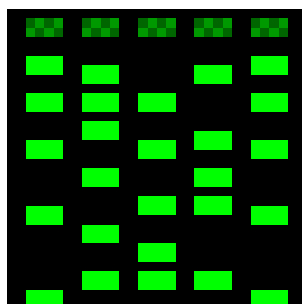


FAMD - Fingerprint Analysis with Missing Data 1.1 β - Manual -

Philipp M. Schlüter

Department of Systematic & Evolutionary Botany
University of Vienna
Austria

March 2006



INDEX

FAMD - Fingerprint Analysis with Missing Data 1.1β Manual

INDEX.....	2
GENERAL.....	4
Purpose.....	4
Author.....	4
Citation.....	4
Licence and Legal Disclaimer.....	4
System requirements.....	4
Source code and compilation.....	5
What's new in FAMD 1.1?.....	5
Limitations and known issues.....	5
Opening the input file.....	5
Groups and LocusSets.....	5
AMOVA.....	5
HOW TO USE.....	6
General directions.....	6
How do I construct a NJ tree?.....	6
DEFAULTS.....	8
Default settings.....	8
Default input format(s).....	9
PROGRAM FUNCTIONS.....	11
File Menu.....	11
Load.....	11
Save DataMatrix.....	12
Select File Names.....	13
Select Analysis Output File.....	13
Select Tree Output File.....	13
Select Consensus Tree Input File.....	13
Select Consensus Tree Output File.....	13
Select Log File.....	13
Export.....	13
Nexus.....	13
Arlequin Project.....	14
Genepop.....	14
NTSys-pc.....	15
List of OTUs.....	15
SynTax.....	15
hindex.....	15
Structure.....	15
Hickory (Nexus).....	15
Turn log on/off.....	15
Exit.....	16
DataMatrix Menu.....	17
Restore Original Matrix.....	17
Matrix Statistics.....	17
Frequency Statistics.....	17
Frequencies per Individuals.....	17
Frequencies per Loci.....	17
Missing Data Statistics.....	18
Replace Missing Data.....	18
Remove Individuals.....	18
With missing data.....	18
With band frequency below.....	18
With band frequency above.....	18
Remove Loci.....	18
With missing data.....	18
Monomorphic.....	18

Monomorphic Absences.....	19
Monomorphic Presences.....	19
With band frequency below.....	19
With band frequency above.....	19
With more missing data than presences	19
Group Manager	19
Defining Groups (groups of individuals)	20
Defining and Selecting LocusSets (groups of loci).....	21
Select Group Only	21
Group-based Profiles	21
Additive Band Profile	21
Fixed Band Profile	22
Analysis Menu.....	23
Standard Similarity	23
Minimum Similarity	24
Maximum Similarity	25
Average Similarity	26
Shannon's Index.....	26
AMOVA.....	27
Trees Menu	28
UPGMA	28
Strict Consensus	28
Majority Rule Consensus	28
Principal Coordinate Analysis	28
Replicate Analyses Menu	30
Bootstrap Shannon's Variance (SH + RMD).....	30
Bootstrap Std Tree (BS + RMD).....	30
Multiple Avg Trees (TR + JC).....	31
Bootstrap Avg Trees (BS + JC).....	31
Estimate R-support (TR)	32
Shannon t-tests	32
Options Menu.....	35
Missing Data Replacement Settings	35
Shannon Scaling.....	35
Replicate Analysis Settings.....	36
MR Consensus and R-support Settings.....	36
Similarity Coefficient Selection	37
Help Menu.....	37
About	37
Help	37

GENERAL

Purpose

This program was written for analysis of RAPD, AFLP or other **dominant** fingerprint data, especially for data sets that contain ambiguous or missing data, so that the impact that missing data may have on the analysis can be better evaluated. The second intention was to provide a means to easily calculate the variance associated with Shannon's index.

Author

Philipp M. Schlüter

Department of Systematic and Evolutionary Botany
Institute of Botany
University of Vienna
Rennweg 14
A-1030 Vienna
Austria

Telephone: +43-1-4277-54149

E-mail: <philipp.maria.schlueter@univie.ac.at>

For any problems or queries about the program please contact me. pms.

Citation

Schlüter, P. M. & Harris, S. A., 2006. Analysis of multilocus fingerprinting data sets containing missing data, *Mol. Ecol. Notes*: doi: 10.1111/j.1471-8286.2006.01225.x.

Licence and Legal Disclaimer

By downloading and using FAMD you accept the following.

FAMD (c) Copyright 2002-2006 Philipp M Schlüter.

This program is distributed in the hope that it will be useful, but WITHOUT ANY WARRANTY, whether expressed or implied; without even the implied warranty of MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE. The author (PMS) will not be liable for any special, incidental, consequential, indirect or similar damages due to loss of data or any other reason, even if he or an agent of his has been advised of the possibility of such damages. In no event shall the author be liable for any damages, regardless of the form of the claim. The person using the software bears all risk as to the quality and performance of the software.

FAMD may be distributed freely for non-commercial purposes provided that the copyright notice is not removed. If you intend to include it in any commercially distributed package, the author should first be contacted. If you wish to modify the source code and rebuild this application so that it fits your own need, you may do so provided you do not remove the original copyright notice and provided you do not intend to distribute it commercially.

System requirements

FAMD was developed for 32-bit Windows operating systems and was tested on Windows XP. The program was developed and tested on an Intel Pentium M system.

FAMD uses floating-point instructions and therefore requires a CPU with a floating-point unit (FPU). FAMD uses instructions such as the CPUID instruction introduced on later 486-processors which therefore represent the minimum processor requirement for FAMD.

Source code and compilation

Source code is available from the author upon request. FAMD was written in Borland Delphi 7. Some routines are based on older ones (Turbo Pascal and Microsoft Macro Assembler). I cannot guarantee that the code will be easy to follow for anybody but you are free to give it a try.

What's new in FAMD 1.1?

Version 1.1 adds a number of features to FAMD. Unlike FAMD 1.0, the program can now...

- read and write unlimited data files
- export data matrices to a number of other file formats
- provide a graphical interface for easily defining groups within a data set
- calculate a number of different similarity/distance measures (Jaccard, Dice, SMC similarities, and NeiLi, Euclidean and Squared Euclidean distances)
- carry out different distance transformations
- carry out AMOVA analysis with all implemented (dis)similarity measures
- construct strict and majority rule consensus trees
- carry out principal coordinate analysis with all implemented (dis)similarity measures
- estimate R_x -support of branches based on the missing data present in the data matrix

Limitations and known issues

Opening the input file

In some cases, it may be difficult to read in an input file generated by e.g. the standard Windows Editor, because it may hide some additional characters in the file that you don't see in the Editor. In such a case, opening the input file fails. The inclusion of remark characters in certain places in the input file, especially if you try to delete a data row by bracketing it with remark characters, may cause the loading of the input file to fail.

Groups and LocusSets

There are limitations in the selection of `Groups` (groups of individuals) and `LocusSets` (groups of loci) at the same time; please see the `Group Manager` section for further details.

AMOVA

The current AMOVA implementation makes certain assumptions about the `Groups` defined in your data. Please see the `AMOVA` section for further details.

HOW TO USE

General directions

To use FAMD:

-) generate an input file that can be read by FAMD
-) load the input file into FAMD
-) specify the file you want your analysis results to be saved to
-) specify the file you want your trees to be saved to
-) if desired, modify your data matrix (`DataMatrix` menu); possibly define groups of individuals or loci using the `DataMatrix` → `Group Manager` function.
-) to change preset parameters, use the `Options` menu
-) carry out your analysis using the `Analysis` or the `Replicate Analyses` menu. The `Analysis` menu writes results to the analysis file you selected and, if you generated a similarity/distance matrix expects you to analyse it further yourself (e.g., tree-building, to be found in the `Trees` menu). The `Replicate Analysis` menu performs more complex analyses that run for some time and handle tree-building themselves.

How do I construct a NJ tree?

Since the current version of FAMD does not implement Neighbour Joining, you need to export your distance matrix to carry out a NJ analysis in another program.

The `Similarity Coefficient Selection` dialogue box in FAMD lets you select the distance transformation you want to have applied to your data. For generating a Neighbour Joining tree from a similarity measure, you would probably wish to use the $d=\sqrt{1-s}$ transformation and CHECK the “write distance matrix to analysis file” checkbox. After calculating a similarity/distance matrix from your data, you can open FAMD’s analysis output file and simply copy the distance matrix you want to use into another file and use this as input data for another program. Some programs will require some additional lines to be added in order to correctly interpret the distance matrix. Should your target file format be nexus, you can simply save your distance matrix to a nexus file using the standard `File` → `Export` → `Nexus` command.

