

Manual for **familias**





Thore Egeland *
Petter F. Mostad**

* Institute of Forensic Medicine
University of Oslo
Rikshospitalet
0027 Oslo - Norway
Email: thore.egeland@medisin.uio.no

** Petter Mostad
Mathematical sciences
Chalmers University of Technology and Göteborg University
SE-412 96 Göteborg, Sweden
Email: mostad@chalmers.se

Version: September 19, 2010

Contents

I	PREFACE	4
1	INTRODUCTION	4
2	EXAMPLE – A PREVIEW OF FAMILIAS	5
	Calculation of the likelihood ratio (LR) by hand	5
	Calculation of the likelihood ratio using familias	5
3	USER'S GUIDE	9
3.1	General DNA data	11
	Reading system data from file	11
	Writing system data to file	12
3.2	Persons 	14
3.3	Known relations 	15
3.4	Case related DNA data 	15
	Read case data	16
	Write case data	17
3.5	Pedigrees 	17
A	APPENDICES	21
A1	Theory and methods	21
A1.1	Prior model	21
A1.2	Posterior model	22
A1.3	Substructure corrections	23
A1.4	Mutation models	24
A2	Solved exercises	29

Manual for **familias**

Answers	41
A3 Generating pedigrees automatically	45
A4 Implementation of prior distribution	46
A5 Description of general input files for familias	48
References	52

i Preface

This document updates the documentation of the **familias** program <http://familias.name> in connection with the 2.0 version released in September of 2010. A list of changes and bug fixes in this latest version appears on the home page. Some material which has previously been scattered (exercises with answers and a description of the file format) has now been included in this document. Comments on the documentation or the program can be sent to thore.egeland@medisin.uio.no or mostad@chalmers.se.

1 Introduction

The **familias** program may be used to compute probabilities and likelihoods in cases where DNA profiles of some people are known, but their family relationship is in doubt. Given several alternative family trees (or pedigrees) for a group of people, given DNA measurements from some of these people, and given a data base of DNA observations in the relevant population, the program may compute which pedigree is most likely, and how much more likely it is than others. Obviously, there are several other programs performing similar tasks. As far we know a distinguishing feature of **familias** is its ability to handle complex cases where potential mutations, silent alleles and population stratification (θ -corrections) are accounted for, together with its ability to handle multiple pedigrees simultaneously. The program has been validated (Drabek 2009). The book (Buckleton, Triggs et al. 2005) provides a general background to forensic genetics.

The reference to **familias** is (Egeland, Mostad et al. 2000) and the program can be freely downloaded from <http://familias.name>. Several example data files, are available from the site. On line help, a short tutorial, is available directly from the help-function of the program. **familias** has been applied in a large number of cases, including identification following disaster, resolving family relations when incest is suspected and determining the most probable relation between a person applying for immigration and claimed relatives of the individual.

The contents of this document are as follows. [Section 2](#) gives a brief introduction to program by means of a simple worked example. Next, [Section 3](#) provides an overview of the options available in the program, along with suggestions for typical values for the various parameters. Some more theory and advanced options are presented in [Appendix A1](#). Solved exercises appear in [Appendix A2](#). These exercises serve also to check the accuracy of the program. Appendices [A3](#) and [A4](#) describe advanced options. Finally, the file format of the input file is described in detail [Appendix A5](#); this is only relevant for programmers as the purpose is to enable programmers to write code producing input files for the **familias** program, on the “**familias** format”.

2 Example – a preview of **familias**

This section presents a very simple case. First, calculations are done by hand and then we demonstrate how the calculations are done using **familias**.

We consider the following hypotheses concerning the relationship between two males M1 and M2:

- H_1 : M1 is the father of M2
- H_2 : M1 is not the father of M2

An illustration of the hypothesised relationship is given in Figure 2.1. F1 is the mother. It is common practice for such illustrations to denote men with squares and women with circles.

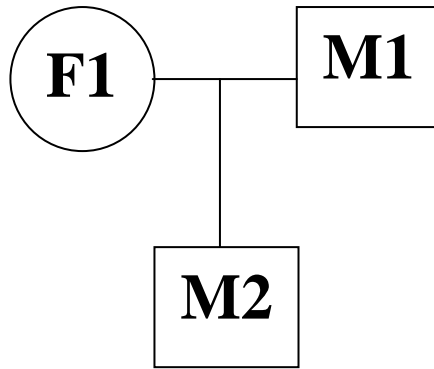


Figure 2.1: The pedigree corresponding to hypothesis H_1 .

The genotypes are $F1 = \{A, A\}$, $M1 = \{B, B\}$ and $M2 = \{A, B\}$. The relative allele frequencies of A and B are $p_A = p_B = 0.05$ and Hardy-Weinberg equilibrium is assumed. The child has inherited the allele A from his mother and the allele B must be inherited from the father.

Calculation of the likelihood ratio (LR) by hand

The likelihood ratio is then given by

$$LR = \frac{P(\text{data} | H_1)}{P(\text{data} | H_2)} = \frac{P(M2 = \{A, B\} | H_1)}{P(M2 = \{A, B\} | H_2)} = \frac{1}{p_B} = \frac{1}{0.05} = 20.$$

Calculation of the likelihood ratio using **familias**

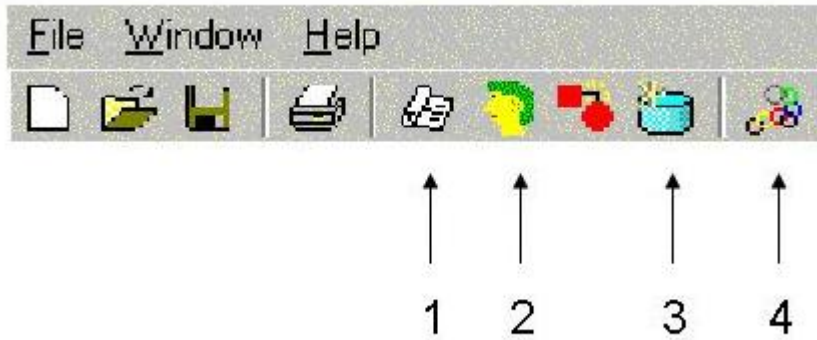



Figure 2.2: *The main menu of familias.*

The calculation in familias is performed by going through the four steps indicated in Figure 2.2.

1. **General DNA data.** The window appearing after clicking  and Add should be completed as shown in Figure 2.3 (the part not shown should be left unchanged for now).

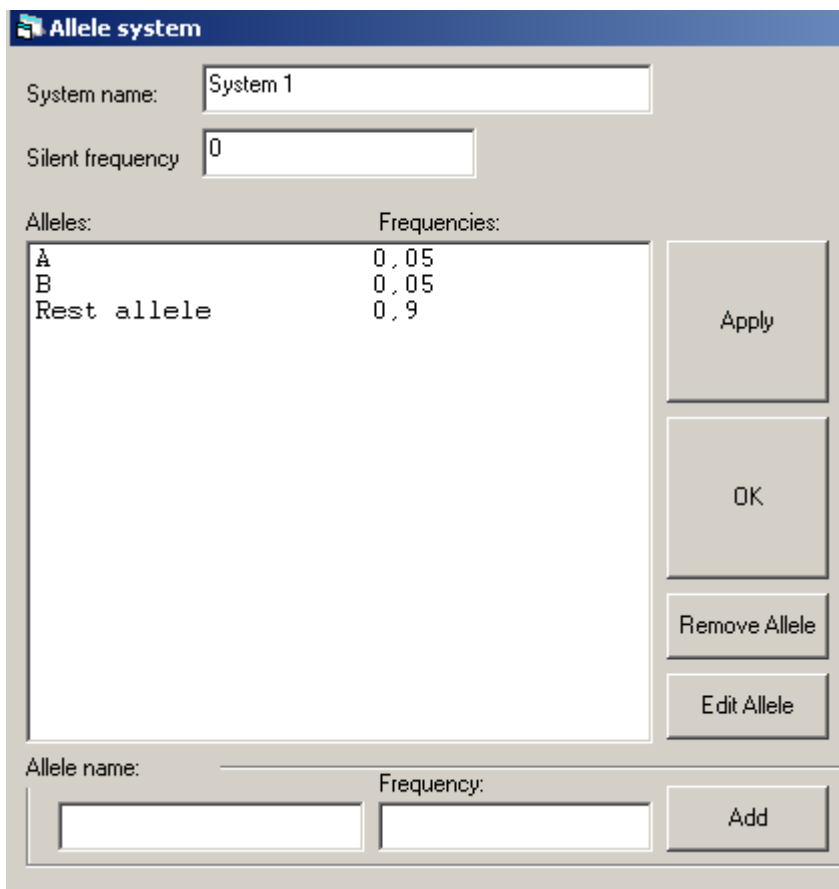


Figure 2.3: *The General DNA data window*

2. **Persons.** The window appearing after clicking  should be completed as shown in Figure 2.4 below.

