POFAD 1.07

Phylogenetic Analysis From Allelic Data

© 2005-2014 Simon Joly
Montreal Botanical Garden, Montréal, Canada
joly.simon@gmail.com

Disclaimer

This program is free software; you can redistribute it and/or modify it under the terms of the GNU General Public License as published by the Free Software Foundation; either version 2 of the License, or (at your option) any later version. This program is distributed in the hope that it will be useful, but WITHOUT ANY WARRANTY; without even the implied warranty of MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE. See the GNU General Public License for more details. You should have received a copy of the GNU General Public License along with this program; if not, write to the Free Software Foundation, Inc., 51 Franklin St, Fifth Floor, Boston, MA 02110-1301 USA

Citing the program

When using the program POFAD, you should cite Joly and Bruneau for the software POFAD using a sentence similar to the following: “ the distance matrix was calculated using the program POFAD (Joly & Bruneau 2006)”. In addition, you should cite the proper reference for the specific distance used (see Distances section below).

Version history

Version	Date released	Modifications
1.07	July 2014	-> PP distance is divided by 2 compared to original description
1.06	April 2014	-> Implementation of FRQ, PBC, MIN, and PP distances
1.05	Sept 2007	-> Implementation of genpofad, matchstates, and MRCA algorithms (from sequences) -> Can compute the POFAD algorithm directly from sequences -> Can create consensus sequences per organisms -> Completely new presentation of the program
1.03	Oct 2006	-> A bug was corrected that crashed the program when a nexus distance file with no diagonals was entered. -> Better handling of nexus distance files: now accepts upper

		triangle matrices and square matrices
1.02	Oct 2005	-> The program now allows that some datasets lack information for organisms (i.e., allows SuperTree reconstruction)
1.01	Sept 2005	-> Give two output matrices: one attributes the same weight to all matrices and one leaves the matrices as is.

Distances available

POFAD implement several methods for estimating distances between organisms between individuals. We formally describe below all the distances implemented. For a comparison on the relative performance of the distance methods, see Joly et al. (2014).

Some distances are nucleotide-based whereas others are sequence-based. Before defining each distance, let us define a few notations. For nucleotide-based distances, let A_X^i be the complete set of nucleotides for individual X at sequence site i and let $\left|A_X^i\right|$ be the number of nucleotide states observed at site i for individual X. For sequence-based method, let A_X be the complete set of alleles (or gene copies) for individual X and let $|A_X|$ be the number of alleles observed in individual X.

Matchstates

The matchstates distance (Joly et al. unpublished) is a nucleotide-based distance method. It looks at each nucleotide present in one individual and tries to see if there is a nucleotide in the other individual that matches. More formally, let A_X^i be the complete set of nucleotides for individual X at sequence site i and let $\left|A_X^i\right|$ be the number of nucleotide states observed at site i for individual X. Then, the matchstates distance between individual X and individual Y is

$$\text{matchstates}_{XY}^i := \frac{\left|A_X^i \Delta\, A_Y^i\right|}{|A_X| + |A_Y|},$$

where $A_X^i \Delta\, A_Y^i$ denotes the set of elements that belong to either A_X^i or A_Y^i, but not in both.

genpofad

The genpofad distance (Joly et al. unpublished) is a nucleotide-based distance method that is called after the POFAD algorithm described by Joly and Bruneau (2006). The genpofad distance can be defined as one minus the ratio of the number of nucleotides shared between two individuals divided by the maximum number of nucleotides observed in either of the individuals. Following the notation introduced above,

$$\text{genpofad}_{XY}^i := 1 - \frac{\left|A_X^i\ \cap A_Y^i\ \right|}{max\left(\left|A_X^i\ \right|, \left|A_Y^i\ \right|\right)}.$$

mrca

The mrca distance (Joly et al. unpublished) is a nucleotide-based distance method. It gives a distance of 0 whenever two individuals share at least one nucleotide at a sequence site and a distance of 1 otherwise. Formally, the mrca distance between individual X and individual Y for site i is

$$\mathrm{mrca}_{XY}^{i} := \begin{cases} 0\ if\ \left|A_X^i \cap A_Y^i\right| \neq \emptyset \\ 1\ if\ \left|A_X^i \cap A_Y^i\right| = \emptyset \end{cases}.$$

Nei's genetic distance

This distance is the application of Nei's genetic distance (1983) at the nucleotide level. The frequency of each nucleotide is estimated per nucleotide for each individual and then Nei's genetic distance between individual X and individual Y is estimated as

$$\mathrm{nei}_{XY}^{i} := 1 - \sum_{j}^{A,C,T,G} \sqrt{p_{j\in A_X}^{i} p_{j\in A_Y}^{i}},$$

where $p_{j\in A_X}^{i}$ is the frequency of nucleotide j in individual X at site i. This formula is flexible as it can be easily applied among individuals from different ploidy levels. Gene dosage is assumed to be known, but it can also be used if it is unknown by giving equal weight to each nucleotide.