For example, if you wished to use manually generate a nexus file from a FAMD-generated distance matrix for NJ tree generation in PAUP* 4.0b10 for Windows, you could use the distance matrix in FAMD’s analysis output file to generate a Nexus file as follows:

```
#NEXUS

BEGIN TAXA;

DIMENSIONS NTAX= your number of taxa here (click DataMatrix → Matrix Statistics in FAMD to find out);

TAXLABELS

... taxon names here (copy the line with taxon labels immediately preceding the numeric values of the data matrix) ...

;

END;
```

```

BEGIN DISTANCES;
DIMENSIONS NTAX= your number of taxa here (click DataMatrix → Matrix Statistics in FAMD to find out);
FORMAT TRIANGLE=LOWER DIAGONAL NOLABELS;
MATRIX
... distance matrix here (copy the distance matrix without the preceding line of taxon labels) ...
;
END;

```

You can now open and execute this .nex file, set the optimality criterion to distance, set the distance measure to user-defined distance (this is important!), and let PAUP* calculate an NJ tree. The following commands should do:

```

set criterion=distance;
dset distance=user;
nj;
savetrees file= name of tree file brelens=yes;

```

Done!

DEFAULTS

Default settings

Individual names given in input file:	yes
Locus names given in input file:	no
Input file has individuals in:	columns
End of data character:	*
Open remark character:	{
Close remark character:	}
Presence:	1
Ambiguity/missing data point:	?
Absence:	0
Characters ignored in input file:	Blank (00h), whitespace, tab, CR, LF
Include groups:	yes
Delimited data:	yes
Output negative ambiguities	no
JC:	100
TR:	50
BS:	1000
SH:	10000
MDR:	yes
Missing data replacement by presences:	50%
Data removal threshold: MD (individuals):	18%
Data removal threshold: MD (loci):	25%
Data removal threshold: Below Frq (loci)	5%
Data removal threshold: Above Frq (loci)	95%
Data removal threshold: Below Frq (indiv.)	30%
Data removal threshold: Above Frq (indiv.)	70%
Shannon Log Base:	2
Frequencies = presences/locus per:	total presences
Majority Rule consensus percentage	70%
Majority Rule consensus for Rx analysis	100%
Rx Consensus Threshold r=	10000
Values of r to be analysed	1 - 10000
Desired precision for R values	1 digit
(Dis)Similarity Coefficient	Jaccard
Distance Transformation	d = 1-s
Write similarities to analysis file	yes
Write distances to analysis file	no
Nei-Li R-value	6.00
Similarity Mode Preference	Standard
Log to file:	no
Analysis output file name:	analysis.txt
Tree output file name:	outtree.ph
Consensus tree input file:	outtree.ph
Consensus tree output file:	constree.ph
Log file name:	famdlog.txt

Default input format(s)

FAMD is quite flexible as regards input files. The default input format is a text file containing a simple data matrix whose end is indicated by a single user-defined `EndOfData` character (default: '*'). The file can contain labels for individuals and/or loci and the data matrix can be in either orientation, individuals in columns or individuals in rows. Finally, data can be delimited by space, tabs, etc., or undelimited. When opening an input file, FAMD will ask you about these things.

There are, however, some limitations: Please avoid empty lines in the file containing your data matrix. Locus and individual names should be UNIQUE names and should NOT contain spaces, the data file should not contain the characters '<' and '>'. The program will not check whether the names you provide are unique, but since some of the internal data handling makes the implicit assumption of unique names, FAMD may behave unexpectedly and possibly provide erroneous results if you violate this assumption.

FAMD does allow for remarks in the input file, i.e., information bracketed by the `OpenRemark` and `CloseRemark` characters (defaults are '{' and '}') is intended solely for your information and is not read in. However, any combination of `OpenRemark` and `CloseRemark` characters must be WITHIN a SINGLE line, and these characters MUST NOT bracket an entire line. Otherwise the input may not be read in correctly and/or FAMD may crash.

The input format is like this:

```
      [Ind1 Ind2 Ind3 ... IndN]
[Loc1]  1    0    1    ...  0
[Loc2]  ?    0    ?    ...  0
  ..    ..    ..    ..    ..
[LocM]  1    0    0    ...  1
*
```

Items in [] can be left out and blanks/tabs separating the data points need not be present.

The asterisk (`EndOfData` character) at the end of the data block is compulsory. Blanks or tabs can delimit the entries; the program is fairly flexible regarding that. You can also swap individuals and loci (i.e. rows and columns) - you just need to set the input file parameters accordingly. You can also choose the characters you use to represent presences, absences and missing data and you can re-save modified data matrices, changing any of those parameters.

Examples for input files are given below:

Example 1 - Delimited data; individuals in columns; individual names present; no locus names present. (This is the default input file format).

```
IndA IndB IndC IndD IndE IndF
1     0     1     1     ?     0
?     0     ?     1     0     0
1     1     0     0     1     1
1     ?     1     1     1     1
*
```

Example 2 - Delimited data; individuals in columns; individual names present; locus names present.

	IndA	IndB	IndC	IndD	IndE	IndF
AA01	1	0	1	1	?	0
AA02	?	0	?	1	0	0
AA03	1	1	0	0	1	1
AA04	1	?	1	1	1	1

*

Example 3 - Delimited data; individuals in rows; individual names present; no locus names present.

```
IndA 1 ? 1 1
IndB 0 0 1 ?
IndC 1 ? 0 1
IndD 1 1 0 1
IndE ? 0 1 1
IndF 0 0 1 1
*
```

Example 4 - Undelimited data; individuals in rows; individual names present; locus names present.

```
AA01 AA02 AA03 AA04
IndA 1?11
IndB 001?
IndC 1?01
IndD 1101
IndE ?011
IndF 0011
*
```

You may put anything after the `EndOfData` character; FAMM doesn't care. However, FAMM does allow to add additional information blocks in an input file AFTER the data matrix/`EndOfData` character. For instance, a block containing information about sample grouping (such blocks may be generated using the `Group Manager` function; see this section for further information). Every block begins with a its block name, contained in square brackets and ends with an asterisk (*). Currently, there are two block types supported, a `Groups` and a `LocusSets` block. Briefly, such blocks in the input file would look like the following:

```
[Groups]
AllData/East= A, C, D, E;
AllData/West= F, G, H, I, J;
AllData/East/PopEast1= A, C, D;
*

[LocusSets]
ExcludeLoc3= Loc1, Loc2, Loc4, Loc5, Loc6;
*
```

PROGRAM FUNCTIONS

File Menu

Load

Displays a file-open dialogue box which lets you select a file to open. For instructions of the supported file formats, please see the `Default input format(s)` section in this manual. The File Parameter Selection box is displayed which lets you specify which characters are used for which purpose in the file:

Individuals in...
 columns Data points in the file are read such that each line represents one locus.
 rows Data points in the file are read such that each line represents one individual.

Header presence for
 individuals If this option is checked, individual labels are assumed to be present. If this is not set, labels will be given as Ind01, Ind02, etc.
 loci If this option is checked, locus labels are assumed to be present. If this is not set, labels will be given as Loc001, Loc002, etc.

Delimited data
 If this option is checked, the program assumes that individual data points in your input data matrix are separated by delimiting characters, such as space or tabs. Unchecking this option essentially allows the input of rows of data points (either loci or individuals) without any delimiting characters. Characters specifying data points MUST be single characters. See the examples in the `Default input format(s)` for the difference between delimited and undelimited data.

Include groups
 The default value is yes (checkbox is checked). This tells FAMD to scan the input file for any information on sample of locus groups defined that may be present in the input file AFTER the EndOfData character (default character: '*'). For more details, please see the `Default input format(s)` and `Group Manager` sections in this manual.

Characters:

These need to be single characters. Characters '<' and '>' should not be used because they are used internally in the program.

Presence: Default value='1'.
 Specifies a band presence.
Absence: Default value='0'.
 Specifies a band absence.
Ambiguity: Default value='?'.
 Specifies a missing data point (e.g., an unclear band).
OpenRemark: Default value='{'.
 After this character, text or numbers will be treated as a remark.
CloseRemark: Default value='}'.
 Defines the end of a remark.
EndOfData: Default value='*'.
 This character represents the end of data in the file. Any data after this character will be ignored.

Note: For limitations in the use of remarks, see the `Default input format(s)` section.

Save DataMatrix

This lets you re-save your original data matrix. In the process, you can change the characters used to specify e.g. band absences and presences, or you can swap data rows and columns (individuals and loci). Remarks that might have been present in the original data file will be lost, as will be all data after the `EndOfData` character (except for `Groups` and `LocusSets` defined, if there are any). If you have removed individuals or loci from the data set, you can use this option to save a sub-set of your original data set.

Use the File Parameter Selection box to specify which characters are used for which purpose in the file:

Individuals in...
 columns Data points in the file are read such that each line represents one locus.
 rows Data points in the file are read such that each line represents one individual.

Header presence for
 individuals If this option is checked, individual labels are assumed to be present. If this is not set, labels will be given as Ind01, Ind02, etc.
 loci If this option is checked, locus labels are assumed to be present. If this is not set, labels will be given as Loc001, Loc002, etc.

Delimited data
 If this option is checked, the program will save your data matrix as delimited data. For more details, please see the `Default input format(s)`.

Include groups
 If this option is checked, FAMMD will save `Groups` and `LocusSets` blocks to the data file, if groups or `LocusSets` have been defined. For more details, please see the `Default input format(s)` and `Group Manager` sections in this manual.