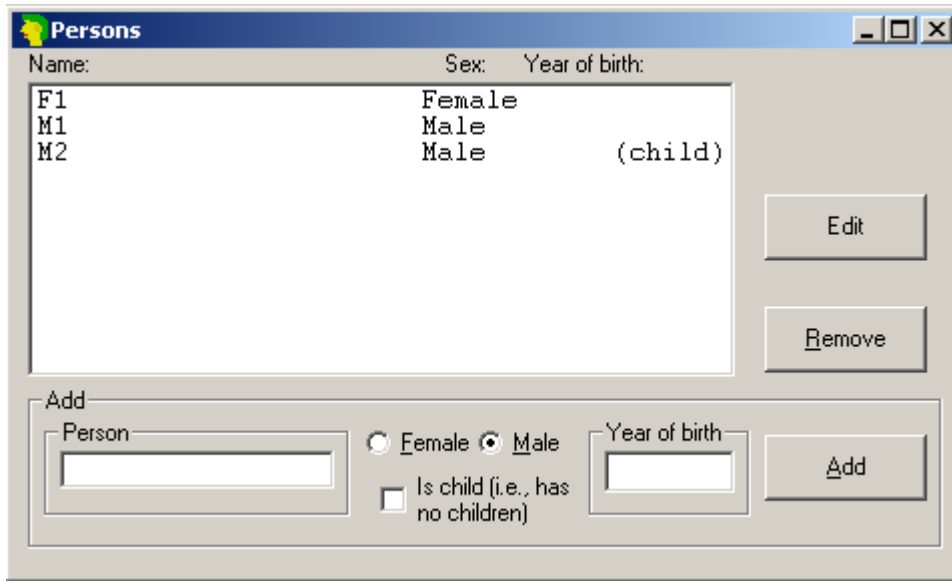



Figure 2.4: *The Persons window.*

3. **Case related DNA.** The window appearing after clicking  should be completed as shown in Figure 2.5 below.

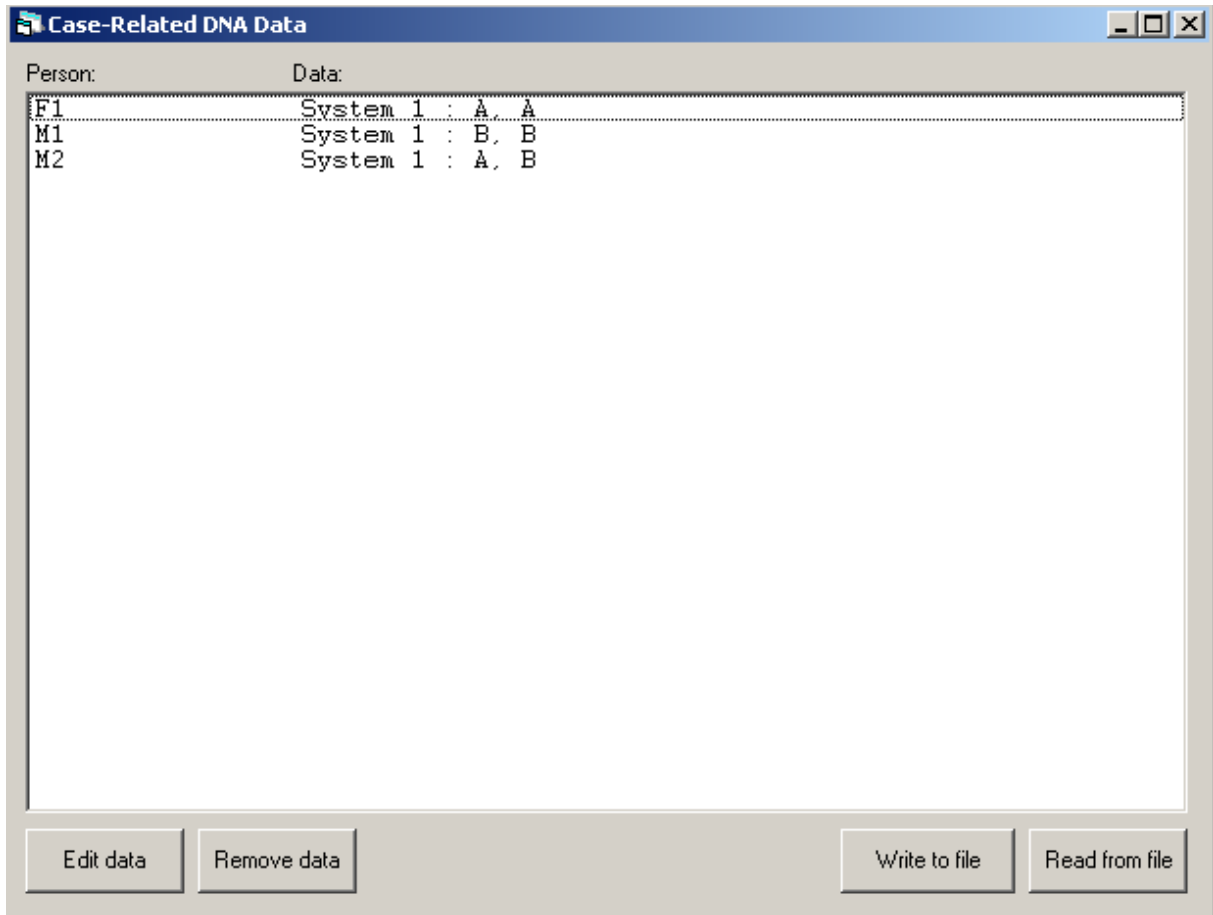



Figure 2.5: *The Case related DNA window.*

The data is entered by clicking the persons and using the menus.

4. **Pedigrees.** Click . First the pedigree corresponding to hypothesis H_2 is entered by clicking Add. This should be completed as shown in Figure 2.6 .