MIN distance

The MIN distance is based on allelic distances and not nucleotides. It was proposed by Göker and Grimm (2008) in the present context, but it had been often used in other contexts as well (e.g., Joly *et al.* 2009; Liu *et al.* 2009; Mossel & Roch 2010). It can be described as

$$\mathrm{MIN}_{XY} := min\left(d_{ij} | i \in A_X, j \in A_Y\right).$$

where d_{ij} is the genetic distance between alleles i and j.

Phylogenetic Bray-Curtis distance (PBC)

The PBC distance is a sequence-based distance defined by Göker and Grimm (2008) as:

$$\mathrm{PBC}_{XY} := \frac{\sum_{i\in A_X} min\left(d_{ij} | j \in A_Y\right) + \sum_{j\in A_Y} min\left(d_{ij} | i \in A_X\right)}{|A_X| + |A_Y|}.$$

PP distance

The PP distance is a nucleotide-based method (Potts *et al.* 2014). Distances between nucleotides are estimated using the step-matrix presented in Figure A.1. The distance output by POFAD differs slightly from the original paper; it is divided by 2 to give a distance of 1 between two standard nucleotides (e.g., A to G).

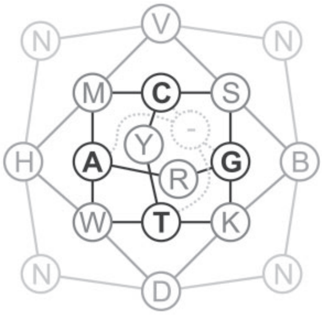

Figure A.1. Step-matrix used for estimating the PP distance (reproducted from (Potts *et al.* 2014)).

POFAD

This is the original algorithm described by Joly and Bruneau (2006). Contrarily to the previous distances, it allows a maximum of 2 alleles per individual and works only with distances files as input (no nucleotide files). The distance matrix can be obtained by any software that calculates distances from sequence data. Then, the algorithm transforms this distance matrix of alleles into a distance matrix of organisms.

To describe the algorithm, let's take an hypothetic example where letters are used to distinguish individuals: capital and lower-case letters represent individuals and alleles, respectively. Alleles within an individual are set apart by a number (1 or 2). Let also *d*(*A*, *B*) be the distance between individuals *A* and *B* and *d*(*a*, *b*) be the distance between alleles *a* and *b*. Moreover, let min[*x*, *y*] be the minimal value of *x* and *y*. When calculating the distance between two organisms, three situations could be encountered.

Both organisms has a single allele

In this case, the distance between the organisms is simply the distance between the alleles. If *A* and *B* are two individuals that each has one allele:

$$d(A, B) = d(a, b).$$

One organisms has one allele and the other has two alleles

If *A* is an individual with one allele (*a*) and *C* is an individual with two alleles (*c*1, *c*2), then

$$d(A, C) = \frac{d(a, c1) + d(a, c2)}{2}.$$

Both organisms have two alleles

Two individuals, *C* and *E*, both have two alleles (*c*1, *c*2 and *e*1, *e*2). There are two pairs of allelic distances possible among these individuals: *d*(*c*1, *e*1) and *d*(*c*2, *e*2) or *d*(*c*1, *e*2) and *d*(*c*2, *e*1). The distance between such organisms is the mean of the shortest pair of distances:

$$d(C, E) = \frac{\min[d(c1, e1) + d(c2, e2), d(c1, e2) + d(c2, e1)]}{2}.$$

Combining multiple genes

The matrix of organisms obtained from one marker can either be used alone or be combined with matrices of organisms obtained from other markers. When combining matrices, two different options are possible.

