多数の配列からの代表配列選択と系統ネットワーク構築

ファイル説明書

2014 年 5 月 29 日

「pack」フォルダに格納したファイルは以下のとおりです。

1. tree_cluster.pl

「作業報告書」2.(1)①で使用した、系統樹を代表するOTUを系統樹全体からなるべく均等に選択するプログラム。

(1) 必要環境

BioPerl 1.0 以上

(2) コマンド

```
$ tree_cluster.pl [--tree | --tree_groups | --nw_groups | --out] <file>
  [--size | --reps] <int> [--interval | --neighbor] <float>
```

引数（必須）

--tree <file>には入力系統樹ファイル（Newick (New Hampshire)形式）を指定します。

--out <file>には拡張子を除く出力ファイル名を指定します。

引数（オプション）

--tree_groups <file>には系統樹用グループ分け一覧ファイルを指定します。

--nw_groups <file>には系統ネットワーク用グループ分け一覧ファイルを指定します。

--size <int>には観察数に基づいて代表クラスタの候補を選択する際の最小観察数を指定します。デフォルト値は 10 です。

--reps <int>には代表クラスタを選択する際の近接する範囲内に存在する代表クラスタ数の上限を指定します。デフォルト値は 1 です。

--interval <float>には間隔に基づいて代表クラスタの候補を選択する際の最小の間隔を指定します。デフォルト値は 1 です。

--neighbor <float>には代表クラスタを選択する際の近接する範囲近隣の範囲指定します。デフォルト値は--interval <float>と同じ値です。

(3) 機能

系統樹ファイル（Newick (New Hampshire)形式）から、指定された引数に基づいて代表 OTU を選択して出力します。

① 代表クラスタの候補の選択

--size <int>で指定された観察数を用いて、「作業報告書」1.(1)②(a)の方法で代表クラスタの候補を選択します。

--interval <float>で指定された間隔を用いて、「作業報告書」1.(1)②(b)の方法で代表クラスタの候補を選択します。

② クラスタの代表 OTU の決定

「作業報告書」1.(1)②(c)の方法でクラスタの代表 OTU を決定します。

③ 代表クラスタの選択

--neighbor <float>および--reps <int>で指定された範囲および上限を用いて、「作業報告書」1.(1)②(d)の方法で代表クラスタを選択します。

④ 一覧の出力

(a) 代表クラスタの候補の一覧

①または②で代表クラスタの候補として選択されたクラスタについて、クラスタの代表、クラスタの全メンバー、③で代表クラスタとして選択されたか否かの一覧を、拡張子.clusters のファイルに出力します。

(b) 代表クラスタの候補の一覧(MEGA用)

(a)の一覧を MEGA[1]で表示するための Group 定義ファイルを、拡張子.mega.txt のファイルに出力します。

(c) RDFファイル作成用配列一覧

③で選択された代表クラスタの代表配列名の一覧を、拡張子.network.id のファイルに出力します。

--tree_groups <file>と--nw_groups <file>のどちらかまたは両方が指定されていた場合は、それらに応じて OTU 名を変換して出力します。

(4) 実行例

```
$ tree_cluster.pl --tree Xp11_hX.tre --tree_groups Xp11_hX.tree.group
  --nw_groups Xp11_hX.nw.group --out Xp11_hX_rep --size 20 --interval 0.0004
  --neighbor 0.0002 --reps 2
```

この例では、指定された引数に基づいて Xp11_hX.tre から代表 OTU を選択して、Xp11_hX_rep.clusters と Xp11_hX_rep.mega.txt と Xp11_hX_rep.network.id に結果を出力します。

2. check_tree.pl

「作業報告書」2.(1)②で使用した、系統樹上でルートからの距離と枝の長さと観察数を集計し、ルートからの距離が長い OTU を検出するプログラム。

(1) 必要環境

BioPerl 1.0 以上

(2) コマンド

```
$ check_tree.pl [--tree | --groups | --out] <file> --place <int> --long <float>
```

引数(必須)

--tree <file>には入力系統樹ファイル(Newick (New Hampshire)形式)を指定します。

--out <file>には拡張子を除く出力ファイル名を指定します。

引数(オプション)

--groups <file>には系統樹用グループ分け一覧ファイルを指定します。

--place <int>にはルートからの距離の階級区分に使用する小数点以下の桁数を指定します。デフォルト値は 5 です。

--long <float>にはルートからの距離が長い OTU として出力する基準となる距離と、観察数が最も大きい階級の距離の、比の値を指定します。デフォルト値は 2 です。

(3) 機能

① 系統樹の集計

系統樹のルートから各ノードまでの長さ、枝の長さ、観察数を集計し、拡張子.table のファイルに出力します。OTU の観察数は--groups <file>に指定されたファイルに基づいて算出します。--groups <file>が指定されなかった場合は、すべての OTU の観察数を 1 として算出します。

(a) 各枝の情報

各枝について、ルートからの距離、枝の長さ、配下の OTU の数、配下の観察数の合計を出力します。

(b) 各OTUの情報

各 OTU について、ルートからの距離、観察数を出力します。

(c) OTUの統計情報

ルートからの距離を階級に区分し、各階級のOTUの数、観察数を出力します。階級はルートからの距離の小数点以下を--place <int>に指定された桁数 n に四捨五入したものを用います。階級の幅は 10^{-n}(n は 0 以上の整数)となります。

② ルートからの距離が長い OTU 名の出力

ルートからの距離が、①(c)で観察数が最大となった階級の距離に--long <float>で指定した値を掛けたものより大きい OTU を MEGA[1]で表示するための Group 定義ファイルを、拡張子.mega.txt のファイルに出力します。

(4) 実行例

```
$ check_tree.pl --tree Xp11_hX.tre --groups Xp11_hX.tree.group
  --out Xp11_hX_treecheck
```

この例では、指定された引数に基づいて Xp11_hX.tre の系統樹を集計したものを Xp11_hX_treecheck.table に出力し、ルートからの距離が長い OTU を Xp11_hX_treecheck.mega.txt に出力します。

3. Xp11_hX/*

Xp11_hX 領域の系統樹および系統ネットワーク

4. dys44/*

dys44 領域の系統樹および系統ネットワーク

5. rrm2p4/*

RRM2P4 領域の系統樹および系統ネットワーク

6. mcph1/*

MCPH1 領域の系統樹および系統ネットワーク

7. 17q21inv/*

17q21inv 領域の系統樹および系統ネットワーク

8. stat2/*

STAT2 領域の系統樹および系統ネットワーク

9. oas/*

OAS 領域の系統樹および系統ネットワーク

10. hyal/*

HYAL 領域の系統樹および系統ネットワーク

[1] MEGA … URI: http://www.megasoftware.net/