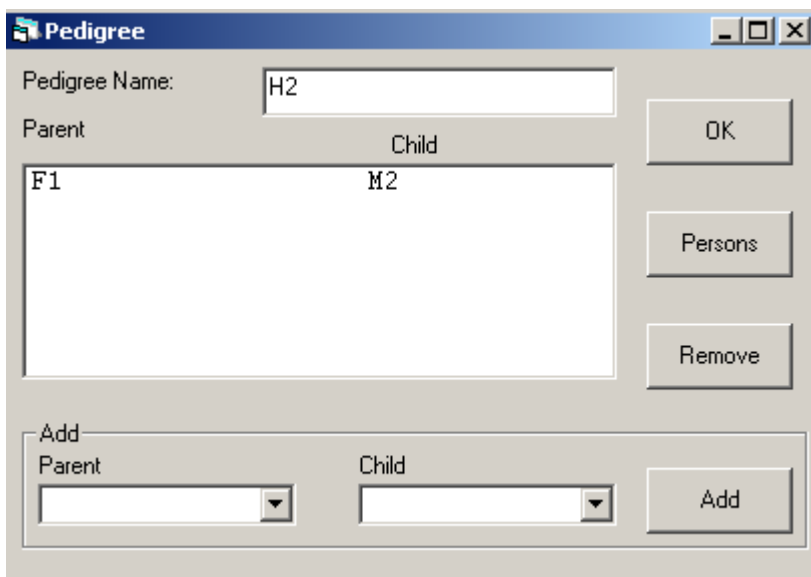


Figure 2.6: *Defining the pedigree corresponding to hypothesis H_2 .*

The next pedigree is defined similarly. The answer, $LR = 20$, appears by clicking LR as shown in Figure 2.7.

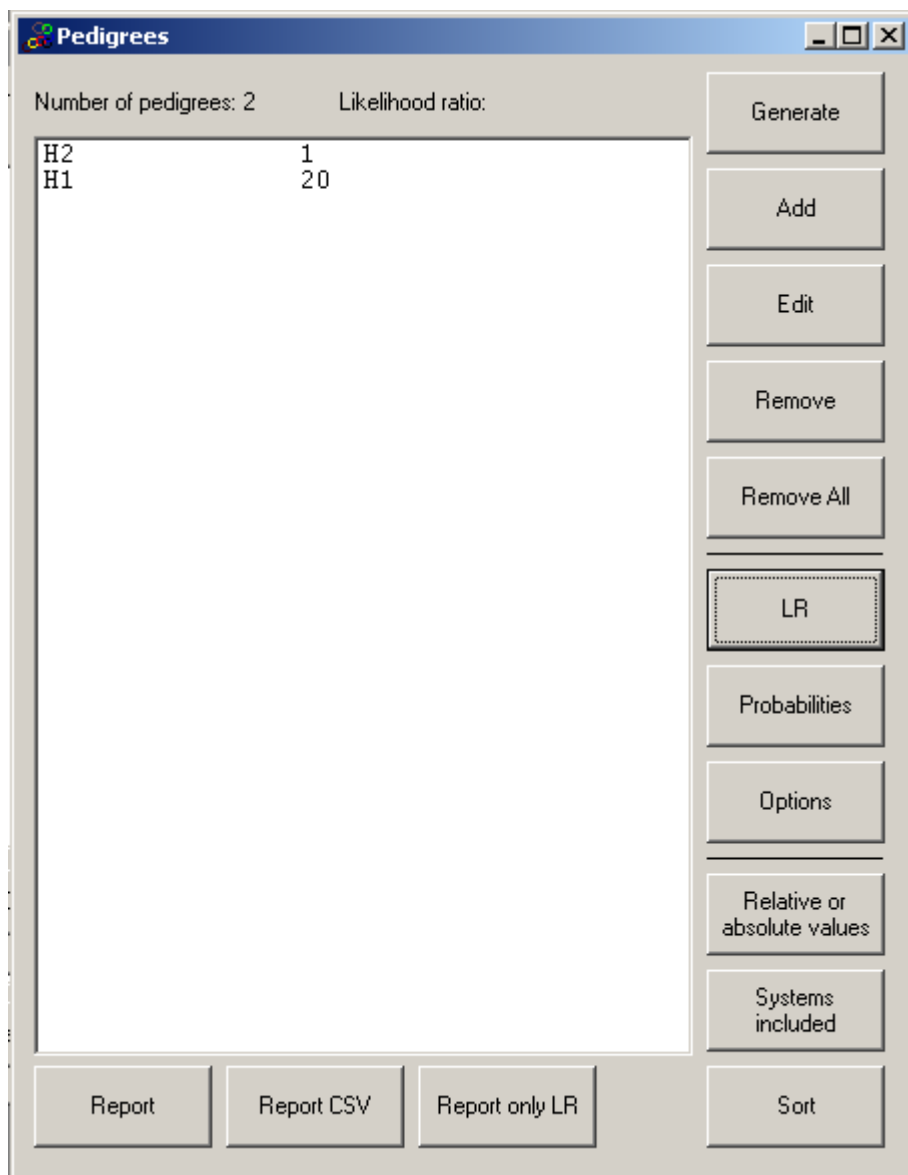


Figure 2.7: *The result as shown in the Pedigrees window.*

3 User's guide

In this section we explain how to use the **familias** program with more details on the functionality. The main menu of **familias** is illustrated in Figure 3.1 below.



Figure 3.1: *Main menus of **familias***

The first four buttons are common to most windows programs: New file, Open file, Save file and Print file. The next five buttons are specific to **familias** and will be treated in the following sections. They are [General DNA data](#), [Persons](#), [Known relations](#), [Case related DNA data](#) and finally [Pedigrees](#). In addition, there is a menu **Casenotes** which can be used to record information regarding the project. Pressing any of these buttons will make a window with the same title appear.

Usually, the user will go through most of the options in a particular fashion. First, the allele systems are defined under **General DNA data**. This is often done manually, but it is also possible to import such data from a database. Secondly, the persons are defined by their gender and age under **Persons**, and possible known relationships are entered under **Known Relations**. Next, under **Case Related DNA Data**, the genotypes of the relevant persons are entered for all available allele systems. Finally, the **Pedigrees** window is used to define pedigrees (either manually or automatically), and perform calculations of probabilities and likelihoods.