Ambiguity negative
 Checking this option will save missing data values preceded by a minus (“-”) sign.

Characters:

These need to be SINGLE characters. Characters ‘<’ and ‘>’ should not be used because they are used internally in the program.

Presence: Default value='1'.
 Specifies a band presence.
Absence: Default value='0'.
 Specifies a band absence.
Ambiguity: Default value='?'.
 Specifies an unclear band; missing data point.
OpenRemark: Default value='{'.
 After this character, text or numbers will be treated as a remark.
CloseRemark: Default value='}'.
 Defines the end of a remark.
EndOfData: Default value='*'.
 This character represents the end of data in the file. Any data after this character will be ignored.

Select File Names

Select Analysis Output File

This lets you select a file to which numerical results can be saved. Note that by selecting a file you do not actually save to this file. If you do not select a file here, FAMd will use the default file name `analysis.txt` in the current directory.

Select Tree Output File

This lets you select a file to which trees generated during the analysis will be saved. Note that by selecting a file you do not actually write to this file. If you do not select a file name here, FAMd will use the default file name `outtree.ph` in the current directory. Trees will be saved in PHYLIP format. Tree names will be stored as root labels. Note that not every program may be able to interpret this correctly. You can use programs such as TreeView to open the tree file and convert trees to another format if desired, and in doing so choose to remove branch labels (i.e. tree names).

Select Consensus Tree Input File

This lets you select a file to which contains the trees that will be used as input information for strict and majority rule consensus trees. If you do not select a file name here, FAMd will use the default file name `outtree.ph` (the default tree output file) in the current directory. FAMd expects input trees to be in PHYLIP format.

Select Consensus Tree Output File

This lets you select a file to which consensus trees will be saved. Note that by selecting a file you do not actually write to this file. If you do not select a file name here, FAMd will use the default file name `constree.ph` in the current directory. Consensus trees will be saved in PHYLIP format with the percentages of cluster occurrences saved as branch labels, and not as branch lengths. You can use programs such as TreeView to open the consensus tree file and convert trees to another format if desired.

Select Log File

This lets you select a file to which FAMd - if told to do so - will log all commands it is told to carry out until logging is stopped. If you do not select a file here, FAMd will use the default file name `famdlog.txt` in the current directory.

Export

This function lets you export your data matrix into a number of file formats that can be read by other analysis program. Please note that support for other file formats cannot be expected to be complete and have only been tested to work with particular versions of targeted programs. However, the different export filters can make your life a lot easier.

Nexus

Exports your data matrix to Nexus format. Programs that have been tested to read exported nexus files are PAUP* 4.0 beta 10 (Windows), MrBayes 3.1 (Windows) and Splitstree 4.2 (Windows). The nexus format comes in two flavours ('old' and 'new' nexus format) and some programs may be able to read one, but not the other of these.

FAMd will export your data matrix as space-delimited data (setting the Nexus INTERLEAVE option set), writing a single row of data per individual. The missing data character will be that defined for FAMd. FAMd will write a TAXA block, and will ask you whether to export your data in a CHARACTERS ('new' nexus, recommended for PAUP*) or DATA block ('old' nexus, recommended for MrBayes). In addition, it will ask you what nexus DATATYPE your data should be assigned to. The available options are: STANDARD and RESTRICTION. STANDARD should be used for use with e.g.,

PAUP*. `RESTRICTION` makes more sense for AFLP data, but is not recognised by PAUP*. MrBayes does recognise this data type.

If you have already generated a similarity/distance matrix in FAMD, the program will ask you whether to a distance matrix to the nexus file as `DISTANCES` block. The distance transformation used as defined under FAMD's Options→Similarity Coefficient Selection dialogue.

Arlequin Project

This option lets you save Arlequin Project (ARP) files. This export filter was tested for use with Arlequin 3.0 (Windows). For this option to work, groups of individuals must be defined (see `Group Manager` section). FAMD does not currently save distance matrices for use with Arlequin.

Depending upon how many hierarchical levels of group you have defined for your data, FAMD may ask you whether to use either the first or the first and the second level of group hierarchy for generating an ARP file. Note that FAMD will export ALL groups at a given level whether or not such groups may 'contradict' each other. FAMD will export all individuals in your current data matrix as haplotypes, even if some of them are not present in any of the exported groups (which should not upset Arlequin). It is the user's responsibility to make sure that the group structure, as exported by FAMD, actually makes sense. FAMD will not check for e.g., individuals which are members of more than one groups (which may be done in FAMD but does not make sense in the context of some of the things you might want to use Arlequin for).

Genepop

This option tells FAMD to generate a genepop-format file. This option was tested using BAPS 3.2 (Windows). Note that FAMD will encode your dominant data as diploid data, the first allele taken from your data matrix, and the second typically encoded as missing data. Since for dominant data, a band absence may be interpreted as a 0/0 genotype, FAMD will ask you whether you wish absences to be treated thus. Please note, however, that most software manuals suggest simply encoding the second allele as missing data. Using this export filter, presences are encoded as 01, absences as 02 and missing data as 00.

The genepop file generated by FAMD does NOT contain any information about populations, i.e., your data matrix is exported as a single population. If you wish to define additional populations, please edit the FAMD-generated text file, introducing population structures manually using the term '`POP`'. An example is provided below

The following is an example of a FAMD-generated genepop-format file with all individuals (IndivA, IndivB, IndivC and IndivD) in one population.

```
FAMD exported data
Locus1, Locus2, Locus3, Locus4, Locus5
POP
IndivA, 0100 0100 0200 0100 0100
IndivB, 0100 0100 0200 0100 0100
IndivC, 0100 0100 0200 0100 0100
IndivD, 0100 0100 0200 0100 0100
```

Suppose you want to have two populations, one consisting of IndivA and Indiv B, and the second one consisting of IndivC and IndivD. To do so, you simply introduce the term '`POP`' before the first individual which should belong to population 2 (i.e., IndivC). Your modified genepop file will look as follows:

```
FAMD exported data
Locus1, Locus2, Locus3, Locus4, Locus5
POP
IndivA, 0100 0100 0200 0100 0100
IndivB, 0100 0100 0200 0100 0100
POP
IndivC, 0100 0100 0200 0100 0100
IndivD, 0100 0100 0200 0100 0100
```

NTSys-pc

Using this option, FAMd will write a text file (.NTS) that can be read by NTSYS-pc. This function was tested using NTSYS-pc 2.1 (Windows). FAMd will output labels for loci and individuals. Presences, absences and missing data will be encoded as 1, 0, and -9, respectively.

List of OTUs

This option simply generates a text file containing the names of individuals in your data matrix. The file will not contain the data matrix itself.

SynTax

Use this option to write files that can be read by SYN-TAX. This option was tested using SYN-TAX 2000. Three files will be generated. The first one (default name `syntax.txt`) will contain your data matrix. The other two files, whose name will be based on the name of your syntax file name, will contain the individual and locus label files for use with SYN-TAX. Their default file names are `syntax_indivlabels.txt` and `syntax_rowlabels.txt`, respectively.

hindex

Use this option to generate an input file for hindex. This option was tested using hindex 1.42 (Linux). For this option to work, groups of individuals must be defined (see *Group Manager* section). FAMd will ask you to select two groups that act as parent 1 and parent 2 for the third group, which is the group of individuals that is to be tested using hindex. Two files will be generated, one `filename_indiv.txt` containing the individuals to be tested. Missing data will be encoded 'NA' in this file. The second file, `filename_parents.txt`, will contain information about the parents, i.e. the frequency of a marker at a given locus in parent 1 and parent 2. Marker frequencies are calculated, treating missing data as defined under the *Options* menu. If the *RMD* (replace missing data) variable under *Options*→*Replicate Analyses Settings* is set, missing data will be replaced according to *Options*→*Missing Data Replacement Settings* before marker frequencies are calculated. If *RMD* is not set, then missing data will be ignored for parents 1 and 2.

Structure

This option allows you to generate an input file for Structure. It was tested using Structure 2.0. FAMd's export filter will ask for the ploidy level of your data. If you enter a value greater than one, your data matrix will be exported, the first allele taken from your data matrix, and the other alleles typically encoded as missing data. Since for dominant data, a band absence may be interpreted as a 0/0 genotype, FAMd will ask you whether you wish absences to be treated thus. Please note, however, that most software manuals suggest simply encoding the additional alleles as missing data. If you have groups defined (see *Group Manager* section), FAMd will ask you whether this information should be exported as population information for use with Structure. Since FAMd allows an individual to occur in more than one group, an individual will be encoded as belonging to the FIRST group it appears in. Structure populations will be consecutively numbered, starting from 1, depending on the order of appearance of groups in FAMd's *Group Manager*. If an individual is not assigned to any group, it will be assigned to Structure's 'Population 0'.

Hickory (Nexus)

To use this option, you need to have groups defined (see *Group Manager* section). This option will generate a nexus (.NEX) file for use with Hickory. The nexus file generated will contain a *TAXA* and an *ALLELES* block.

