DistAFLP

Why was written and what does "DistAFLP"

The program 'DistAFLP' has been developpped to calculate current and evolutionnary genome divergences from AFLP data (Amplified Fragment Length Polymorphism [Vos & Zabeau 1993]), based on the number of nucleotides involved in restriction and selective PCR, which constrain AFLP results (Mougel *et al.* 2002).

The program can also be used to calculate genome divergences from other DNA profiling methods such as RAPD or IRS-PCR providing that the number of "constraining nucleotides" is known.

An option "NL" for Nei & Li is given to calculate the rate of nucleotide substitution from RFLP banding data according to Nei & Li (1979).

With data for which the concept of "constraining nucleotides" is not applicable (for instance morpho-biochemical traits), the program can be used to calculate the simple distance, which is "one minus similarity". In this case, the number of constraining nucleotides will be "1" (default value).

The 'DistAFLP' program calculates similarities (Jaccard or Dice index), current genome mispairing and evolutionary genome divergence with or without bootstrap resamplings for the purpose of genomic phylogeny. For this reason, the main output file is in a format suitable for using the phylogenetic inference package PHYLIP of Felsenstein (1993).

Particularities of "DistAFLP" (adapted from Mougel *et al.* 2002)

Similarities between pairs of OTUs.

Similarities can be calculated in two ways. The pattern similarity used for the comparison of current traits is determined with the Jaccard index (S_{Jxy}). The similarity with the common ancestor of two OTUs required for phylogenetic studies is determined by the Dice index (S_{Dxy}).

$$S_{Jxy} = n_{xy} / (n_{xy} + \Delta_{xy}) \qquad \text{equ. 1}$$

$$S_{Dxy} = n_{xy} / [n_{xy} + (\Delta_{xy}/2)] \qquad \text{equ. 2}$$

in which n_{xy} is the number of traits (i.e. fragments in the case of AFLP) common to both OTUs x and y, and Δ_{xy} is the number of traits found only in x or only in y.

Estimation of genome divergence from AFLP data.

The occurrence of a common AFLP fragment in two OTUs is assumed to require the identity of r nucleotide sites (i.e. constraining nucleotides) involved in both restriction and amplification. As a consequence, the proportion of common AFLP fragments pF is given by:

$$pF = (1 - d)^r \qquad \text{equ. 3}$$

where d is the rate of base mispairing between two genomes (i.e. genome mispairing).

Conversely, the proportion of nucleotide differences is given by:

$$d = 1 - (pF)^{1/r} \qquad \text{equ. 4}$$

Current genome mispairing and evolutionary genome divergence.

The current genome mispairing suitable to look for current differences between genomes is given by d_{Jxy} after substitution of S_{Jxy} for pF in equation 4, because the likelihood for the current number of sites n_{xy} and Δ_{xy} is maximum by using the Jaccard index.

The evolutionary genome divergence suitable for phylogenetic studies, d_{Dxy} is calculated after substitution of S_{Dxy} for pF in equation 4, because the Dice index is suited for phylogenetic studies by assuming that all sites that are shared between two strains were present at a common ancestor halfway between them.

Jukes-Cantor correction for mutiple substituions.

The evolutionary distance can be corrected to account for the unobserved substitutions using the standard Jukes-Cantor correction, which assumes equal rates of substitution between all pairs of bases. The evolutionary genome divergence, expressed as number of nucleotide substitution per site is estimated by:

$$\hat{t} = -(3/4) \ln [1 - (4/3)\, d_{Dxy}] \qquad \text{equ. 5}$$

Correction for fragment dependence.

Simple computation

Bands seen on electrophoregrams that differ between strains x and y (∂xy) are more numerous than DNA fragments that really differ between strains (Δxy) because with some point mutations bands move but do not disappear. A correction for fragments dependence is given by K such as:

$$\Delta_{xy} = K\, \partial_{xy} \qquad \text{equ. 6}$$

In the case of AFLP, the correction factors K for fragment dependance is given by:

$$K = (1/r)\, [\mathrm{E}/(1 + p\mathrm{EE}.p\mathrm{E}) + \mathrm{M}/(1 + p\mathrm{MM}.p\mathrm{M}) + \mathrm{N}] \qquad \text{equ. 7}$$

in which :

E and M are the numbers of nucleotides in the restriction sites (e.g. 6 and 4, respectively), and N the number of selective nucleotides (NB: $\mathrm{E} + \mathrm{M} + \mathrm{N} = r$);

$p\mathrm{E}$ and $p\mathrm{M}$ are the probability of the hexacutter and quadracutter sites in both genomes, respectively.

pEE and pMM are the proportions of fragments that produce two different bands in AFLP patterns following mutations in the hexacutter or the quadracutter sites, respectively;

NB: pEE and pMM were estimated by determining the proportion of fragments predicted to have identical hexacutter site extremities, and identical quadracutter site extremities, respectively, by simulation using large DNA sequences such as complete genome sequences.

NB: pE = $(1 - d)^E$ and pM =$(1 - d)^M$ are maximal for nearly identical strains and decrease with genome mispairing, thus fragment dependence tend to be disregardable between species. For simplification, we use pE = pM = 1 in the present version of DistAFLP.

Bootstrap computation

With bootstrap, distance data were not corrected for fragment dependence, but involved sampling K% fragments at random among the 100%.