Figure 3.1: *The main menu of **familias**.*

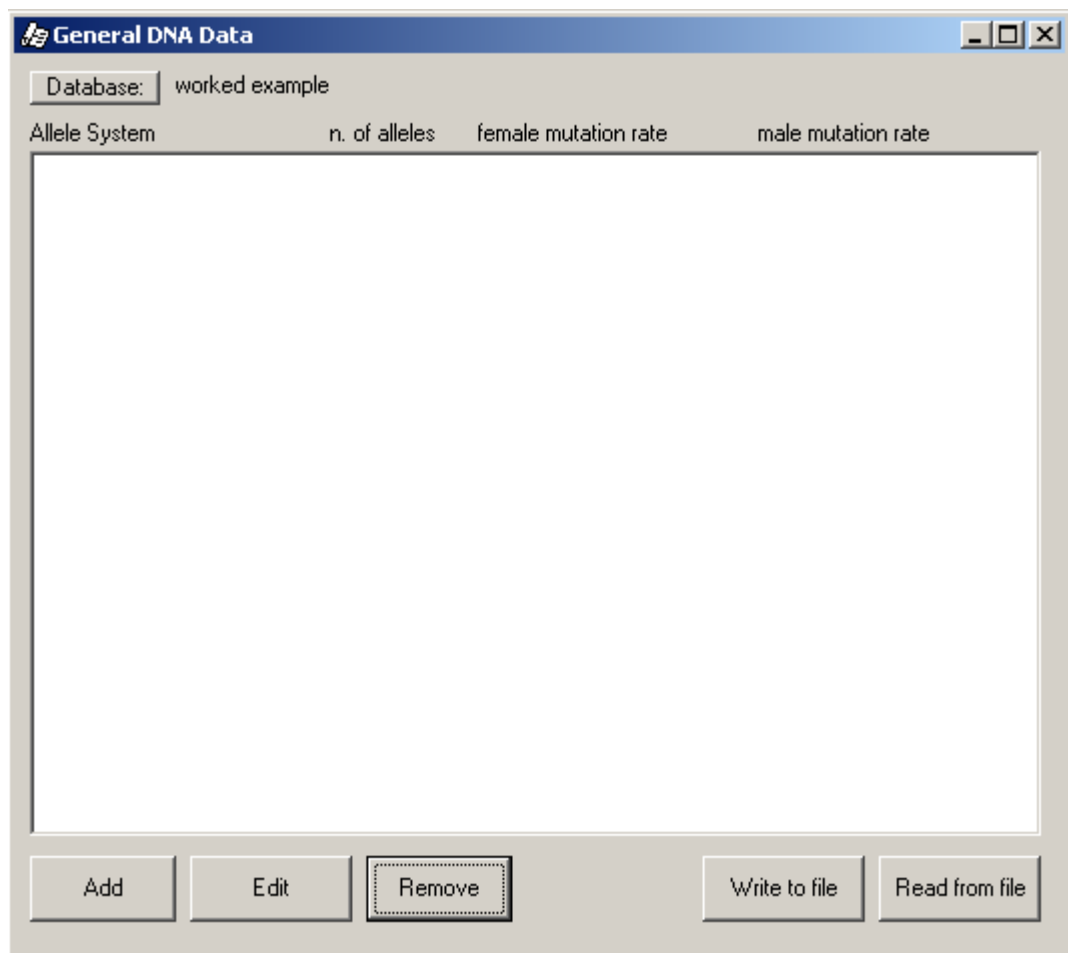


Figure 3.2: *A system may be added, edited or removed manually.*

3.1 General DNA data



This window provides options for adding, editing, removing, reading and writing allele systems. An illustration of the window is given in Figure 3.2. Systems may be edited manually or by reading files.

Entering data manually

To enter a new allele system manually, press **Add**. Then the **Allele system** window, illustrated in Figure 3.3, appears. Here you enter the system name, the alleles and their respective frequencies. It is recommended to ensure that the allele frequencies sum to 1. If necessary, an extra allele, called say ‘Rest allele’ can be added as in Figure 3.3.

Sorting

Alleles are sorted according to name. This is essential when using mutation models that depend on the ordering of the alleles. It also means that the alleles in such a system must be named alphabetically. For example, if your alleles are named 8, 9, 10, 11, 12, you should give them the names 08, 09, 10, 11, 12 when describing them to the program, otherwise they will be sorted alphabetically as 10, 11, 12, 8, 9. If the first character of a system name is a letter, it is probably wise to use small or capital letters consistently. For instance alleles a and E are sorted E, a whereas a, e are sorted a, e.

Reading system data from file

The file below corresponds to the output from an Excel file, with tabs as separators. The different systems are listed below each other, separated by at least one blank line. The listing for each system starts with the name of the system, followed by a number of lines, each containing the name of the allele, and as the following item, the frequency. The alleles are sorted by the program, alphabetically according to name, to correspond to the corresponding sorting when inputting alleles manually. The data is read in, and is added to the current allele systems. The name of the file read is recorded in the upper left corner adjacent to the field **Database**. This field can be edited to keep track of the database used or modified. If allele systems with the same names already exist, these are replaced. The systems are created with zero mutation rates and no silent alleles. If the frequencies listed are not positive, an error is issued, and the reading of data stops. If the frequencies do not add to 1, they are adjusted to do so, with a warning. An example of input is given below.

```
SYS1
A    ,002
B    ,096
C    ,119
D    ,225
```

Manual for **familias**

E	,326
F	,163
G	,056
H	,013

SYS2

06	,056
07	,073
08	,190
09	,192
10	,253
11	,143
12	,089

Table 3.1: Example of system data that can be read into **familias** from the General DNA Data window. You can load the data between the lines below by cutting and pasting in an editor like word or excel. It is important that you save the data as a text file, from excel you should use tab delimited text file, as mentioned previously. The allele frequencies of SYS2 do not sum to 1. On reading into **familias** a warning will be given for this system before the allele frequencies are scaled to add to 1.

Writing system data to file

The system data can be written to a file on the same format as used for input.

Silent alleles

There is also a possibility to specify a frequency for a silent allele. This refers to alleles that for some reason or other are not detected with the common methods. With a positive silent allele frequency, you cannot know whether an identified homozygote really is homozygote or if he is heterozygote with the other allele being a silent allele. The silent allele frequency and the other allele frequencies should add to 1. Further details on silent alleles are given in [Exercise 2](#) of Appendix A2.

Specifying mutation models

The default value for mutation rates is zero. However, if it is known or reasons to suspect that there is a non-zero mutation rate, it should be specified here. A reasonable mutation rate could be around 0.005. The program offers the possibility to distinguish between male and female mutation rates. The reason for this is that paternal alleles tend to mutate more often than maternal alleles. There are 4 different mutation models to choose from

- 1) Prob decreasing with range (stable)
- 2) Prob decreasing with range (equal).

- 3) Probability proportional to frequency (stable)
- 4) Equal probability (simple and fast)

Mathematical details are provided in [Appendix A1](#), along with an example of analytical calculations incorporating the various models. However, to use the program all you really need to know is the following: If a model is stable this implies that adding irrelevant persons will not affect the result. Conversely, for unstable models adding irrelevant persons may lead to slightly different results.

For models 1 and 2 the probability of mutation depends on the size of the mutation. For example, if you have an allele with 14 repetitions, this allele will be more likely to mutate into an allele with 13 or 15 repetitions than to an allele with 12 or 16 repetitions. For models 1 and 2, **familias** the user must supply a parameter. A typical Mutation range is 0.1. This value corresponds to a mutation probability that decreases by one tenth for each additional unit length difference between the parent allele and the offspring allele. Be aware that the “length” of the alleles is only decided by the order in which they are entered. The difference in length between two subsequent alleles is taken to be 1, which means that it in some circumstances it will be necessary to enter unobserved alleles.

Consider next model 3 Probability proportional to frequency. Here the probability of mutating *to* an allele is proportional with this allele’s frequency in the population. This means that if you have, e.g., an allele A with frequency 0.05 and another allele B with frequency 0.1, then the probability for a mutation leading to a new allele B is larger than one resulting in a new allele A.

In the model Equal probability (simple and fast) the probability of mutating from one allele to another allele is the same independently of the frequency and the range of the alleles.

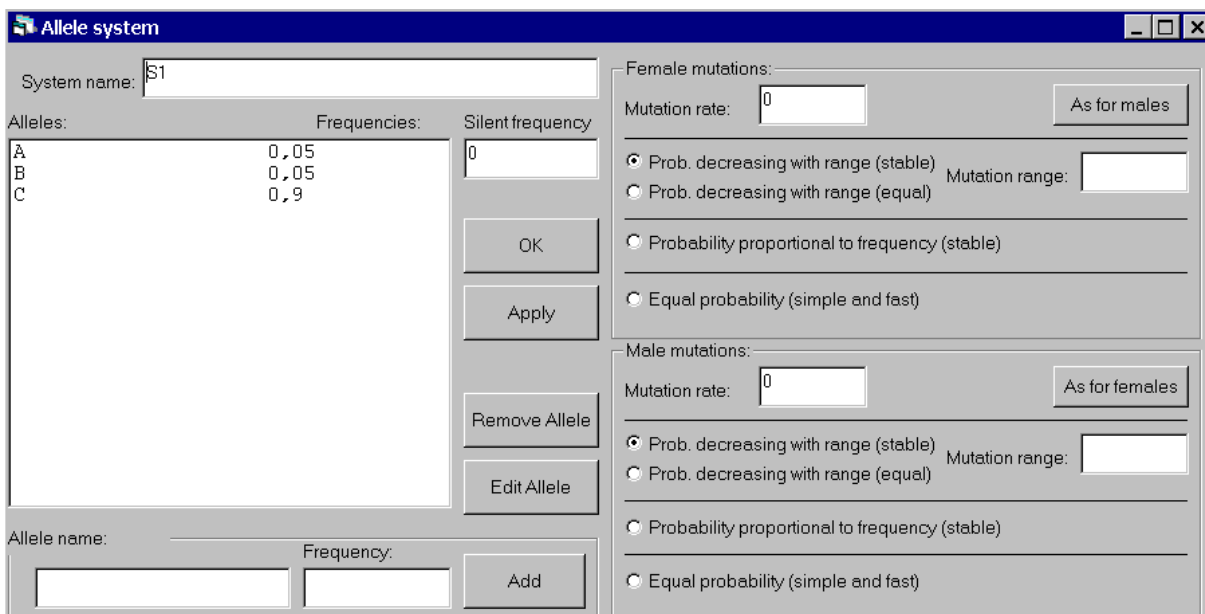


Figure 3.3: *The window for entering an allele system.*

3.2 Persons



By pressing this button, the window shown in Figure 3.4 appears. Here you define the persons involved in the case. For each person a name and gender must be specified. For most applications this is the only information needed and used. In addition, it is possible to enter a year of birth, and you may also specify if the person is a child or in effect has no children. Concerning the year-of-birth specification: as **familias** only makes use of the relative dates, it is possible to use this option to specify age differences even when the exact year-of-birth is unknown. The Is Child-option is used to limit the number of possible pedigrees. The list of persons is edited by means of **Edit** and **Remove**.