Turn log on/off

This option tells FAMd to turn logging on or off. If enabled, FAMd will log all commands it carries out to a log file (default is `famdlog.txt`). This is useful for purposes of documenting what you are doing. The messages saved to the log files are identical to those displayed on the screen.

Exit

Quits the program. FAMD will not ask you to save data prior to quitting.

DataMatrix Menu

The DataMatrix menu lets you view and modify different aspects of your data. Generally, results are displayed on the screen and you are asked whether you would like to save them to the analysis file. If this file already exists you are asked whether data should be appended to it or whether the file should be overwritten.

Restore Original Matrix

This option lets you restore the original data matrix as first loaded when an input file was opened. It is useful, e.g. after data points have been removed from or replaced in the data matrix.

Matrix Statistics

This option tells you how many individuals and loci were detected in your input file. It is often useful to use this option to check that the program loads exactly the data set you wanted it to load, or that the data set was loaded completely. In addition to the number of individuals and loci, also the size of the data matrix (i.e. number of data points, given by #individuals × #loci) and the amount of missing data in the data set are displayed.

Frequency Statistics

Frequencies per Individuals

This will save the frequency $p(i)$ of band presences in each individual i to the analysis output file.

$$p(i) = \frac{N_1(i)}{L} \quad \text{where: } N_1(i) \text{ is the number of band presences (1) in locus } i$$

L is the number of loci

Frequencies per Loci

This will save the frequency $p(i)$ of band presences in each locus i to the analysis output file. You are first asked how you want frequencies to be defined. The first option is:

$$p(i) = \frac{N_1(i)}{n} \quad \text{where: } N_1(i) \text{ is the number of band presences (1) in locus } i$$

n is the number of individuals

This is intuitive. The second option is:

$$p(i) = \frac{N_1(i)}{\sum_{k=1}^s N_1(k)}$$

where: $N_1(i)$ is the number of band presences (1) in locus i

s is the number of loci in the data set

$\sum N_1(i)$ is the total number of presences in the data set

The reason for giving this option of calculating frequencies lies in the formula by Bowman *et al.* (1969) for calculating the variance associated with Shannon's index.

Reference

Bowman, K. O., Hutcheson, K., Odum, E. P. and Shenton, L. R., 1969, Comments on the distribution of indices of diversity, *Proc. Intl. Symp. Stat. Ecol.* **3**: 315-359.

Missing Data Statistics

This will write information about missing data to your analysis output file. The output values are the percentages of missing data points in each individual and in each locus.

Replace Missing Data

Selecting this option replaces missing data in your data set according to the settings specified under Options → Missing Data Replacement Settings. By doing so, the data matrix is modified and missing data in it replaced by discrete characters. Therefore, if you wish to proceed with the analysis using the original data matrix, you must first restore it using DataMatrix → Restore Original Matrix.

Remove Individuals...

This option removes individuals from the data matrix that match the criteria applied. If you wish to undo such a data removal to continue with the original data matrix, you must restore it using DataMatrix → Restore Original Matrix.

With missing data...

This will remove individuals that have a percentage of missing data that is greater than the specified threshold percentage.

With band frequency below...

This will remove individuals whose frequency of band presences is below the specified threshold percentage. This may be useful e.g. to remove individuals whose data stems from poor AFLP reactions which (if they were scored) will often display a strikingly lower number of band presences.

With band frequency above...

This will remove individuals whose frequency of band presences is greater than the specified threshold percentage.

Remove Loci...

This option removes loci from the data matrix according to your choice of options. If you wish to undo such a data removal to continue with the original data matrix, you must restore it using DataMatrix → Restore Original Matrix.

With missing data...

Removes loci that have a percentage of missing data that is greater than the specified threshold percentage.

Monomorphic...

Removes monomorphic loci from the data set. This should be done for instance prior to calculation of Shannon's index (non-bootstrapping version). A locus with a monomorphic presence or absence is defined as follows:

Monomorphic presence: IF $(m \leq o)$ AND $(z=0)$.

Monomorphic absence: IF $(m \leq z)$ AND $(o=0)$.

Where $m = N_2(i)$, the number of missing/ambiguous data in locus i
 $o = N_0(i)$, the number of band absences in locus i
 $z = N_1(i)$, the number of band presences in locus i

All other bands are considered polymorphic.

Monomorphic Absences...

Removes monomorphic absences from the data set, where a band is considered to be a monomorphic absence

IF $(m \leq z)$ AND $(o=0)$.

Where $m = N_2(i)$, the number of missing/ambiguous data in locus i
 $o = N_0(i)$, the number of band absences in locus i
 $z = N_1(i)$, the number of band presences in locus i

Monomorphic Presences...

Removes monomorphic presences from the data set, where a band is considered to be a monomorphic presence

IF $(m \leq o)$ AND $(z=0)$.

Where $m = N_2(i)$, the number of missing/ambiguous data in locus i
 $o = N_0(i)$, the number of band absences in locus i
 $z = N_1(i)$, the number of band presences in locus i

With band frequency below...

This will remove loci whose frequency of band presences is below the specified threshold percentage.

With band frequency above...

This will remove loci whose frequency of band presences is greater than the specified threshold percentage.

With more missing data than presences

This will remove loci if the number of missing data points is greater than the number of band presences.

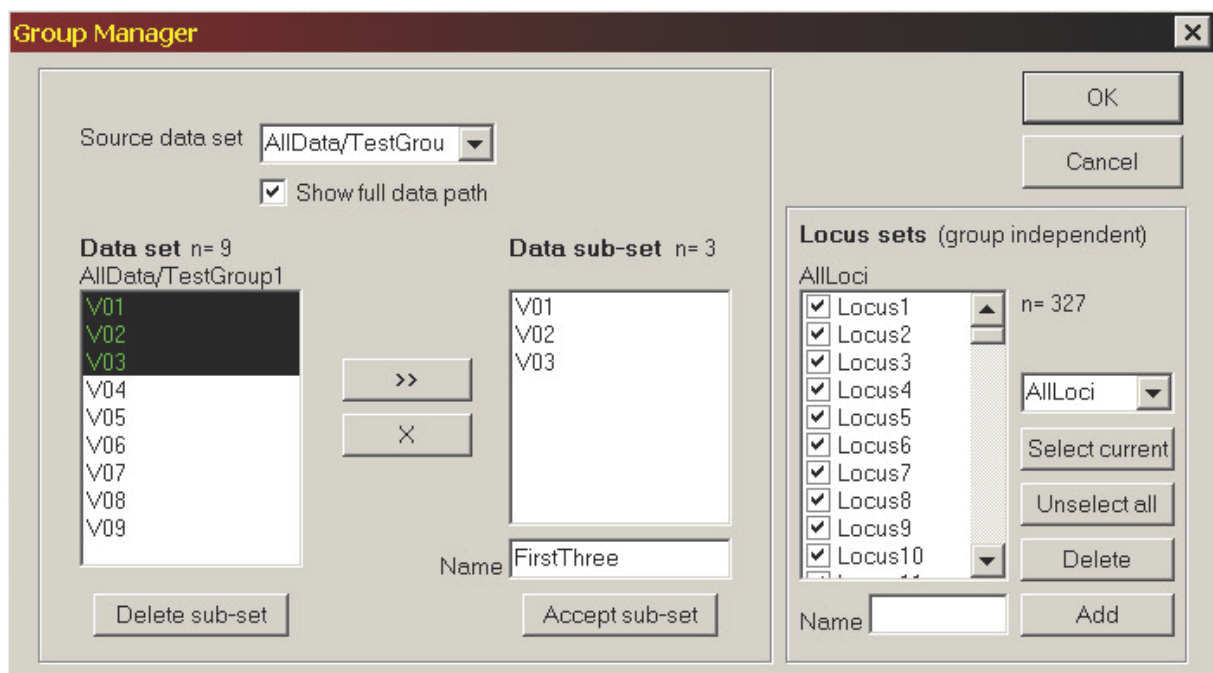
Group Manager

The group manager serves 3 functions:

- i) to define groups of individuals (herein termed 'Groups')
- ii) to define groups of loci (herein termed 'LocusSets')
- iii) to allow the selection of a LocusSet

If Group or LocusSet information was read in from the input data file, then these groups will appear be available in the Group Manager. Groups of samples can then be selected using the `DataMatrix → Select Group Only → [Group Name]` command. Support for sample or locus groups is in many ways not yet satisfactory and may therefore be subject to change in future versions of FAMd.

To aid description of the Group Manager, a screen-shot is provided below.



Defining Groups (groups of individuals)

The left side of the Group Manager dialogue box lets you define groups. This allows for selections to be made and enables the comparisons of e.g., different populations. The topmost control lets you select a source data set. If you have no groups defined yet, this source data set will be “AllData”. This is a built-in group that cannot be modified. It is equivalent to all individuals in the current data set. Groups of individuals are hierarchical, with “AllData” acting as the base group. By contrast, none of the selections made using a the Group Manager affect “AllData”.

