# 1000 Genomes Project data を用いた
# ヒトゲノム遺伝子流動候補領域における、
# ハプロタイプ間系統樹作成

# ファイル説明書

2013年8月5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　add_bootstrap.pl

「作業報告書」8で使用した、別途計算した Bootstrap 値を系統樹に書き込むプログラム。

### (1) 機能

Bootstrap値が書かれていない系統樹ファイル（Newick形式）とBootstrap値を記述したファイル（PHYLIP Consense 形式）をもとに、Bootstrap値を書いた系統樹ファイル（Newick形式）を出力します。

Bootstrap値を記述したファイル（PHYLIP Consense 形式）は、PHYLIP Consense または EMBASSY fconsense で作成したものを使用してください。この形式のファイルには Bootstrap data から構築したいずれかの系統樹に1回以上現れるすべてのノードの Bootstrap 値が記載されていますので、Bootstrap consensus tree に現れないノードであっても Bootstrap 値を知ることができます。

### (2) 必要環境

BioPerl 1.0 以上

### (3) コマンド

```
$ ./add_bootstrap.pl <input_tre> <input_consense> <output_tre>
```

<input_tre>には入力系統樹ファイル名を指定します。

<input_consense>には入力するBootstrap値を記述したファイル名を指定します。

<output_tre>には出力系統樹ファイル名を指定します。

### (4) 実行例

```
$ ./add_bootstrap.pl input.tre input.consense output.tre
```

この例では、input.consenseに記載されているBootstrap値をinput.treに書きこんだものを、output.treに出力します。

## 2　fasta2tre_bs.bsh

「作業報告書」9で使用した、アライメントから系統樹構築とBootstrap値の計算を連続して行うシェルスクリプト。

### (1) 機能

FASTA形式の配列をもとに、アライメントを行い、系統樹構築とBootstrap値の計算を行います。

シェルスクリプトから実行したプログラムが0でない値を返して終了したときは、その時点でシェルスクリプトの

実行を終了します。

UGE (Univa Grid Engine) を使用したスーパーコンピュータのジョブスクリプトとして使用することもできます。

**① アライメント**

MAFFTを使用して、入力FASTAファイルをFFT-NS-2法でアライメントします。

**② 元配列の距離の計算**

PHYLIP Dnadist を使用して、①の配列からOTU間の距離を計算します。

**③ 元配列の系統樹の作成**

PHYLIP Neighbor を使用して、②の距離行列から系統樹を作成します。

**④ Bootstrap data の作成**

PHYLIP Seqboot を使用して、①の配列から500個の Bootstrap data を作成します。

**⑤ Bootstrap data の距離の計算**

PHYLIP Dnadist を使用して、④の配列から500個の Bootstrap data のOTU間の距離を計算します。

**⑥ Bootstrap data の系統樹の作成**

PHYLIP Neighbor を使用して、⑤の距離行列から500個の Bootstrap data の系統樹を作成します。

**⑦ Bootstrap data の Bootstrap 値の計算**

PHYLIP Consense を使用して、⑥の500個の系統樹から Bootstrap 値を計算します。

**⑧ Bootstrap 値の書き込み**

add_bootstrap.pl を使用して、⑦の Bootstrap 値を③の系統樹に書きこみます。

### (2) 必要環境

MAFFT

PHYLIP

add_bootstrap.pl(カレントディレクトリに置いてください。)

BioPerl 1.0 以上(add_bootstrap.pl 内で必要)

### (3) コマンド

```
$ ./fasta2tre_bs.bsh <input_FASTA> <output_tre>
```

<input_FASTA>には入力FASTAファイル名を指定します。

<output_tre>には出力系統樹ファイル(Newick形式)名を指定します。

### (4) 実行例

```
$ ./fasta2tre_bs.bsh 1000g.fa 1000g.bv.tre
```

この例では、1000g.faを入力FASTAファイル読み込み、アライメントを行い、系統樹構築とBootstrap値の計算を行った結果を1000g.bv.treに出力します。

## 3 rm_zerodist.pl

「作業報告書」10(2)②で使用した、距離が0同士の複数のOTUを1つにまとめるプログラム。

### (1) 機能

アライメントファイル(PHYLIP形式)と距離行列ファイル(PHYLIP形式)をもとに、距離が0である複数のOTUを1つにまとめたアライメントファイル(PHYLIP形式)を出力します。また、どのOTUがどのグループにまとめられたかをログファイルに出力します。

複数のOTUが1つにまとめられたもののID名は以下のようになります。

- ID名は3つのパートに分かれます: [種類][番号]-[数]