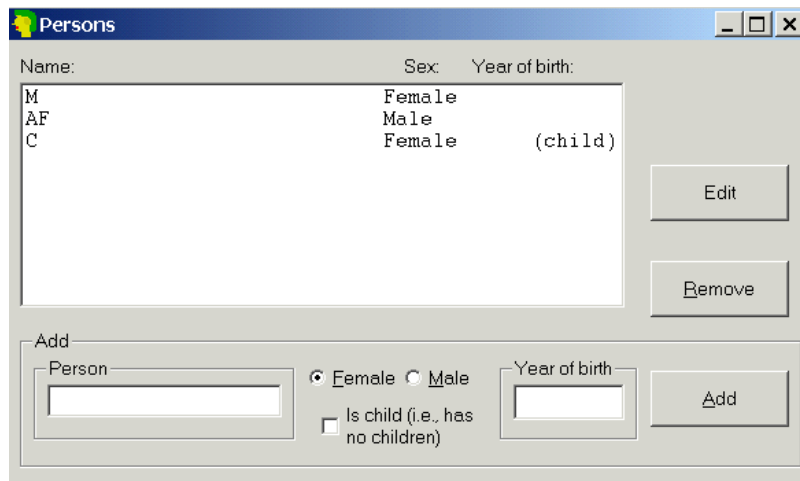


Figure 3.4: *The window for entering the persons involved in a case.*

3.3 Known relations


This is where known relations are defined. If it is certain that, e.g., F is the father of D, then this could be specified here. It is only possible to define parent-child relations. This means that if, for example, two girls are known to be sisters, this cannot be defined straightforward, but through their relations with the common parents. The window is illustrated in Figure 3.5. The menu  is not strictly needed as this information can be provided also when the pedigrees are defined.



Figure 3.5: *The window for entering known relations.*

3.4 Case related DNA data

In this form you enter the DNA data for the persons for whom this information is available. This can be done manually or by reading from a file.

Manually

By marking one of the persons on the list in the window shown in Figure 3.6, and pressing **Edit data**, a new window appears (see Figure 3.7). Here you enter, for the selected person, the DNA data of all the investigated allele systems. For persons for whom there are no available DNA data, just leave it open. Apparent homozygotes are entered with two of the same allele, also in the cases where there could be silent alleles.

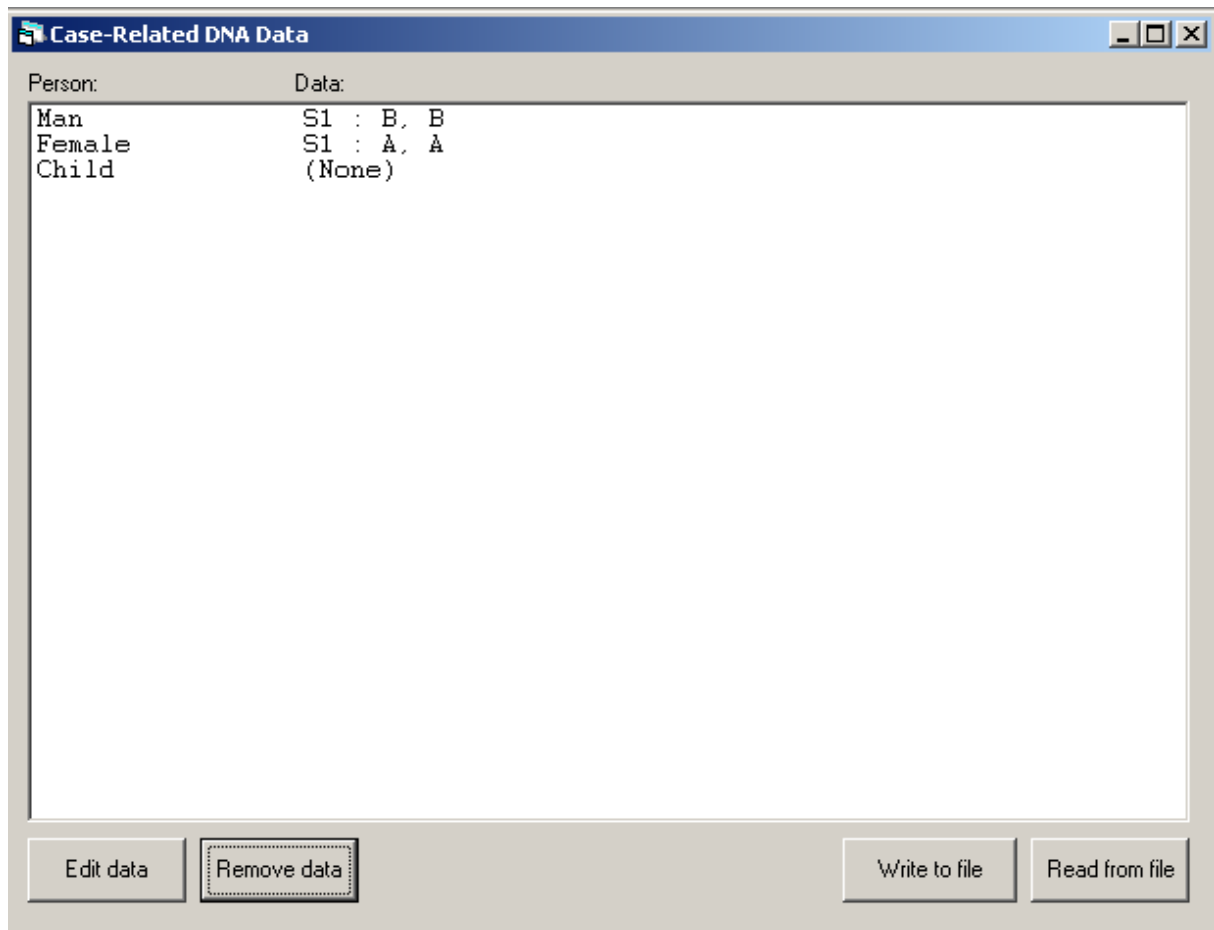


Figure 3.6: *Selecting the persons to assign genotype data is done in this window.*