You can select a group that is a subset of a given group (AllData or any of the groups you have defined) by selecting that group in the topmost control titled “**Source data set**”. All individuals contained in this group are then displayed in the leftmost list, titled “**Data set**”. You can select one or multiple individuals in this list, press the “>>” button, and it/they will appear in the “**Data sub-set**” list. To remove an individual from that list, select it and press the “X” button. If the “**Data sub-set**” list contains those individuals that you wish to be included in your group, enter a name for the group in the empty field titled “**Name**” below the “**Data sub-set**” list and click on the “**Accept sub-set**” button. The “**Data sub-set**” list will then be cleared, and the newly defined group available in the “**Source data set**” control. Groups will appear in that control with their name indicating their placement in the hierarchy in the form of a path name. For instance, if you define a group named, “FirstGroup”, it will have a complete path name of “AllData/FirstGroup”. The same name will also be displayed in other FAMD functions using groups. You can choose to display only the last part of the group name (i.e., the bit you entered) in the Group Manager (and other functions in FAMD) by unchecking the “**Show full data path**” checkbox under the “**Source data set**” control.

To delete a group from the group manager, select the group using the “**Source data set**” control and then press “**Delete sub-set**”.

NB: Your selections will only be kept if you press “**OK**” to quit the Group Manager, clicking “**Cancel**” will discard all changes you made. Please also note that FAMD will NOT check whether any of your groups contains one individual several times (although this may result in errors later on).

Note that, “AllData” may or may not be equivalent to the data file initially loaded: Any loci or individuals removed by any functions of the DataMatrix submenu are not present in “AllData”. (If necessary, you can restore the original data matrix via the DataMatrix → Restore Original

Matrix command.) Please also note that when saving a file including group information, “AllData” is NOT saved.

Defining and Selecting LocusSets (groups of loci)

The right side of the Group Manager dialogue box lets you define `LocusSets`, that is, groups of loci. This allows the inclusion/exclusion of loci from the data set. In contrast to groups of individuals, `LocusSets` are not hierarchical but simply reflect selections of certain loci that can be stored and loaded. The rightmost list is a list of all loci in the data set. “AllLoci” is a predefined `LocusSet` that cannot be modified; it selects all loci in the data set. You can unselect a single locus by unchecking it, or unselect all loci by clicking on the “**Unselect all**” button. If you are satisfied with your selection, enter a name in the field titled “**Name**” and click on the “**Add**” button. The locus set should now be selectable from the combo box above the “**Select current**” button. To delete a `LocusSet`, select it using that same combo box and click on the “**Delete**” button.

To select a `LocusSet`, select it from the combo box above the “**Select current**” button and then click on the “**Select current**” button. This will exclude unselected loci from any analysis you perform subsequently (until you again select the “AllLoci” `LocusSet`).

Limitation.

Note that in the current version of FAMD, `LocusSets` can only be changed when the currently selected group is “AllData”. However, it is possible to select different groups after a `LocusSet` has been selected.

Select Group Only

This option lets you select a group that you have either previously defined using the `Group Manager` or that was read in with the input file with the effect that subsequent analyses will be restricted to the currently selected group of individuals. The `AllData` group refers to all the individuals present in your current data matrix. Note, however, that individuals or loci that have been removed using the `DataMatrix → Remove Loci` or `Remove Individuals` functions have been excluded also from the `AllData` data structure and that once missing data has been replaced using the `DataMatrix → Replace Missing Data` functions is not restored by selecting `AllData`. To restore a data matrix in its entirety, use the `DataMatrix → Restore Original Matrix` function.

Limitations. (FAMD version 1.1β).

A) It is currently not possible to first select a `Group` and then a `LocusSet`. If you wish to use `LocusSets`, you must select your current locus set BEFORE selecting any groups.

B) Groups that can be selected should appear as sub-menus of `DataMatrix → Select Group Only`. However, they may only appear as submenu items AFTER the `Group Manager` has been started once (irrespective of whether you actually did anything in the `Group Manager` or not).

Group-based Profiles

To use this function, groups must be defined (see `Group Manager` section). Band profiles will be generated as detailed below and every defined group (irrespective of its hierarchical level) will be represented as one ‘pooled’ individual whose genotype is the band profile calculated from the information of all individuals in the group. Note that by using this feature, your data matrix (`AllData`) containing ‘real’ individuals will be overwritten. The original data matrix can be reloaded using the `DataMatrix → Restore Original Matrix` function.

Additive Band Profile

This is (at least theoretically) the *in silico* equivalent of pooling DNA of a group of individuals and subjecting the pooled DNA to a dominant fingerprinting method.

The resulting band profile will have a band presence at a given marker locus if at least one individual of the group has a band presence at this marker locus. The resulting band profile will NOT contain any missing data.

Fixed Band Profile

The resulting band profile will only contain band presences at those loci that are fixed in the group. A band presence will be considered fixed if there are no band absences at the locus and the number of actual band presences is greater than the number of missing data points. The resulting band profile will NOT contain any missing data.

Analysis Menu

The **Analysis** menu lets you perform different calculations on your current data set, such as different calculation of similarity/distance matrices which can then be subjected, e.g. to UPGMA tree reconstruction in the **Trees** → **UPGMA** menu. To control which (dis)similarity measures and, if applicable, distance transformations are being used, use **Options** → **Similarity Coefficient Selection**.

Standard Similarity

This calculates a similarity (or distance) matrix from your current data matrix, based on the coefficient selected under **Options** → **Similarity Coefficient Selection**. The same dialogue lets you check options to save the similarity matrix - and/or the matrix after the distance transformation has been applied - to the analysis output file.

The default coefficient selected is Jaccard's similarity coefficient.

Jaccard's similarity coefficient for a pair of individuals i and j is defined as:

Standard Jaccard:

$$S_{ij,Jaccard} = \frac{n_{11}}{n_{11} + n_{01} + n_{10}}$$

where n_{xy} is the number of characters that have state x in individual i and state y in individual j . Possible character states are band presence (1), band absence (0) and missing data (?).

Similarly, other the similarity coefficients are defined as follows.

Standard Dice/Sørensen:

$$S_{ij,Dice} = \frac{2n_{11}}{2n_{11} + n_{01} + n_{10}}$$

Standard SMC (Simple Matching Coefficient):

$$S_{ij,SMC} = \frac{n_{11} + n_{00}}{n_{11} + n_{01} + n_{10} + n_{00}}$$

The following are distance measures rather than similarity measures:

Nei-Li distance (following Nei & Li, 1979) for restriction-site data:

$$d_{ij,NeiLi} = -\frac{3}{2} \ln \frac{4(\frac{2n_{11}}{2n_{11} + n_{10} + n_{01}})^{1/2r} - 1}{3} = -\frac{3}{2} \ln \frac{4(s_{ij,Dice})^{1/2r} - 1}{3}$$

where r is the length of the restriction enzyme's recognition sequence, e.g. for *EcoRI*, recognising GAATTC, r would be 6.

NB: The implementation in PAUP* 4.0 beta 10 assumes that $r = 6$ (J. Wilgenbusch, pers. comm.). For comparability, this is also the default value for r in FAMD, but you can change it using the **Similarity Coefficient Selection** dialogue box.

Standard Euclidean distance:

$$d_{ij, Euclid} = \sqrt[2]{n_{10} + n_{01}}$$

Standard Squared Euclidean distance:

$$d_{ij, SqEuclid} = d_{ij, Euclid}^2 = n_{10} + n_{01}$$

References

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.

Minimum Similarity

This calculates a minimum similarity (or maximum distance) matrix from your current data matrix, based on the coefficient selected under Options → Similarity Coefficient Selection. The same dialogue lets you check options to save the similarity matrix - and/or the matrix after the distance transformation has been applied - to the analysis output file.

The default coefficient used the minimum Jaccard's coefficient, taking into account missing data; for a pair of individuals i and j , it is defined as:

Minimum Jaccard:

$$s_{ij, \min} = \frac{n_{11}}{n_{11} + n_{01} + n_{10} + n_{21} + n_{12} + n_{20} + n_{02}}$$

where n_{xy} is the number of characters that have state x in individual i and state y in individual j . Possible character states are band presence (1), band absence (0) and missing data (?).

If the data set does not contain missing data, the minimum Jaccard value will be identical to the standard Jaccard's coefficient.

Similarly, other the minimum similarity coefficients are defined as follows:

Minimum Dice:

$$s_{ij, \min} = \frac{2n_{11}}{2n_{11} + n_{01} + n_{10} + n_{12} + n_{21} + n_{02} + n_{20}}$$

Minimum SMC:

$$s_{ij, \min} = \frac{n_{11} + n_{00}}{n_{11} + n_{01} + n_{10} + n_{12} + n_{21} + n_{02} + n_{20} + n_{00}}$$

The maximum distance coefficients are defined as follows:

Maximum Nei-Li distance:

$$d_{ij,\max} = -\frac{3}{2} \ln \frac{4(s_{ij,Dice,\min})^{1/2r} - 1}{3}$$

where r is the restriction enzyme's recognition site length

Maximum Euclidean distance:

$$d_{ij,\max} = \sqrt[2]{n_{10} + n_{01} + n_{1?} + n_{?1} + n_{0?} + n_{?0}}$$

Maximum squared Euclidean distance:

$$d_{ij,\max}^2 = n_{10} + n_{01} + n_{1?} + n_{?1} + n_{0?} + n_{?0}$$

Maximum Similarity

This calculates a maximum similarity (or minimum distance) matrix from your current data matrix, based on the coefficient selected under Options → Similarity Coefficient Selection. That same dialogue lets you check options to save the similarity matrix - and/or the matrix after the distance transformation has been applied - to the analysis output file.

