Ghid de proceduri bioinformatice de analiză comparativă a amestecurilor bacteriene complexe folosind UniFrac

UniFrac

UniFrac este un instrument de măsurare a β-diversității care utilizează informații filogenetice pentru a compara probe de mediu. UniFrac, împreună cu tehnici statistice multivariate standard, inclusiv analiza coordonatelor principale (PCoA), identifică factorii care explică diferențele dintre comunitățile microbiene. UniFrac măsoară diferența dintre două colecții de secvențe (de exemplu, molecule de ARNr 16S secvențiate din probe microbiene diferite) exprimată drept cantitate de istorie evolutivă; aceasta este unică pentru oricare dintre cele două colecții și este măsurată ca fracțiunea lungimii ramurilor dintr-un arbore filogenetic care duce la descendenți ai unui eșantion sau ai altuia, dar nu ai ambelor.

Aplicația pe scară largă a UniFrac este facilitată de interfețe web ușor de utilizat, care gestionează seturi mari de date generate prin tehnologii de secvențiere de ultimă generație. UniFrac este, de asemenea, implementat în diverse pipeline-uri de analiză a secvenței comunității microbiene, precum QIIME și mothur. În plus, acesta este disponibil a fi utilizat direct din platforma Galaxy, fără a fi necesare cunoștințe prealabile de programare.

Distanța UniFrac este o distanță ponderată filogenetic între două comunități de organisme. Măsura aceasta a fost extinsă de mai multe ori pentru a include versiuni ponderate în funcție de abundență și ajustate în funcție de varianță.

Utilizarea UniFrac in R Studio

Pentru a calcula distanța UniFrac în R se folosește funcția UniFrac. Acesta este disponibilă în pachetul `phyloseq`.
Ceea ce face funcția UniFrac este să calculeze rapid distanțele UniFrac (Fast UniFrac) pentru toate perechile de probe dintr-un obiect de tip `phyloseq-class`.

Sintaxa funcției este următoarea:

```
UniFrac(physeq, weighted=FALSE, normalized=TRUE,
parallel=FALSE, fast=TRUE)
```

Argumentul `physeq` este obligatoriu, în timp ce argumentele `weighted`, `normalized`, `parallel` și `fast` sunt opționale. Cele 5 argumente se folosesc astfel:

- `physeq` – Este un argument obligatoriu, de tip `phyloseq-class`. Trebuie să conțină cel puțin un arbore filogenetic (`phylo-class`) și un tabel de contingență (`otu_table-class`).
- `weighted` – Este un argument opțional, de tip logic, care în mod implicit este setat pe valoarea FALSE. Opțiunea implicită FALSE presupune că pentru toate perechile de probe se determină distanța UniFrac neponderată. UniFrac ponderat (weighted) ia în considerare abundența relativă a speciilor/taxa împărțite între eșantioane, în timp ce UniFrac neponderat (unweighted) ia în considerare doar prezența/absența.

- `normalized` – Este un argument opțional, de tip logic, care în mod implicit este setat pe valoarea TRUE. Opțiunea implicită FALSE presupune că ieșirea este normalizată astfel încât valorile sale variază de la 0 la 1 independent de valorile lungimii ramurilor. Este important de reținut faptul că UniFrac neponderat este întotdeauna normalizat prin lungimea totală a ramurilor și, prin urmare, această valoare este ignorată atunci când `weighted` == FALSE.
- `parallel` – Este un argument opțional, de tip logic, care în mod implicit este setat pe valoarea FALSE. Opțiunea implicită FALSE presupune că UniFrac va înregistra un backend paralel, astfel încât commanda *`foreach::%dopar%`* să nu returneze un avertisment. Setarea argumentului *`parallel`* pe FALSE duce la executarea calculului în paralel, folosind simultan mai multe nuclee CPU. Acest lucru poate grăbi dramatic timpul de calcul pentru această funcție. Totuși, pentru a fi posibilă executarea paralelă, este necesar ca utilizatorul să fi înregistrat un „parralel backend" înainte de a apela această funcție.
- `fast` – Este un argument opțional, de tip logic, acceptă doar opțiunea TRUE. În prezent acest argument nu este folosit. El este implementat nativ în pachetul `phyloseq`.