Format of input and output files

Input files

The input file is a tabular binary matrix in PHYLIP sequencial format, with the first line containing the number of OTUs and the total number of different DNA fragments separated by a space. The following rows correspond to OTUs with OTU names on 10 characters as in the input file and binary data not separated by spaces. Commonly, the input file is generated by the program 'LecPCR'.

Input files must be saved in "text only" format.

Output files

Output files are upper semi-triangular matrices in PHYLIP format.
Depending of chosen options, the outputfile are:

Simple computation

- with Jaccard (usually without JC correction):
filename.matJ matrix of Jaccard distances (e.g. current genome mispairing)
filename.matJP matrix of Jaccard similarities
filename.nxy matrix of number of common fragments
The same files can be obtained in ADE-4 format

- with Dice (usually with JC correction):
filename.matDJC matrix of corrected Dice's distances (e.g. evolutionary genome divergence)
filename.matDP matrix of Dice similarities
filename.nxy matrix of number of common fragments
The same files can be obtained in ADE-4 format

- with NL (only for RFLP banding patterns)
filename.matNL matrix of NL's distances (e.g. evolutionary gene divergence)

Bootstrap computation
(generally with Dice and JC correction)

filename.matDJC matrices of corrected Dice distances (e.g. evolutionary genome divergence)

Running DistAFLP

At opening, a dialog box is open. The input file must be in the current directory, else the directory can be selected using "Preference" in the "Edit" menu.

'Simple computation' or 'Bootstrap computation' are selected using the "Options" menu.

The dialog box asks for :
'Restriction sites length': indicate the value "r", or "1" for simple distance.
'K value': indicate the value calculated (for instance 0.89 with Agrobacterium using *Eco*RI+2 nucleotides and *Mse*I), or let the default value "1".
'Index": choose between Jaccard "1", Dice "2", Nei & Li "3".
'Jukes-Cantor correction': indicate "1" to choose this option preferably with Dice for evolutionary studies.
'ADE-4': indicate "1" to choose this option for further multivariate analyses using the ADE-4 package (Thioulouse *et al.* 1997)

NB: the NL option is not valid with bootstrap computation.

Example

Input file : ex: testfile01 (PHYLIP format)

```
------------------------------------------------------------------------------------------------------
9 55
name1     0101110100010000110110000000101001000110101011000000000
name2     0101010001010110011111000000000000001000000101000010111
name3     0001101100010000110111000000111001000000000011000010000
name4     0001010001011100111100000000101110011000000011000010000
name5     0101001000000000011111000111101100101000010010000000000
name6     0101000000011100000111101000101100101000010011000000000
name7     0011011010010010010110010000000000001001001011100011000
name8     0101010000111000111100000011011000001000000011000110000
name9     1101010000110001010011000011011000001000000000011011000
------------------------------------------------------------------------------------------------------
```

Output files:

ex: testfile01.matDJC

```
------------------------------------------------------------------------------------------------------
9
name1     0.0679 0.0249 0.0499 0.0657 0.0562 0.0583 0.0499 0.0788
name2            0.0635 0.0423 0.0657 0.0562 0.0499 0.0499 0.0583
name3                   0.0455 0.0517 0.0517 0.0540 0.0455 0.0635
name4                          0.0562 0.0401 0.0583 0.0291 0.0788
name5                                 0.0311 0.0766 0.0477 0.0562
name6                                        0.0766 0.0562 0.0766
name7                                               0.0583 0.0679
name8                                                      0.0354
```

```
name9
----------------------------------------------------------------------------------
```

ex: testfile01.matDP

```
----------------------------------------------------------------------------------
9
name1     0.4497 0.7450 0.5552 0.4612 0.5156 0.5028 0.5552 0.3959
name2            0.4734 0.6071 0.4612 0.5156 0.5552 0.5552 0.5028
name3                   0.5841 0.5433 0.5433 0.5291 0.5841 0.4734
name4                          0.5156 0.6223 0.5028 0.7088 0.3959
name5                                 0.6920 0.4062 0.5693 0.5156
name6                                        0.4062 0.5156 0.4062
name7                                               0.5028 0.4497
name8                                                      0.6583
name9
----------------------------------------------------------------------------------
```

ex: testfile01.nxy

```
----------------------------------------------------------------------------------
9
name1        19     8    13    10     8     9     9    10     7
name2              19     8    11     8     9    10    10     9
name3                    17    10     9     9     9    10     8
name4                          19     9    11     9    13     7
name5                                18    12     7    10     9
name6                                      18     7     9     7
name7                                            19     9     8
name8                                                  19    12
name9                                                        19
----------------------------------------------------------------------------------
```

References

Felsenstein, J. (1993). PHYLIP (Phylogeny Inference Package) version 3.5c, Distributed by the author, Department of genetics, University of Whashington, Seattle, USA.

Mougel C., Thioulouse J., Perrière G., Nesme X. 2002. A mathematical method for determining genome divergence and species delineation using AFLP. Int. J. Syst. Evol. Microbiol. 52:573-586.

Nei, M., Li, W.-H. 1979. Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc. Natl. Acad. Sci. USA 76:5269-5273.

Thioulouse, J., Chessel, D., Dolédec, S. & Olivier, J.M. (1997). ADE-4: a multivariate analysis and graphical display software. Statistics and Computin*g* 7:75-83.

Zabeau, M. & Vos, P. (1993). Selective restriction fragment amplification: a general method for DNA fingerprinting, *Publication* 0 534 858 A1, Münich, Germany: European Patent Office.