The default coefficient used the maximum Jaccard's coefficient, taking into account missing data; for a pair of individuals i and j , it is defined as:

Maximum Jaccard:

$$s_{ij,\max} = \frac{n_{11} + n_{?1} + n_{1?} + n_{??}}{n_{11} + n_{01} + n_{10} + n_{?1} + n_{1?} + n_{??}}$$

where n_{xy} is the number of characters that have state x in individual i and state y in individual j . Possible character states are band presence (1), band absence (0) and missing data (?).

If the data set does not contain missing data, the maximum Jaccard value will be identical to the standard Jaccard's coefficient.

Similarly, other the maximum similarity coefficients are defined as follows:

Maximum Dice:

$$s_{ij,\max} = \frac{2(n_{11} + n_{1?} + n_{?1} + n_{??})}{2(n_{11} + n_{1?} + n_{?1} + n_{??}) + n_{01} + n_{10}}$$

Maximum SMC:

$$s_{ij,\max} = \frac{n_{11} + n_{1?} + n_{?1} + n_{0?} + n_{?0} + n_{??} + n_{00}}{n_{11} + n_{01} + n_{10} + n_{1?} + n_{?1} + n_{0?} + n_{?0} + n_{??} + n_{00}}$$

The minimum distance coefficients are defined as follows:

Minimum Nei-Li distance:

$$d_{ij,\min} = -\frac{3}{2} \ln \frac{4(s_{ij,Dice,\max})^{1/2r} - 1}{3}$$

where r is the restriction enzyme's recognition site length

The standard Euclidean and squared Euclidean distances already represent the minimum possible value for these distances, so the minimum [squared] Euclidean measure is equivalent to the standard [squared] Euclidean distance measure.

Average Similarity

This function calculates a similarity matrix from your current data matrix based on the average similarity s_{ij}^* (or distance d_{ij}^*) coefficient you have selected under `Options → Similarity Coefficient Selection`. For this, minimum ($s_{ij,\min}$) and maximum ($s_{ij,\max}$) similarity coefficients are calculated. The average similarity coefficient is defined as the arithmetic average of values drawn randomly (uniformly) from the interval $[s_{ij,\min}; s_{ij,\max}]$. The number of random draws, JC can be defined in the `Options → Replicate Analysis Settings` dialogue box. In addition to the average similarity value, the associated variance and standard deviation (square root of variance) is calculated and output to the analysis file (provided the respective option is turned on). If the data set does not contain missing data, the average Jaccard value will be identical to the standard similarity coefficient and variance and standard deviation will be zero. The same procedure is followed also if the selected coefficient is a distance rather than a similarity.

Shannon's Index

This function calculates Shannon's index and its variance from your current data set. You should always first remove monomorphic loci manually from the data matrix, using the respective function in the `DataMatrix` menu. This is because depending on your definition of band frequencies, monomorphic loci may still contribute to the sum calculated (although they aren't supposed to), which will artificially change Shannon's index. FAMD 1.1 will, however, warn you if your data matrix still contains monomorphic bands. Note that monomorphic bands, if removed will not be present in the data matrix after this function has completed the calculation of Shannon's index.

Shannon's index is defined as:

$$I \approx -\sum_{i=1}^s p_i \log_2 p_i$$

where I is Shannon's index (often also referred to as H_{Sh})
 p_i is the frequency of band presences in locus i , as defined as in the `Shannon Scaling...` dialogue box
 s is the number of loci
 $\log_2 x = \lg x$ is the logarithm to base 2.

The associated variance is calculated using the formula of Bowman *et al.* (1969):

$$\text{var}(I) \approx \frac{\sum_{i=1}^s p_i \log_2^2 p_i - (\sum_{i=1}^s p_i \log_2 p_i)^2}{n} + \frac{s-1}{2n^2} = \frac{\sum_{i=1}^s p_i \log_2^2 p_i - I^2}{n} + \frac{s-1}{2n^2}$$

where I is Shannon's index
 p_i is the frequency of band presences in locus i , as defined as in the `Shannon Scaling...` dialogue box
 s is the number of loci
 n is the number of individuals
 $\log_2 x = \lg x$ is the logarithm to base 2.

The calculation of this variance will only work if p_i is defined as band presences in a locus relative to all band presences in the data set. Otherwise, doing the calculation may result in a negative value. The standard deviation is calculated as the square root of the variance.

References

Bowman, K. O., Hutcheson, K., Odum, E. P. and Shenton, L. R., 1969, Comments on the distribution of indices of diversity, *Proc. Intl. Symp. Stat. Ecol.* **3**: 315-359.

AMOVA

This function carries out an AMOVA (Analysis of Molecular Variance) analysis as described in Excoffier *et al.* (1992). For this to work you must have groups defined (see `Group Manager` section). Note that for AMOVA results to be meaningful, group must be defined that correctly reflect the population structure you wish to be tested. This means that all individuals that are present in the data set (in the `AllData` structure) must be partitioned into one and only one group. If there are more levels of group hierarchy, again all individuals in all groups must be assigned to one and only one sub-group. FAMD will not check whether these criteria are met - this is your responsibility!

The AMOVA calculation does not require you to previously calculate a distance matrix since the function calculates its own distances as defined under `Options` → `Similarity Coefficient Selection`. The similarity preference selection in this dialogue box defines whether an AMOVA will be conducted on standard, minimum, maximum or average similarities. Note that a minimum similarity naturally goes together with a maximum distance if your selected coefficient is a distance measure.

FAMD will ask you whether or not to save AMOVA results to your analysis file.

The AMOVA SSD terms (see Excoffier *et al.*, 1992) operate on squared distances (δ_{ij}^2). FAMD will square any input distance value, so there is no need for you to manually select the squared distance transformation mode. In other words, the relationship is as follows: (i) FAMD calculates a similarity (s_{ij}) or distance (d_{ij}) matrix. (ii) FAMD carries out a distance transformation (a function of the similarity or primary distance) as $\delta_{ij}=d'_{ij}=f(s_{ij})$ or $\delta_{ij}=d'_{ij}=f(d_{ij})$ and (iii) FAMD performs the AMOVA on δ_{ij}^2 .

You should be aware that (a) the calculated Φ_{ST} etc. values (genotypic variation) cannot be directly compared with F_{ST} etc. values estimated by different procedures, (b) different distance measures will obviously lead to different AMOVA values, and (c) that depending on your combination of parameters, δ_{ij} may not be a Euclidean metric which could potentially impact on the AMOVA.

Please note also that AMOVA values can differ between FAMD and, e.g. Arlequin if a data set contains missing data because Arlequin treats missing data differently. If a data set does not contain missing data and all groups are set up correctly, then AMOVA results between FAMD and Arlequin should be identical. Arlequin uses the standard Euclidean distance with no distance transformation ($d'_{ij}=d_{ij}$).

References

Excoffier, L., Smouse, P. E. & Quattro, J. M., 1992, Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data, *Genetics* **131**: 479-491.

Trees Menu

This menu lets you carry out tree-based analyses (UPGMA and consensus methods) as well as principal coordinate analysis which, admittedly, has very little to do with trees.

UPGMA

This generates a UPGMA (unweighted pair group method using arithmetic averages) tree from a distance [default distance= 1-similarity(Jaccard)] matrix that you have generated previously and stores the tree in the tree output file you have defined (or else to the default tree output file). The distance transformation applied can be changed under `Options → Similarity Coefficient Selection`.

The algorithm implemented is a modified UPGMA algorithm that can generate multifurcating trees, in contrast to a strictly bifurcating implementation. This means that if there are two or more equally good choices for clustering groups of individuals, this will be realised as a multifurcation in the tree, rather than randomly choosing one of the possible choices to generate a strictly bifurcating tree. However, this seldom occurs in real data sets.

FAMD produces trees in the PHYLIP format. You can use other software, such as TreeView to convert between formats and to view/print trees.

Strict Consensus

This function calculates a strict consensus tree from the trees present in the consensus input trees file and saves the consensus to the consensus output trees file. The input file is expected to contain trees in PHYLIP format. The output file likewise will contain PHYLIP format trees. You can use other software, such as TreeView to convert between formats and to view/print trees.

Majority Rule Consensus

This function calculates a majority rule consensus tree from the trees present in the consensus input trees file and saves the consensus to the consensus output trees file. You can set the majority rule consensus threshold (default 70%) under `Options → MR Consensus and R Support Settings`. If a consensus threshold <50% is selected, a 50% threshold will be used. The input file is expected to contain trees in PHYLIP format. The output file likewise will contain PHYLIP format trees. Output trees will contain the percentage of occurrence of nodes saved as node labels (not as branch lengths). You can use other software, such as TreeView to convert between formats and to view/print trees.