Utilizarea UniFrac pe platforma Galaxy

Platforma Galaxy poate determina distanța UniFrac atât ponderat (weighted, – luând în considerare abundența relativă a speciilor), cât și neponderat (unweighted – luând în considerare doar prezența/absența speciilor).

Utilizatorul trebuie să furnizeze argumente pentru următoarele câmpuri, în funcție de tipul de analiză urmărită

- `tree` – arborele filogenetic
- `group` – fișierele grup pentru arborele filogenetic
- `groups` – Selectarea grupurilor pentru comparația perechilor de date
- `name` – Fișierul de nume pentru arborele filogenetic
- `iters` – Numărul de iterații (implicit 1000):
- `random` – Compararea propriilor arbori cu alți arbori generați aleator (nu se folosește dacă subsample=TRUE)
- `use subsampling of groups?:` (în locul analizei comparative aleatorii)
- `distance` – creearea unei matrici de distanțe
- `root` – întregul root pentru analiza de calcul
- count – a count_table: (generat de count.seqs)
- `Output logfile?:` – generarea unui fișier log

GUniFrac

GUniFrac este o suită de metode pentru analiza puternică și robustă a datelor legate de microbiom. Această analiză include normalizarea datelor, simularea datelor, testarea asocierii la nivel de comunitate și analiza abundenței diferențiale. GUniFrac implementează distanțe UniFrac generalizate, normalizarea mediei geometrice a raporturilor perechi (GMPR), simulator de date semiparametrice, metode statistice bazate pe distanță și metode statistice bazate pe caracteristici.

Metodele statistice bazate pe distanță includ trei extensii ale PERMANOVA:

1. PERMANOVA utilizând schema de permutare Freedman-Lane
2. testul omnibus PERMANOVA folosind mai multe matrice
3. abordare analitică pentru aproximarea valorii p PERMANOVA

Metodele statistice bazate pe caracteristici includ metode bazate pe modele liniare pentru analiza abundenței diferențiale a datelor compoziționale de tip zero-inflated highdimensional.

Utilizarea GUniFrac in R Studio

GUniFrac este o versiune generalizată a distanțelor UniFrac utilizate în mod obișnuit. Distanța UniFrac generalizată conține un parametru suplimentar α care controlează ponderea speciilor mai abundente. Motivul este acela de evita ca distanța UniFrac să fie influențată exclusiv de către speciile foarte abundente. UniFrac neponderat ("d_1") și ponderat ("d_UW") sunt de asemenea implementate.

Pentru utilizarea GUniFrac in R Studio este nevoie de instalarea și activarea următoarelor pachete disponibile in Cran Repository:

- `vegan`
- `ggplot2`
- `matrixStats`
- `Matrix`
- `ape`
- `parallel`
- `stats`
- `utils`
- `statmod`
- `rmutil`
- `dirmult`
- `MASS`
- `ggrepel`
- `foreach`
- `modeest`
- `inline`
- `methods`

Sintaxa funcției GUniFrac este următoarea:

```
GUniFrac(otu.tab, tree, size.factor = NULL, alpha = c(0, 0.5, 1), verbose = TRUE)
```

Cele 5 argumente se folosesc astfel:

- `otu.tab` – Este o matrice, tabelul de numărare OTU, rând - n eșantion, coloană - q OTU
- `tree` – Este un arbore filogenetic înrădăcinat din clasa R „`phylo`"
- `size.factor` – Este un vector numeric al factorilor de normalizare pentru a împărți numărările. Lungimea reprezintă numărul de probe. Acest lucru oferă flexibilitatea de a normaliza datele folosind metoda de normalizare preferată (de exemplu, factorul de normalizare GMPR). Dacă nu este furnizată, se va folosi suma totală.
- `alpha` – Este un vector numeric cu parametrii care controlează greutatea pe liniile abundente
- `verbose` – Este o valoare logică, în funcție dacă se dorește tipărirea sau nu de mesaje

Funcția returnează o listă care conține `unifracs` – matrici tridimensionale care conțin toate matricele de distanță UniFrac.