Standardized matrices

With this option, all individual matrices are given the same weight by giving a distance of 1 to the largest distance of the matrix and by scaling the others appropriately. So if *X* and *Y* are two individuals, their distance in the final matrix will be:

$$d_{XY} = \frac{1}{n}\sum_{i=1}^{n} d_{XY}^{i}/\max\left(d_{XY}^{i}\right)$$

where n is the number of datasets and where $\max\left(d_{XY}^{i}\right)$ is the largest distance for dataset i. Note that if you use this option, the final distances cannot be interpreted anymore as number of substitutions per site.

Non-standardized matrices

The other option is to not standardize the data. In such a case, the final distance between individuals is simply the mean distance over all matrices:

$$d_{XY} = \frac{1}{n}\sum_{i=1}^{n} d_{XY}^{i}$$

where n is the number of datasets. With this option, the individual matrices won't have the same weight and the most variable matrices will have a bigger weight in the final matrix.

Installing the program

The program POFAD is written in C++. Binaries of the program for Windows 32 and MAC OSX can be found in the "bin" folder. The source code is provided with the program, and makefiles are provided for compiling the program using Borland (Makefile.bcc) or g++ (Makefile). If only type

```
>make
```

The Makefile will give you different options. For example, if you type

```
>make install
```

then pofad will be compiled, moved to the /bin folder, and remove the object files in the /src folder. If you choose to use Borland, you should type

```
>make -f Makefile.bcc install
```

Running the program

The program needs to be in the same folder as the files to be analysed. POFAD has an interactive menu where it allows you to enter different options, but it can also be run using command line entered arguments (batch mode, see below). To run the program, you will need to move to the folder containing the program and the files in the terminal and type “pofad” (With MAC OSX, you need to type ./pofad to indicate that the program is in the present directory). Alternatively, you may also double-click on the icon. Once the program is executed, you will see the following “phylip type” menu:

```
************************  POFAD version 1.07 ***********************

 Select one of the following options:

 A     Distance algorithm                   matchstates
 I     File containing the organisms        (no file entered)
 D     File containing the datasets         (no file entered)
 O     Name of the output file              pofad_output.nex
 W     Distance standardization             Standardized distances
 M     Multiple hits correction             Jukes and Cantor
 G     Gap handling                         missing
 ?     Missing nucleotides                  Infer
 Z     Missing distances                    Estimate

 Other options:

 Q     Quit the program

 ********************************************************************

 Type "Y" to accept these settings or the letter for one to change:
```

To change one of the options, just type the letter at the beginning of the line (case-insensitive). Once everything is OK and you are ready to run the program, type ‘Y’. Let’s look at the different options one by one.

Distance algorithm

The different algorithms are described above.

Create consensus sequences

This creates a consensus sequence for each organism from its alleles.

File containing the organisms

This file is required for all methods implemented in POFAD. It consists of a nexus-type block that looks like this:

```
Begin organisms;
   Dimensions Norg=4;
   OrgLabels blanda foliolosa nitida palustris;
end;
```

It is very similar to a nexus taxa block. The number of organisms need to be given after “Norg=”. The organisms need to be separated by white spaces (or tab or new line). POFAD converts all organism and haplotype names to capital letters. Therefore, organisms that only differ by case won’t be differentiated and will lead to bad results. *(see organisms.nex in folder /example_files)*

File containing the datasets

This file contains a nexus-type block that list the datasets to be used by the program for distance calculations. It is not required by the program; if no dataset file is given the program will prompt you for the names of all datasets (which is much more likely to lead to typo errors). The “Begin datasets” block looks like this:

```
Begin datasets;
   Dimensions NDatasets=3;
   DatasetsLabels gapdh_dist.nex tpi_dist.nex ms_dist.nex;
end;
```

The number of datasets need to be given after “NDatasets=”. The datasets need to be separated by white spaces (or tab or new line). When using UNIX, the names of the files are case-sensitive, so be careful (this is not the case when using Windows). The format for each dataset file is described below under the section “Format for input files”. *(see datasets.nex in folder /example_files)*

Name of the output file

The program outputs a distance file in nexus format, or alternatively a sequence nexus file if the ‘consensus sequence’ option is chosen. The default file name is “pofad_output.nex.

The program also outputs a log file. I strongly suggest that you look at the log file to make sure everything seems fine. Amongst other information found in the log file is a table listing how many alleles of each individual were found for each dataset. This is useful to make sure none was forgotten because of a mistake in naming alleles, for example.