Principal Coordinate Analysis

This function carries out a principal coordinate analysis (PCoA or PCO; also referred to as metric multidimensional scaling or MDS; Gower 1966) on the current (dis)similarity matrix. The distance transformation used to generate the distance matrix can be changed under `Options → Similarity Coefficient Selection`.

PCoA is calculated from this distance (d_{ij}) matrix as follows:

- a new matrix **A** is calculated whose elements (a_{ij}) are given by $a_{ij} = -0.5 d_{ij}^2$
- from this, the centred matrix Δ is derived whose elements δ_{ij} are calculated as $\delta_{ij} = a_{ij} - \bar{a}_i - \bar{a}_j + \bar{a}$, where the last 3 terms are the row, column and total means of all a_{ij} values in the matrix **A**, respectively.

-) eigenvalues λ_i and normalised eigenvectors \mathbf{V}_i of the centred matrix Δ are computed using the iterative method of Jacobi implemented in TPMath by Jean Debord. (FAMD calls the Jacobi routine in J. Debord's *eigen* unit to do this.)

-) normalised eigenvectors are then scaled so that $\left| \overrightarrow{V}_k \right| = \sqrt{\lambda_k}$. This is done by multiplying all vector components with the square root of their associated eigenvalue.

-) these scaled eigenvector components are the final PCoA coordinates.

Please note that there is no correction for negative eigenvalues, they (and their associated eigenvectors) are simply ignored.

The first (largest) eigenvalue is output to the analysis file, as is the number and sum of all positive eigenvalues and the PCoA coordinates (scaled eigenvector components). The eigenvalue percentages given in the output file are calculated relative to the sum of all positive eigenvalues found.

Apparently, the output for PCoA analyses are not exactly identical between different pieces of software (i.e., there seem to be some alternative implementations around). For instance, PCoA implemented in the R-package 4.0d9 (Mac) yields different coordinates than SynTax 2000 (Windows). During test runs, FAMD consistently gave results similar to the R-package, but different from SynTax.

References

Gower, J. C., 1966, Some distance properties of latent root and vector methods in multivariate analysis, *Biometrika* **53**: 325-338.

Replicate Analyses Menu

The functions of the `Replicate Analyses` menu work with replicates of your current data set and modify these replicate data sets automatically. Before carrying out any replicate analysis, you should define the parameters for analysis using the `Options → Replicate Analysis Settings` dialogue. The parameters used by the individual functions are indicated in brackets after their name.

Bootstrap Shannon's Variance (SH + RMD)

This option estimates Shannon's index and its variance by data resampling. Unlike the function `Shannon's Index` that operates on your current data set and for which you should first remove monomorphic loci from the data set, this option generates resampled data sets from your current data set and automatically removes monomorphic loci from it before calculating Shannon's index. Since every resampled data set may contain a different configuration of loci from the original data matrix and since missing data may be replaced randomly, it may be that loci that are treated as monomorphic are not monomorphic in another data set replicate. Therefore, you **SHOULD NOT** remove monomorphic bands manually from your data set (or replace missing data manually) for carrying out this function, since thereby you will limit the variation generated during bootstrapping.

This function uses the parameters `SH` and `RMD` set in the `Options → Replicate Analysis Settings` dialogue box. It is implemented as follows:

Repeat the following `SH` times:

-) Generate a resampled data set by randomly choosing `s` loci from your current data set (every locus can be picked in every random draw), where `s` is the number of loci present in your current data set.
-) If `RMD` is set, replace missing data so that monomorphic loci are removed from the newly generated replicate data set (according to the definitions given under `Remove Loci...`)
-) From this data matrix, Shannon's index is calculated as described in the respective section

The series of values for Shannon's index are averaged and the variance calculated.

Please be aware that for high `SH` values, this procedure may take a considerable time.

Bootstrap Std Tree (BS + RMD)

This function can be used to generate multiple UPGMA trees based upon the standard selected similarity of distance coefficient from resampled data matrices and stores them in the tree output file. You can then use the `Trees → Strict Consensus` or `Trees → Majority Rule Consensus` functions, or alternatively use other software to analyse the trees.

This function uses the parameters `BS` and `RMD` set in the `Options → Replicate Analysis Settings` dialogue box. It is implemented as follows:

Repeat the following `BS` times:

-) Generate a resampled data set by randomly choosing `s` loci from your current data set (every locus can be picked in every random draw), where `s` is the number of loci present in your current data set.
-) If `RMD` is set replace missing data so that monomorphic loci are removed from the newly generated replicate data set (according to the definitions given under `Remove Loci...`)

-) From this data matrix, calculate similarity matrix based upon the selected coefficient of similarity or distance (standard definition)
-) Perform the selected distance transformation
-) Perform the UPGMA clustering algorithm on the distance matrix
-) Write a tree to the tree output file.

Multiple Avg Trees (TR + JC)

This function can be used to generate multiple UPGMA trees based upon the average selected similarity (or distance) coefficient and stores them in the tree output file. You can then use the `Trees → Strict Consensus` or `Trees → Majority Rule Consensus` functions, or alternatively use other software to analyse the trees.

This function uses the parameters `TR` and `JC` set in the `Options → Replicate Analysis Settings` dialogue box. It is implemented as follows:

Repeat the following `TR` times:

-) Randomly draw `JC` values from the interval $[s_{ij,min}; s_{ij,max}]$ (minimum to maximum possible similarity value) and calculate average similarity values from this series of numbers. The same is done if the selected coefficient directly produces a distance measure
-) Generate a (dis)similarity matrix from the average Jaccard values
-) Perform the selected distance transformation
-) Perform the UPGMA clustering algorithm on the distance matrix
-) Write a tree to the tree output file

Bootstrap Avg Trees (BS + JC)

This function can be used to generate multiple UPGMA trees based upon the average selected similarity (or distance) coefficient from resampled data matrices and stores them in the tree output file. You can then use the `Trees → Strict Consensus` or `Trees → Majority Rule Consensus` functions, or alternatively use other software to analyse the trees.

This function combines the data resampling of loci used for bootstrapping values and the calculation of average Jaccard coefficients by sampling from the interval of possible values.

This function uses the parameters `BS` and `JC` set in the `Replicate Analysis Settings` dialogue box. It is implemented as follows:

Repeat the following `BS` times:

-) Generate a resampled data set by randomly choosing `s` loci from your current data set (every locus can be picked in every random draw), where `s` is the number of loci present in your current data set.
-) Randomly draw `JC` values from the interval $[s_{ij,min}; s_{ij,max}]$ (minimum to maximum possible similarity value) and calculate average similarity values from this series of numbers. The same is done if the selected coefficient directly produces a distance measure.
-) Generate a similarity matrix from the average [dis]similarity values
-) Perform the selected distance transformation

-) Perform the UPGMA clustering algorithm on the distance matrix
-) Write a tree to the tree output file

Estimate R-support (TR)

This function attempts to estimate the R_x -support for clusters and generates a R_x -consensus tree (written to the consensus tree output file) according to the options set under `Options → MR and R-support Settings` and the `TR` parameter set in the `Options → Replicate Analysis Settings` dialogue. The output tree will be in PHYLIP format with node labels (e.g., “R1.8”) indicating estimated R_x support for a given cluster. Clusters appearing only at r values higher than the specified threshold value (or the maximum r value analysed, whichever is the smaller value) will be unresolved and branches collapsed.

NB:

A) Even in relatively small data sets, this function can be very resource-intensive. It requires a lot of computing (CPU time = user waiting time), and memory. In some instances, FAMMD may ask Windows for more memory than Windows is willing to give it. In this case, there will be an “Out of memory” error. This is not a program error in FAMMD, but simply says that the computer system’s resources are insufficient to carry out the analysis given the selected parameters. Performing the analysis on a newer computer system may resolve this problem, if it occurs.

B) FAMMD is a single-threaded program and will NOT respond to user input until it has finished its calculations (or encounters an error).

This analysis is designed to estimate R_x -support for clusters found during UPGMA clustering. Average similarity (or distance) coefficients are calculated by averaging r values drawn from the interval $[s_{ij,min}; s_{ij,max}]$. Consensus trees are then constructed from `TR` trees calculated from average similarity matrices generated using different values of r . Given the consensus threshold of x %, the value r_x for a cluster considered will be the smallest value of r in that a cluster appears in the consensus tree. Since these values r_x may be large numbers, it is convenient to define the replicate support number R_x at the given consensus threshold percentage of x % as

$$\text{Replicate support:} \quad R_x = \log_{10} r_x$$

The smallest possible values are $r_x = 1$ and $R_x = 0$. The smaller the R_x value obtained for a given cluster, the more highly supported this cluster is based upon data and missing data in the data matrix.

When r becomes very large, sampling from $[s_{ij,min}; s_{ij,max}]$ is expected eventually lead to convergence of average similarity, s_{ij}^* , on arithmetic mean of minimum and maximum similarity values. Since replicate support is estimated essentially using a stochastic process, slight differences between runs may be expected. It is also possible that, very rarely, a tree is found which is apparently inconsistent with previous trees, i.e., does not contain a cluster the present r value which was found in a tree with a smaller r value. The current version of FAMMD will report such trees - should they be found - but essentially ignores them.