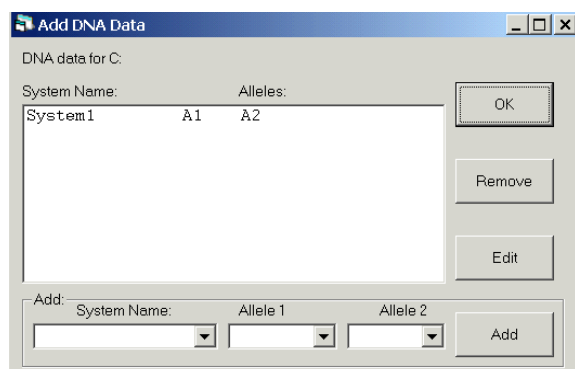


Figure 3.7: *Adding DNA data for a selected person is done in this window.*

Read case data

Data for specific samples can now also be read from files. The data for specific samples should be given as a table. The format can be outputted from Excel, using tabs as separators (from Excel, save as Text (Tab delimited)). The table should have a line with headings, and the following lines should each represent a sample source, i.e., a person. Blank lines (i.e., lines where there is nothing in the first column) will be ignored. The first column should list

the names of the sample sources, i.e., the persons. If the names correspond to names of persons already entered, the data will be added to the data for this person. Otherwise, the persons will be added as they are read in. The data for the systems must be provided prior to reading case data. There must be two columns specifying sex chromosomes in the table. These columns must be beside each other, the first must contain the letter “X” as all entries, and the second must contain either “X” or “Y”, depending on the sex. (Remember that **familias** is case sensitive. The X and Y should be in capitals.) When new persons are added, they will be given the sex specified by these columns. For existing persons, the data is ignored. Except for the three columns described above, all columns must come in pairs of two, beside each other, with the headings specifying the name of the allele system the columns contain data for. The headings for each pair must be identical, except for the last character (which could be, for example “1” and “2”). After the last character has been removed, the remaining name (removing blanks at the end) must correspond exactly to the name of an already entered allele system. The two columns below then contain the names of the alleles observed in this system, for the respective persons. Note that homozygotes must have alleles entered twice, once in each column. Missing data are coded with a ‘*’. Both (or none) alleles must be missing for a marker. An example of an input file is given below.

name	ame11	ame12	SYS1	1	SYS1	2	SYS2	1	SYS2	2
na1	X	X		F		G		08		09
na2	X	Y		G		G		10		11
Jakob	X	Y		G		G		09		10

Table 3.2: Example of case data that can be read into **familias** from the Case Related DNA Data window. The system called **SYS1** and **SYS2** must be given on beforehand, for example by reading the data of Table 3.1 above. The names (*na1*, *na2* and *Jakob*) may or may not be given. The loading of the data is explained previously.

Write case data

The case data can be written to a file of the same format as used for input files.

3.5 Pedigrees



In this form you may add your own pedigrees or you may use **familias** to generate pedigrees, this latter option is rarely used and is discussed in [Appendix A3](#). After having generated the pedigrees, one can calculate probabilities and likelihoods ratios and produce reports. In the following we will go through the entire set of buttons and options of the window shown in Figure 3.8 starting in the upper right corner.

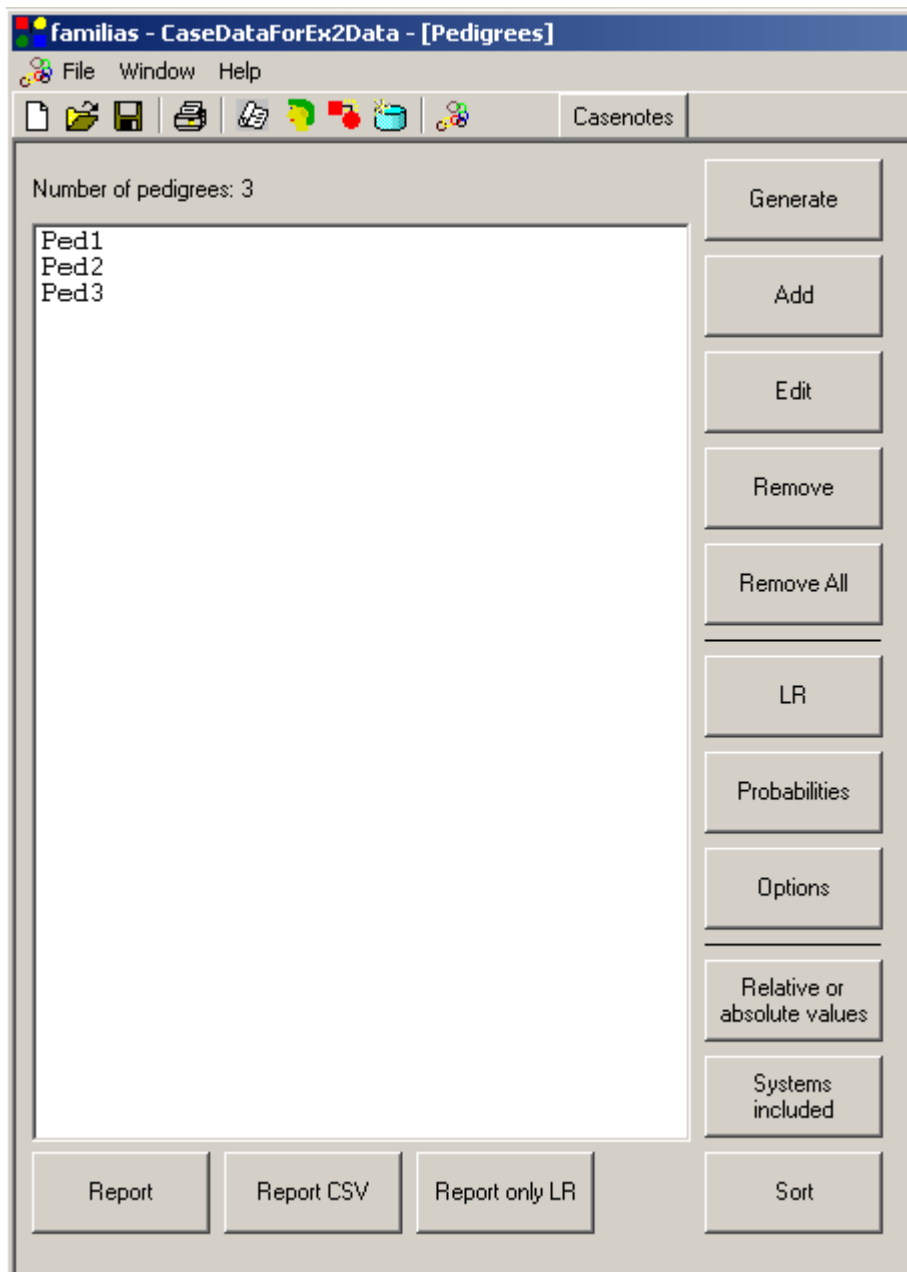


Figure 3.8: *The Pedigrees window with 3 generated pedigrees.*

Generate: Generate pedigrees automatically

See [Appendix A3](#).

ADD: Creating pedigrees

Usually, the first thing to do is to create a set of pedigrees manually by clicking ADD. The pedigree is defined by giving the parent child relations as exemplified in Figure 3.9. The **Persons** button is used to introduce individuals needed to define a pedigree. For instance, such extra persons may be needed to define cousin relations. There are various examples of

Manual for **familias**

pedigrees can be defined in the exercises and solutions provided in Appendix 2, see for instance [Exercise 2](#). The pedigree name can be edited.

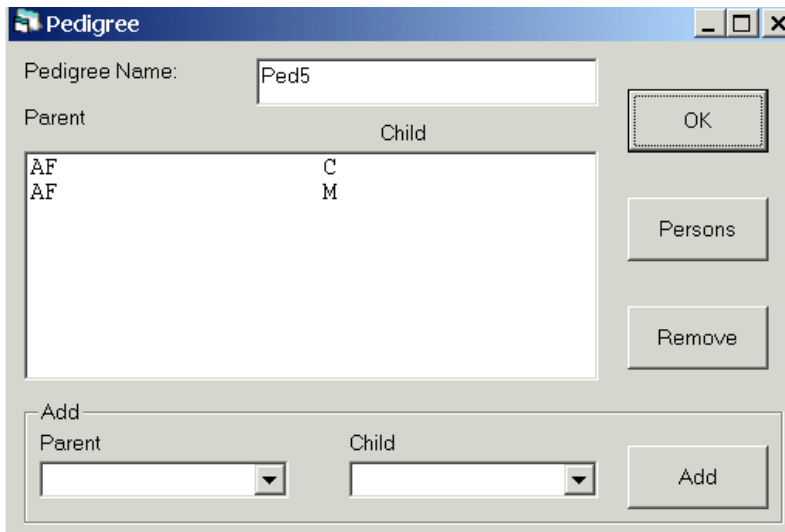


Figure 3.9: *AF is defined as the father of C and M. C and M are thus half-sibs. An extra parent is needed to define full sibs.*

EDIT: Editing pedigrees

The pedigrees are edited using this button.