Distance standardization

This determines how the distances of the original datasets are scaled when calculating the final distance of organisms. The default is to use standardize distances where for each dataset, the maximum distance is given a value of 1 and all other distances are scaled proportionally to this value (see above). The other option is to use raw distances, where no changes are performed to the distances of each dataset. ***Note that if you choose to standardize the distances, the resulting distances cannot be interpreted as substitutions per site anymore.***

Gap handling

The default is to treat gaps as missing characters. You may also choose to assume that gaps represent a fifth character state.

When gaps are treated as a fifth character state, the methods matchstates, genpofad, and mrca use the approach of Bandelt & Dür (2007) to include gaps in the IUPAC nucleotides. This is not the case for 2ISP as the inclusion of gaps is not straightforward in that case.

Multiple hits correction

This determines whether the distances are corrected for multiple hits or not. The Jukes and Cantor distance is used for taking into account multiple hits (Jukes & Cantor 1969). If this option is selected and assuming that the distance obtained from the algorithm is p, then the corrected distance d will be:

$$d = -\frac{3}{4} log_e \left(1 - \frac{4}{3} p\right).$$

This correction is very simple, yet it is difficult to think of more complex ones when you calculate distance from polymorphic states (e.g., R,Y,N). Calculating the number of transition and transversion, which are required for more complex models of nucleotide evolution (e.g., HKY, TN, etc.), is not straightforward.

Note that you should not proceed to a Jukes and Cantor correction if you are using distances that were already corrected for multiple hits.

Missing nucleotides

There are two options for missing states: ignore them or infer them. If the missing states are ignored, sites for which one or both sequence have a '?' are ignored and are not taken into account when computing the distance between sequence. When they are inferred, sites with missing data are included in the distance calculation and missing states are assign a state as to minimize the distance. In other words, when one of the two sequence have a '?' at a site, the distance is always zero for this site when missing data are inferred.

Missing distances

The program allows that some of the datasets lack information concerning certain organisms, as long as the distance between any two individuals is available in at least one of the datasets. When there are missing data, the final distance between organisms is the mean of the distances in the datasets that have this information. The program also outputs the percentage of missing distances present in each dataset.

If some distances are missing in the final matrices, the distances reported are '–999' for these.

Inferring missing distances in the final matrix

The program offers to infer the missing distances (if present) in the final matrix. This uses the additive method of Landry *et al.* (1996), using the implementation of Vladimir Makarenkov from the T-Rex software (Makarenkov 2001). Note that this inference of missing distances assumes an additive property of the distance matrix. Therefore, the results might be highly biased if one suspects that the matrix should not fulfill this criteria, such as then there are some hybrid individuals in the matrix. We recommend using this option with great care.

Batch File Mode

POFAD can also be run in batch file mode. Here are the different arguments:

```
pofad [-i] [file_containning_the_organims]
      [-b] [Run in batch file mode]
      [-d] [name of file containing datasets names]
      [-a] [distance method: 0=genpofad (default), 1=matchstates,
               2=mrca, 3=classic pofad, 4=FRQ, 5=PBC,
               6=Output consensus sequences, 8=MIN, 9=Nei, 10=2ISP]
      [-c] [name of file for consensus sequence]
      [-m] [multiple hits correction: 0=none, 1=JC model]
      [-w] [distance standardization: 0=false, 1=true]
      [-g] [Gives gap handling: 0=missing, 1=fifth state]
      [-?] [missing nucleotides: 0=ignore, 1=infer]
      [-z] [missing distances: 0=leave them, 1=infer]
```

```
            [-o] [name_for_the_output_file]
            [-v] [sets the program in verbose mode]
```

If the different files or names are not entered as arguments, the program will ask you the names of the different files.

Input file format

File of organisms

A file containing the name of the individuals between which we want to calculate the distance. The file must be in a nexus style. You need to call a block 'begin organisms' that contains a 'dimensions' line with the keyword 'NOrg=' followed by the number of organisms in the file and a semicolon. Then there need to be a OrgLabels line followed by the names of the individuals included separated by a new lane or by a space. The OrgLabel command must finish by a semicolon and the whole block must finish with an 'end;'. *(see organisms.nex in folder /example_files)*

File containing the datasets (this file is not required)

The program needs a file that contains the names of the datasets to be included in the analysis. The file must be in a nexus style. You need to call a block 'begin datasets' that contains a 'dimensions' line with the keyword 'NDatasets=' followed by the number of datasets in the file and a semicolon. Then there need to be a DatasetsLabels line followed by the names of the individuals included separated by a new lane or by a space. The DatasetsLabel command must finish by a semicolon and the whole block must finish with an 'end;'. *(see datasets.nex in folder /example_files)*. This file is not required and if not provided, the program will prompt you for the dataset file names.