- [種類]は含まれる配列の種類により次の文字が入ります。
  - ヒトのreferenceとDenisovaのいずれも含まれないグループ … H
  - ヒトのreferenceが含まれ、Denisovaが含まれないグループ … HrH
  - Denisovaが含まれ、ヒトのreferenceが含まれないグループ … DH
  - ヒトのreferenceとDenisovaの両方が含まれるグループ … HrDH
- [番号]はグループが見つかった順番に整数が入ります。
- [数]はグループに属すOTUの数が入ります。
- ただし、これらの規則によって命名したものが11文字以上になるときは、先頭から10文字だけが使われます。11文字目以降は捨てられます。

例

(a) H3-10：3番目に見つかったグループで、10個のOTUが属し、ヒトのreferenceとDenisovaのいずれも含まれないグループ

(b) HrH12-3 ： 12番目に見つかったグループで、3個のOTUが属し、ヒトのreferenceが含まれ、Denisovaが含まれないグループ

### (2) コマンド

```
$ ./rm_zerodist.pl <input_distance> <input_phy> [<output_phy> [<output_log>]]
```

<input_distance>には入力距離行列ファイル（PHYLIP形式）名を指定します。

<input_phy>には入力アライメントファイル（PHYLIP形式）名を指定します。

<output_phy>には出力アライメントファイル（PHYLIP形式）名を指定します。

<output_log>には出力ログファイル名を指定します。

<output_phy>と<output_log>の両方を、または<output_log>のみを、省略することができます。

<output_phy>が省略されたときのファイル名は、<input_phy>が“.phy”で終わるときは<input_phy>の“.phy”を“.uniq.phy”に変えたものになり、そうでないときは<input_phy>の後に“.uniq.phy”を加えたものになります。

<output_log>が省略されたときのファイル名は、<output_phy>が“.uniq.phy”で終わるときは<output_phy>の“.uniq.phy”を“.zerodist”に変えたものになり、そうでなく<output_phy>が“.phy”で終わるときは<output_phy>の“.phy”を“.zerodist”に変えたものになり、そうでないときは<output_phy>の後に“.zerodist”を加えたものになります。

### (3) 実行例

```
$ ./rm_zerodist.pl old.dist old.phy new.phy
```

この例では、old.distを入力距離行列ファイル（PHYLIP形式）として、old.phyを入力アライメントファイル（PHYLIP形式）として読み込みます。次にold.distにおいて距離が0になっているOTUを1つのグループにまとめ、old.phyのうち同一グループに属すOTUを1つにまとめたものをnew.phyに出力します。グループ名とOTU名の対応はnew.zerodistに出力します。

### (4) 制限事項

#### ① 距離行列ファイルの形式

入力距離行列ファイル（PHYLIP形式）の行列の形状は“Square form”である必要があります。“Lower-triangular form”または“Upper-triangular form”になっていると正しく動作しません。

#### ② 距離行列の値

距離行列の左上と右下を結ぶ対角線上にある要素には0が指定されていなければなりません。0でないとき

はエラーになります。

距離行列の左上と右下を結ぶ対角線に対して対称の位置にある要素には同一の値が指定されていなければなりません。同一でないときはエラーになります。

③　距離が0であるOTU

あるOTUと、別の複数のOTUとの間の距離が0のとき、その別の複数のOTU同士の距離も0でなければなりません。例えばOTU-AとOTU-Bとの間の距離が0で、OTU-AとOTU-Cとの間の距離も0のとき、OTU-BとOTU-Cとの間の距離も0でなければなりません。そうでないときはエラーになります。

4　1000g.tre

「作業報告書」6(1)で作成した 1000 Genomes Project data およびオランウータン・チンパンジー・デニソバのXp11_hX領域のゲノム配列の系統樹ファイル（Newick形式）。

5　1000g.consensus.tre

「作業報告書」7で作成した 1000 Genomes Project data およびオランウータン・チンパンジー・デニソバのXp11_hX領域のゲノム配列の Bootstrap consensus tree の系統樹ファイル（Newick形式）。

6　1000g.bv.tre

「作業報告書」8で作成した 1000 Genomes Project data およびオランウータン・チンパンジー・デニソバのXp11_hX領域のゲノム配列の系統樹に Bootstrap 値を書き込んだ系統樹ファイル（Newick形式）。

7　1000g.mts

「作業報告書」10(1)で作成した系統樹ファイル（MEGA形式）。

8　1000g.uniq.mts

「作業報告書」10(2)③で作成した系統樹ファイル（MEGA形式）。