Remove

This button is used to remove a single marked pedigree.

Remove all

This button is used to remove all pedigrees.

LR

The likelihood ratio is calculated wrt to the first pedigree.

Probabilities

The posterior probabilities are calculated.

Options

Various parameters can be set. The most frequently used is the kinship parameter ($\theta = F_{ST}$). The remaining parameters are explained in [Appendix A4](#).

Relative or absolute values

The LR button previously described gives the LR wrt the first pedigree. The option **Relative or absolute values** gives the LR wrt to a marked pedigree. Pressing this button twice gives the likelihoods.

Systems included

By default all systems are used for calculation. This option can be used to give results for selected systems.

Report only LR

Writes a selected LR to a file.

Report CSV

Writes a CSV file (which can be read into e.g. Excel) containing the overall LR and also for individual systems. The reference is the first pedigree.

Report CSV

A complete report is written summarising input and output.

A Appendices

A1 Theory and methods

The method **familias** is based on may be divided into the following stages: First, we describe the set of possible pedigrees involving the relevant persons. This may sometimes be a very large number. Secondly, we assign a prior probability distribution to this set of pedigrees, based on non-DNA evidence. Finally, we introduce DNA measurements and mutation parameters, obtaining a posterior probability distribution on the pedigree set. Likelihood ratios (LR-s) may also be calculated and then prior distributions are not needed.

familias determines relationships between persons through parent-child relations. When you define persons in **familias**, you distinguish persons based on those who may have children and those you know do not have children. This distinction will typically be made based on age. It is thus possible to define a person as a child. If no such information is available, then the safest alternative is to classify all the persons as adults. Next, the persons involved are characterised according to gender.

Based on the information above, one may generate all possible pedigrees containing only these individuals. However, one will frequently be interested in pedigrees involving persons not included in the original group. For example, to describe that a woman has three children with the same man, it is necessary to include this man in the pedigree, even though his DNA is unavailable. The implemented approach introduces a number of “extra” men and “extra” women and generates all possible, different pedigrees.

A1.1 Prior model

The set of pedigrees generated should contain the pedigrees we consider probable given the background information, but will also contain a large number of pedigrees that are unlikely for different reasons. For example, many very incestuous pedigrees will be generated; in most cases, they should not be considered a priori as likely as non-incestuous pedigrees. Similarly, most pedigrees will indicate a more promiscuous behaviour than is usual in most cultures.

familias generates a probability distribution on the set of pedigrees reflecting such considerations. Starting with an equal probability distribution on the pedigree set, we may choose to modify the prior probabilities of different pedigrees using the three options *inbreeding*, *promiscuity* and *generations*. The first parameter may be used to increase or decrease the probabilities of pedigrees involving inbreeding. A similar comment applies to promiscuity, while generations allude to the modification of probabilities of pedigrees extending over several generations. The prior distribution is proportional to

$$M_I^{b_I} M_P^{b_P} M_G^{b_G} \quad (1.1)$$

where M_I , M_P and M_G are non-negative parameters provided by the user of the program. The subscripts refer to the three mentioned options. The corresponding integer exponentials b_I , b_P and b_G explained next are calculated by **familias**. b_I is the number of children whose

parents have a common ancestor in the pedigree. For *promiscuity*, the number of pairs having precisely one parent in common is calculated and denoted b_p . The number of persons in the longest chain of generations starting with a named person and ending in an adult named person is calculated and assigned the value b_G . In addition, it is possible to discard automatically all pedigrees where the number of generations b_G exceeds a prescribed level.

Letting $M_I = 0$, the prior probability of all incestuous pedigrees is 0. A value of the parameter between 0 and 1 decreases the probability of incestuous alternatives in comparison to non-incestuous ones, while a value exceeding 1 increases the probability of incestuous constellations. A similar comment applies to the other options. A small, artificial example illustrates some of the concepts above. Assume three men, M1, M2 and M3 are found dead and two alternatives are considered: H_1 : M1 is the father of M2 who is the father of M3 and H_2 : M1 is the father of M2, while M3 is unrelated to M1 and M2. The ratio of the priors corresponding to alternatives H_1 and H_2 follows from Equation (1) as

$$\frac{M_I^0 M_p^0 M_G^3}{M_I^0 M_p^0 M_G^2} = M_G$$

We emphasise that this prior is but one pragmatic suggestion among many others possible; in many cases they are not needed. The default of the parameters M_I , M_G and M_p is by **familias** set to equal 1 and therefore implies that all pedigrees have, a priori, the same probability.

A1.2 Posterior model

According to Bayes' theorem the *posterior probability ratio* (PPR) may be written as

$$\text{Posterior probability ratio} = \text{Likelihood ratio} \times \text{Prior probability ratio}$$

In a more mathematical terminology

$$\frac{\Pr(H_p | E, I)}{\Pr(H_d | E, I)} = \frac{\Pr(E | H_p, I)}{\Pr(E | H_d, I)} \times \frac{\Pr(H_p | I)}{\Pr(H_d | I)} \quad (1.2)$$

where E typically stands for evidence, more precisely DNA-data, and I is some conditioning information like for example age. Relating to forensic medicine, the term H_p is the prosecution hypothesis and the defendant hypothesis is denoted H_d . Usually it is the likelihood ratio (LR) that is reported in court.

It remains to explain the calculation of the likelihood $\Pr(E | H, I)$. A version of the Elston-Stewart algorithm is implemented (Elston and Stewart 1971). The algorithm is extended to account for possible substructure, silent alleles and mutations and these extensions are explained in the coming sections.

A1.3 Substructure corrections

The probability of a set of DNA-data is calculated by looking at the different loci separately before multiplying the results. For all individuals, a locus of the DNA consists of two alleles, which can be either equal, constituting a homozygous locus, or different, giving a heterozygous locus. The probability of a particular combination of alleles (the genotype) is in the simplest cases calculated by means of Hardy-Weinberg's law. This law states that the probability of being either heterozygote A_iA_j or homozygote A_iA_i is given by

$$\Pr(A_iA_j) = \begin{cases} P_{ii} = p_i^2 & \text{if } i = j \\ P_{ij} = 2p_i p_j & \text{if } i \neq j \end{cases} \quad (1.3)$$

where p_i is the frequency of allele A_i in the population.

Assuming the following conditions are satisfied:

- i. random mating,
- ii. no selection,
- iii. no mutation,
- iv. no migration,

the population in question is at so-called Hardy-Weinberg equilibrium, and Equation (3) is valid.

In situations where mutations and non-random mating occur, the assumptions in Hardy-Weinberg's law are no longer necessarily satisfied. As mentioned, Hardy-Weinberg's law may not apply in the presence of population stratification and relatedness. To handle this, **familias** incorporates a kinship parameter, which is set by the user. The parameter corresponds to the traditional F_{ST} known from population genetics (see, e.g., [1]). It takes into consideration that within a subpopulation there tends to be a higher frequency for homozygosity than if Hardy-Weinberg equilibrium is obtained.