Data files

For each marker, a data file in nexus format must be given. This can be in nucleotide format for all methods, or in distance format for the methods based on distances (POFAD, PBC, MIN).

Allele's names

In order for Pofad to correctly associate the alleles to the proper organism, it is very important to follow a strict rule. Each allele name (in the data files) needs to start by the exact organism name, followed by and underscore character ('_'), and then by some character that identifies the alleles. In other words, the underscore character is used to identify the organism name in the allele name. For this reason, ***it is very important that the organisms names do not contain underscore characters***. Note that the maximum length of haplotypes (and of organisms) is of 99 characters.

Distance files

For distance nexus files, a taxa block needs to be provided and interleaved matrices are not supported. Also, the order of the taxa in the matrix has to be the same as in the "taxa block".

Sequence files

This is a typical sequence nexus file. POFAD accepts both sequential and interleaved files. Only IUPAC nucleotides are accepted, e.g., one of {A,C,G,T,R,Y,W,C,…,N}. Binary character {0, 1, 2, etc.} are not allowed. Also, note that for POFAD, 'N' and '?' are very different. '?' means a lack of information whereas 'N' means that all four states {A,C,G,T} are present at this position.

For the present, I do not recommend that you include sequences with polymorphisms when treating gaps as a fifth state. The problem is that Pofad automatically converts nucleotides to capital case, and lower cases are needed to implement polymorphic nucleotides including gaps (see Bandelt & Dür 2007). This behaviour is good to minimize unintentional errors, but doesn't allow including polymorphic nucleotides including gaps (i.e., that need to be in lower case). Please contact me is you would like to do so.

Important notes

Distance precision

The maximum length of each distance output by the program is set to "8" by default. If you want to modify this number (e.g., increase the number of decimals), change the '8' for the number you want at line 91 of the file pofad.h:

```
#define PRECISION 8
```

Then recompile the program.

Reporting Bugs

Please, report any suspicious behaviour of POFAD. Send an Email to Simon Joly: joly.simon@gmail.com. By doing so, you will also help other users of POFAD.

References

Bandelt H-J, Dür A (2007) Translating DNA data tables into quasi-median networks for parsimony analysis and error detection. *Molecular Phylogenetics and Evolution*, **42**, 256–271.

Göker M, Grimm GW (2008) General functions to transform associate data to host data, and their use in phylogenetic inference from sequences with intra-individual variability. *BMC Evolutionary Biology*, **8**, 86.

Joly S, Bruneau A (2006) Incorporating allelic variation for reconstructing the evolutionary history of organisms from multiple genes: an example from Rosa in North America. *Systematic Biology*, **55**, 623–636.

Joly S, Bryant DJ, Lockhart PJ (2014) Flexible methods for estimating genetic distances from nucleotide data. *bioRxiv*. doi: http://dx.doi.org/10.1101/004184.

Joly S, McLenachan PA, Lockhart PJ (2009) A statistical approach for distinguishing hybridization and incomplete lineage sorting. *The American Naturalist*, **174**, e54–e70.

Jukes TH, Cantor CR (1969) Evolution of protein molecules. *Mamalian protein metabolism* pp. 21–123. Academic Press, New York.

Landry P-A, Lapointe F-J, Kirsch JAW (1996) Estimating phylogenies from lacunose distance matrices: additive is superior to ultrametric estimation. *Molecular Biology and Evolution*, **13**, 818–823.

Liu L, Yu L, Pearl DK, Edwards SV (2009) Estimating species phylogenies using coalescence times among sequences. *Syst Biol*, **58**, 468–477.

Makarenkov V (2001) T-REX: reconstructing and visualizing phylogenetic trees and reticulation networks. *Bioinformatics*, **17**, 664–668.

Mossel E, Roch S (2010) Incomplete lineage sorting: consistent phylogeny sstimation from multiple loci. *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, **7**, 166–171.

Potts AJ, Hedderson TA, Grimm GW (2014) Constructing Phylogenies in the Presence Of Intra-Individual Site Polymorphisms (2ISPs) with a Focus on the Nuclear Ribosomal Cistron. *Systematic Biology*, **63**, 1–16.