najpierw ściągamy spakowany katalog z programem przy użyciu komendy:

wget http://orthomcl.org/common/downloads/software/v2.0/orthomclSoftware-v2.0.3.tar.gz

Następnie rozpakowujemy:

tar -zxf orthomclSoftware-v2.0.3.tar.gz

W rozpakowanym katalogu **orthomclSoftware-v2.0.3** znajdują się pliki binarne programu oraz pliki konfiguracyjne. Do poprawnego działania programu trzeba stworzyć bazę danych w **mysql**. Włączamy program
mysql -u root -p
wpisujemy to samo hasło co przy instalacji **mysql'a**.
Tworzymy nowego użytkownika, bazę danych i nadajemy konkretne uprawnienia:

mysql> CREATE DATABASE orthomcl;

mysql> GRANT SELECT,INSERT,UPDATE,DELETE,CREATE VIEW,CREATE, INDEX, DROP on orthomcl. TO orthomcl@localhost identified by 'orthomcl1';*

mysql>exit;

Aby zalogować się później jako użytkownik wpisujemy:
mysql -u orthomcl -p (hasło: orthomcl1).

W katalogu **orthomclSoftware-v2.0.3/bin/**:
- tworzymy katalog **myorthomcl**, w którym będziemy przechowywać wyniki;
- z katalogu **orthomclSoftware-v2.0.3/doc/OrthoMCLEngine/Main/** przekopiowujemy do **myorthomcl** plik **orthomcl.config.template**;

- zmieniamy nazwę **orthomcl.config.template** z katalogu **myorthomcl** na **orthomcl.config** i w linijkach *dbLogin=* oraz *dbPassword=* dopisujemy *orthomcl*, *orthomcl1* odpowiednio. Z lini rozpoczynającej się *dbConnectString=* usuwamy *:3307*

3 Schemat

Wszystkie binarne pliki uruchamiamy z katalogu **orthomclSoftware-v2.0.3/bin/** komendą:
./nazwa_programu.
Aby wyświetlić pomoc/opis programu, uruchamiamy go bez argumentów.

1. Instalowanie schematu mysql.
   Program **orthomclInstallSchema** instaluje schemat orthomcl w wcześniej utworzonej bazie danych mysql. Argumentami programu są plik konfiguracyjny i plik do którego zostaną zapisane log'i.

   ./orthomclInstallSchema myorthomcl/orthomcl.config myorthomcl/install_schema.log

2. Dopasowanie plików Fasta.
   W katalogu **myorthomcl** tworzymy katalog **genomy**, w którym umieszczamy pliki z sekwencjami aminokwasowymi do analizy w formacie Fasta. Program **orthomclAdjustFasta** dopasowuje plik Fasta do dalszych etapów. Argumentami programu są kolejno:
   - trzy- lub czteroliterowy kod taksonu dla rozróżnienia organizmu;
   - plik fasta;
   - numer identyfikacyjny sekwencji (numer pola w opisie sekwencji, która odpowiada tej szczególnej sekwencji w genomie; pola rozdzielane są spacją lub znakiem '|').
   Przykładowo dla sekwencji rozpoczynającej się:

   >gi|387825663|ref|YP_005805116.1| Glycosyl ... Borrelia burgdorferi JD1

   możemy uruchomić program w następujący sposób:

   ./orthomclAdjustFasta BORJ myorthomcl/genomy/NC_017403.faa 4

   Procedurę powtarzamy dla wszystkich badanych genomów. Wynikiem są pliki Fasta z prostym opisem identyfikującym sekwencje w obrębie organizmu. Nowopowstałe pliki przenosimy do katalogu **myorthomcl/genomy/nowe/**.

3. Filtrowanie plików Fasta.
   Program **orthomclFilterFasta** filtruje pliki w odniesieniu do argumentów- minimalna długość białka (sugerowane 10), maksymalny procent kodonów STOP (sugerowane 20), oraz skleja wszystkie pliki w jeden **goodProteins.fasta**.

   ./orthomclFilterFasta myorthomcl/genomy/nowe/ 10 20

4. BLAST (każda z każdą).
   Uruchamiamy program BLAST dla pliku **goodProteins.fasta** z opcją zapisu tabelarycznego. Należy pamiętać o utworzeniu bazy danych dla programu BLAST. Domyślnie można ustawić evalue: 1e-5, powinno się ustawić SEG i maskowanie. Aby utworzyć bazę danych należy wpisać:

makeblastdb -in goodProteins.fasta -dbtype prot

5. Program **orthomclBlastParser** przerabia formę tabularyczną z BLAST'a tak, aby mógł być załadowany do bazy danych *mysql* oraz oblicza procent dopasowania wszystkich tzw. hit'ów. Wywołujemy przykładowo:
   ./orthomclBlastParser wynik_z_Blast podobneSekwenjce.txt .

6. Loadowanie wyników BLAST do bazy:
   ./orthomclLoadBlast myorthomcl/orthomcl.config podobneSekwencje.txt

7. Krok, w którym znajdowane są pary białek. Dodatkowe opcje to:
   cleanup= yes- wyczyścić wszystkie tabele;
   no- nie czyścić wszystkich tabeli;
   only- tylko wyczyścić tabele z końcowymi wynikami;
   all- tylko wyczyścic tabele końcowe z wynikami + czyszczenie tabeli Inparalogów, Ortologów i Koortologów.
   oraz *startAfter* =*TAG*, która pozwala się podnieść w miejscu, gdzie program skończył czy z jakichkolwiek powodów się wyłączył.
   ./orthomlcPairs myorthomcl/orthomcl.config myorthomcl/orthomcl_pairs.log cleanup=yes

8. Wyszukiwanie potencjalnych ortologów:
   ./orthomlcDumpPairsFile myorthomcl/orthomcl.config

9. Klastrowanie programem Markov Cluster Algorithm:
   mcl mclInput –abc -I 1.5 -o mclOutput
   gdzie *I* to parametr inflacji.

10. Tworzenie grup orthomcl z pliku wyjściowego z klastrowania, generowanie ID dla każdej grupy:
    ./orthoMCLToGroups prefix(np.OG) id(np.1000) MCLOutput >groups.txt.