Note that the replicate number, r , is the same as the variable `JC` defined for use in other functions in e.g. the `Replicate Analyses` menu with the only difference that `JC` remains constant during these analyses but different values of r are considered during the `Estimate R-support` routine.

Shannon t-tests

This option lets you carry out t-tests comparing different Shannon values. In order to use this function, groups must be defined in your data matrix (see `Group Manager` section). FAMMD will display a dialogue box asking you to check/uncheck the groups whose Shannon indices you wish to compare.

The dialogue box also asks you what data you wish to see written to your analysis file. You can make the following choices:

-) Should Shannon's measure and variance will be calculated using the Bowman *et al.* (1969) formula?
-) Should Shannon's measure and variance will be estimated by bootstrapping?
-) Should FAMD output the p-value or simply tell you significant vs. insignificant at a given p-value?
-) Should FAMD output the p-value or simply tell you significant vs. insignificant at a given p-value?
-) Should FAMD output t(df) values?
-) Should FAMD output the actual Shannon values and variances and group sizes?

Shannon's index will be calculated according as specified under `Options → Shannon Scaling`, and, if the bootstrapping method is selected, the additional settings under `Options → Replicate Analysis Settings`.

NB: If you use the Shannon/Bowman variance option and band frequencies $p(i)$ are defined on a per-locus basis, the variances may become negative which precludes t-testing. FAMD will not stop you to use such settings, but will display error messages if variances do get negative.

This routine operates on transient copies of the input data matrix and takes care of the removal of monomorphic bands in these replicate data sets.

The implementation for Shannon/Bowman variance is as follows:

For every selected group, do the following:

-) If `RMD` is set replace missing data so that monomorphic loci are removed from the newly generated replicate data set (according to the definitions given under `Remove Loci...`)
-) From this data matrix, calculate Shannon's index is calculated using the scaling options selected and the Bowman variance

The implementation for Shannon/bootstrapped variance is as follows:

For every selected group, repeat the following `SH` times:

-) Generate a resampled data set by randomly choosing s loci from your current data set (every locus can be picked in every random draw), where s is the number of loci present in your current data set.
-) If `RMD` is set replace missing data so that monomorphic loci are removed from the newly generated replicate data set (according to the definitions given under `Remove Loci...`)
-) From this data matrix, Shannon's index is calculated using the scaling options selected

The series of values for Shannon's index are averaged and the variance calculated.

T-Tests are carried out as follows:

t-values are calculated as

$$t = \frac{I_1 - I_2}{\sqrt[2]{\text{var}(I_1) + \text{var}(I_2)}}$$

where...

I_X is Shannon's index for group X.

$\text{var}(I_X)$ is the variance of I_X .

the degrees of freedom are calculated as

$$df = \frac{(\text{var } I_1 + \text{var } I_2)^2}{\frac{\text{var}^2(I_1)}{n_1} + \frac{\text{var}^2(I_2)}{n_2}}$$

where...

I_X is Shannon's index for group X.

$\text{var}(I_X)$ is the variance of I_X .

n_X is the sample size (number of individuals) of group X.

p-values are calculated as implemented in TPMath by Jean Debord. [FAMD calls the PStudent (df, tvalue) function in J. Debord's *fspec* unit.]

Reference

Bowman, K. O., Hutcheson, K., Odum, E. P. and Shenton, L. R., 1969, Comments on the distribution of indices of diversity, *Proc. Intl. Symp. Stat. Ecol.* **3**: 315-359.

Options Menu

Functions in this menu allow you to set parameters that are use by different routines in the program. Changing any of these parameters does not actually result in any calculations or modifications done to your data matrix.

Missing Data Replacement Settings

This displays a dialogue box that lets you select how missing data should be treated by those routines that deal with missing data. No action is performed on the data matrix. The options are to replace all missing data points by presences, by absences or to randomly replace missing data by x % of presences and $(100-x)$ % of absences, where x is the value entered by you. The default is missing data replacement by 50 % presences.

Shannon Scaling...

Here you can define which frequency definition and which logarithm base should be used for calculation of Shannon's index.

The frequency of band presences $p(i)=p_i$ can be defined:

-) relative to all presences (frequency per data set):

$$p(i) = p_i = \frac{N_1(i)}{\sum_{k=1}^s N_1(k)}$$

where $p(i)$ is the band frequency in locus i
 $N_1(i)$ is the number of band presences in locus i
 s is the number of loci in the data set
 $\sum N_1$ is the total number of band presences in the data set.

-) divided by the number of individuals (frequency per locus):

$$p(i) = p_i = \frac{N_1(i)}{n}$$

where $p(i)$ is the band frequency in locus i
 $N_1(i)$ is the number of band presences in locus i
 n is the number of individuals in the data set

The will program will always calculate Shannon's index using $\log_2 x = \lg x$ (the dual/binary logarithm), because (i) FAMD/Shannon's index deals with binary data and (i) it is the native logarithm for the computer's processor (FPU):

$$I_2 = I \approx -\sum_{i=1}^s p_i \log_2 p_i$$

where I_2 represents the fact that logarithm with base 2 was used

However, since in principle, any logarithm can be used, the program can re-scale Shannon's index accordingly by multiplying with a correction factor:

$$I_A = I_2 \log_A 2$$

where I_2 represents Shannon's index based on $\log_2 x$
 I_A represents Shannon's index based on $\log_A x$

You can select the following options:

-) use $\log_2 x = \lg x$
-) use $\log_e x = \ln x$
-) use $\log_{10} x = \lg x$
-) use $\log_A x$, where A is user-defined

Replicate Analysis Settings

Here, you can define parameters for different analysis functions that operate on multiple data set replicates. It is advisable if you do this before you start your analyses. The available parameters are:

- SH: Resampled data matrix replicates for Shannon
 This number defines how many replicates of your current data set should be generated for estimating Shannon's index by data resampling (i.e. the number of bootstrap replicates).
- RMD: Replace missing data (as set in respective dialogue)
 If checked, missing data in the data matrix will be replaced according to the parameters defined in the Options → Missing Data Replacement Settings dialogue box.
- BS: Resampled data matrix replicates for similarity analyses
 Defines from how many data set replicates UPGMA trees should be generated.
- JC: Average Jaccard from how many random draws
 Defines the number of random draws from the interval $[s_{ij,min}; s_{ij,max}]$ that is used for calculating an average similarity value and its variance.
- TR: Number of average similarity trees to generate
 Defines how many UPGMA trees based upon average Jaccard-derived distance matrices should be generated.

MR Consensus and R-support Settings

Here you can define parameters that are used by those routines that require consensus tree methods, especially the Replicate Analyses → Estimate R-support function.

You can define

-) the threshold percentage used for the majority rule consensus function. Clusters occurring at a frequency less than the threshold frequency will be unresolved in the resulting consensus tree. (Changing this parameter does not affect R_x -analysis.)
-) the threshold percentage, x , for the majority rule consensus routine used by FAMD used internally during R_x -support analysis. 100% (the default) means that a strict consensus is used (and R_{100} calculated).
-) the R_x threshold r -value (corresponding to variable JC for average similarity calculations) for the majority rule consensus routine used by FAMD used internally during R-support analysis. Clusters will be unresolved on the R_x -consensus tree if

they occur only in consensus trees generated at r-values greater than or equal to the specified values.

-) specify the range of values that should be considered for R-support analysis, i.e. the minimum and maximum r-values to be considered.
-) the desired precision to be obtained for R_X values. A precision of 1 digit here would specify an output of values with one digit after the decimal point, e.g. $R_X=1.3$, where the analysis should be sufficiently precise that run-to-run variation between R_X -values should be limited to the last digit displayed. Note that increasing the precision of the analysis may drastically increase computing time and memory used by FAMD.

Similarity Coefficient Selection

Using this dialogue box, you can select the similarity or dissimilarity coefficient to be used by FAMD, which distance transformation to be applied and whether you wish to use standard, minimum, maximum or average similarity coefficients for AMOVA. Please note that, when using a distance measure, a minimum distance would correspond to a maximum similarity.

Jaccard, Dice and SMC coefficients produce similarities, NeiLi, Euclidean, and squared Euclidean produce distances. Distance measures are marked with an asterisk (*) in the dialogue box.

For details of the different similarity coefficients, please see the `Standard`, `Minimum`, `Maximum`, and `Average Similarity` sections in this manual.

The distance transformations available are the following.

-) $d_{ij} = 1 - s_{ij}$ for similarities or $d'_{ij} = d_{ij}$ for distances (i.e., distance measures are not modified further)
-) $d_{ij} = \sqrt[3]{1 - s_{ij}}$ for similarities or $d'_{ij} = \sqrt[3]{d_{ij}}$ for distances
-) $d_{ij} = (1 - s_{ij})^2$ for similarities or $d'_{ij} = d_{ij}^2$ for distances

Help Menu

About

Displays a message box with the authors's contact address, a copyright notice, the legal disclaimer.

Help

Displays this help file (famdhelppdf).