Bioinformatyka: Analizy genomiczne (Borsnip)

Krótki opis usługi

Usługa pozwala wyszukiwać asocjacje w danych GWAS, wykorzystując metodę selekcji zmiennych Boruta i szybki, stochastyczny klasyfikator, paprocie losowe. Formalnie program duplikuje funkcjonalność kanonicznych implementacji tych narzędzi, pozwala jednak analizować dane GWAS szybciej i z użyciem mniejszych zasobów.

Aktywowanie usługi

Aby skorzystać z usługi należy mieć konto w projekcie PL-Grid (Zakładanie konta w portalu PL-Grid) wraz z aktywnym grantem (Granty obliczeniowe). Oprogramowanie składa się z kilku aplikacji konsolowych, można z nich korzystać zarówno bezpośrednio na systemie obliczeniowym (w trybie interaktywnym lub w skryptach) lub w skryptach uruchamianych za pośrednictwem systemu UNICORE.

Pierwsze kroki

Pierwszym krokiem jest konwersja danych GWAS na format *bsi* natywny i zrozumiały dla narzędzia borsnip; służą do tego programy pomocnicze `plink2bsi` i `gen2bsi` które akceptują binary plikozbiór narzędzia Plink (trójka `.bed` `.bim` i `.fam`) oraz format GEN (`.gen`). Dane rozbite na podzbiory (np. oddzielnie chromosomy albo oddzielnie obiekty z różnych grup) należy wcześniej połączyć korzystając z innego narzędzia.

Decyzję należy zachować oddzielnie, jako plik tekstowy z klasami zakodowanymi jako liczby od 0..(#klas-1), po jednej na linię. Połączenie z danymi genomicznymi odbywa się na zasadzie wspólnej kolejności. Jeśli plik/i/ danych zawiera już decyzję, zostanie ona zignorowana.

Przykładowo, zakładając że decyzja jest w pliku `input.dec`, jeśli mamy dane w formacie Plink jako pliki `input.bed`, `input.bim` i `input.bed`, a chcemy przepakować dane do pliku wejściowego borsnip `test.bsi`, należy wywołać:

```
plink2bsi input input.dec test.bsi
```

Jeśli mamy plik w formacie GEN, input.gen, należy wywołać:

```
gen2bsi input.gen input.dec test.bsi
```

W zasadniczym zadaniu należy zamieścić wywołanie głównego programu, borsnip.

```
borsnip -D 5 -N 100000 test.bsi
```

Parametr N (liczba paproci) powinien być nie mniejszy niż liczba analizowanych zmienności, podczas gdy parametr D (głębokość paproci) należy dopasować do złożoności problemu.

Wynikiem będzie tabela podobna do

```
rs999969     Confirmed        25         25
rs9999966    Rejected          0         11
rs9999662    Rejected          0         11
rs9999931    Confirmed        25         25
rs9999604    Rejected          0         11
```

Pierwsza kolumna to identyfikator zmienności, skopiowany z pliku wejściowego, druga to decyzja narzędzia (*Confirmed* dla istotnej asocjacji, *Rejected* dla nieistotnej, *Tentative* dla braku decyzji). Kolejne dwie kolumny to odpowiednio liczba zaliczonych i przeprowadzonych testów istotności dla zmienności.

Wywołanie borsnip z parametrem -h wyświetli listę dostępnych opcji. W szczególności, flagą -s można ustawić ziarno generatora pseudolosowego by zapewnić reprodukowalność obliczeń. Flaga -j przełącza output z tabeli na czytelny maszynowo format JSON, zawierający również dodatkowe informacje.

Gdzie szukać dalszych informacji?

Metoda Boruta jest opisana w artykułach http://www.jstatsoft.org/v36/i11/paper i http://www.biomedcentral.com/1471-2105/15/8 oraz na stronie metody https://m2.icm.edu.pl/boruta, zaś metoda paproci losowych w artykule http://www.jstatsoft.org/v61/i10/paper. Wszelkie pytania prosimy kierować do opiekuna aplikacji, na adres M.Kursa@icm.edu.pl.