If p_i is the frequency of A_i in the population, then the genotypic frequencies are described by

$$\Pr(A_iA_j) = \begin{cases} F_{ST} p_i + (1 - F_{ST}) p_i^2 & \text{if } i = j \\ 2(1 - F_{ST}) p_i p_j & \text{if } i < j \end{cases} \quad (1.4)$$

Generally, the complete correction (sometimes referred to as θ -correction) described in (Balding and Nichols 1994) is implemented.

The differences between probabilities calculated with and without incorporating kinship can be quite large. For example, the probability of a genotype (A, A) when $p_A = 0.05$, is 0.00250. However, using a kinship parameter of 0.01, this probability becomes 0.00298.

It can be problematic to decide an appropriate value for the kinship parameter. One suggestion is to use 0.01-0.05 for Europeans and 0.1-0.3 for more divergent populations.

A1.4 Mutation models

There are four different mutation models available in **familias** [3]. The mutation model is specified for each allele system, and can be different for males and females. The alternative models are:

1. Prob decreasing with range (stable)
2. Prob decreasing with range (equal).
3. Probability proportional to frequency (stable)
4. Equal probability (simple and fast)

The stable models (1 and 3) are stationary. This means that if the allele probabilities for an individual are the population allele frequencies, then allele probabilities will still be allele frequencies after one or more generations of mutations. Equivalently, the allele probabilities will converge towards the population frequencies as the number of generations with mutations increase. An unpleasant consequence of using a non-stationary (non stable) model is that the likelihood ratio will change by including extra irrelevant persons in the calculation. We provide details of the models below.

A1.4.1 Prob decreasing with range (equal).

It is convenient to first describe model 2. In the decreasing model we assume that the list of alleles is expanded to include all “possible” alleles, and that they are listed by increasing lengths. The probability of mutation from allele a to allele b decreases in this model as a function of the difference in length between the alleles. This property is illustrated in Figure A.1, where the thickness of the arrows illustrates the probability of the transitions. The transition matrix M for this model is given by:

$$M = \begin{bmatrix} 1-R & k_1 r^{|1-2|} & \cdot & \cdot & k_1 r^{|1-N|} \\ k_2 r^{|2-1|} & 1-R & \cdot & \cdot & k_2 r^{|2-N|} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ k_N r^{|N-1|} & \cdot & \cdot & \cdot & 1-R \end{bmatrix},$$

where R is the overall mutation rate, r is a constant between 0 and 1 ($0 < r < 1$). The r parameter is provided by the user and is **Mutation range** in **familias**. k_i is chosen such that $\sum_{j=1}^N m_{ij} = 1$.

A calculation gives $k_i = \frac{R(1-r)}{r(2-r^{i-1}-r^{N-i})}$.

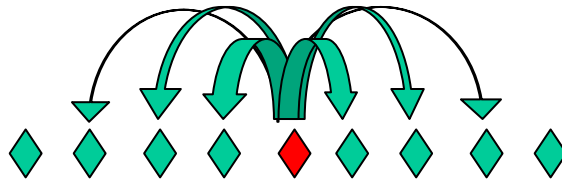


Figure A.1: *Mutation model*

A1.4.2 Prob decreasing with range (stable)

This model is explained in (Dawid, Mortera et al. 2002) and is a stationary version of the previously described model. Below we provide some details beyond those presented in the mentioned paper.

We want to generate stable mutation models. Recall that a mutation model can be represented as a square matrix $M = [m_{ij}]$ where m_{ij} is the probability of mutating from allele i to allele j . The fact that these values are probabilities is contained in the requirement $M1 = 1$ where 1 is the column vector of ones, and in the requirement that all elements of M are non-negative. Let p be the column vector of allele population frequencies and p' the transposed (row) vector. Then M is stable iff $p'M = p'$

How can one modify a mutation model so that it becomes stable? Clearly this can be done in many ways, but an attractive alternative would be to adjust, for each allele, the probability that a mutation occurs, while keeping unchanged the relative probabilities of the identities of the resulting mutated alleles after a mutation. In terms of a mutation model matrix, this corresponds to adding (or subtracting) various values along the diagonal, while adjusting the remaining values so that the numbers on each line still sum to 1. Technically, let A be a mutation model, i.e., $A1 = 1$ and all elements non-negative. Then we will find a stable version of it by writing $M = DA + I - D$, where D is a diagonal matrix. We get $M1 = DA1 + 1 - D1 = 1$, so M is a mutation matrix, as long as D is defined so that the elements of M are non-negative: This means that $d_{ii} \geq 0$ and $d_{ii} \leq 1/(1-a_{ii})$. M is also stable iff $p'M = p'$, that is, if $p'DA + p' - p'D = p'$, i.e., iff $p'DA = p'D$, i.e., iff $v = Dp$ is a right eigenvector of A' belonging to the eigenvalue 1.

Assume A is symmetric, as it is in our examples. Then 1 is such an eigenvector, and we get a solution by defining D such that $1b = Dp$, where b is some positive scalar. Note that b must be small enough so that

$$d_{ii} \leq 1/(1-a_{ii}), \text{ i.e., } b \leq \min_i (p_{ii}/(1-a_{ii})).$$

Thus we can always generate a stable mutation model from a symmetric mutation model matrix, in the manner above.

Manual for **familias**

Define A by defining $a_{ij} = c^{|i-j|}$ for $i \neq j$ for some constant c , and define a_{ij} for $i = j$ so that $AI = I$. Then the stabilized matrix M becomes defined by $m_{ij} = ba_{ij} / p_i$ for $i \neq j$ and m_{ij} for $i = j$ again computed so that $MI = I$. We get

$$m_{ii} = 1 - [bc / p_i(1-c)](2 - c^{i-1} - c^{n-i})$$

The parameter c is assumed input from biological knowledge, while b is computed from the overall mutation rate R , using the following relation:

$$1 - R = p_1 m_{11} + p_2 m_{22} + \dots + p_n m_{nn}$$

giving

$$b = R(1-c)^2 / [2c(n-cn-1+c^n)] \quad (1.5)$$

With the user giving as input R and c , the program computes the mutation model M by first computing b as above, then computing the off-diagonal elements of M , and then the diagonal by requiring the rows to sum to 1. Note that the requirement that b cannot be too large translates to the requirement that for all i

$$R \leq 2(n-cn-1+c^n) / [(1-c)(2-c^{i-1}-c^{n-i})] p_i.$$

As another example, define $A = Ip'$. then clearly $AI = I$. To stabilize it, we choose a D such that $p'DA = p'D$. We may choose $D = kI$ for some constant k . We get that we must have $k \leq 1/(1-p_i)$ for all i , and, defining R as above, we get that

$$R \leq \frac{\sum_i (1-p_i) p_i}{(1-p_i)}$$

for all i .

A1.4.3 Probability proportional to frequency (stable)

In this model the probability of mutating to an allele is proportional to that allele's frequency. This model is as mentioned stationary. The transition matrix M for this model is given by:

$$M = \begin{bmatrix} 1-k+kp_1 & kp_2 & \cdot & \cdot & kp_N \\ kp_1 & 1-k+kp_2 & \cdot & \cdot & kp_N \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ kp_1 & \cdot & \cdot & \cdot & 1-k+kp_N \end{bmatrix}$$

where k is a constant. This model satisfies the stationarity condition $\sum_{i=1}^N p_i m_{ij} = p_j$. The overall mutation rate becomes $R = k \sum_{i=1}^N p_i (1 - p_i)$, therefore we must set the constant to be

$$k = \frac{R}{\sum_{i=1}^N p_i (1 - p_i)}.$$

Note that if the frequency of the entered alleles do not sum to 1, **familias** will assume there is a single extra allele making up for the rest of the probability when computing k . If this is not the case, k will be slightly wrong. Thus the frequencies of all the alleles in the system should be entered when using the proportional model.

A14.4 Equal probability (simple and fast)

In this model we assume that there are Q different alleles observed in a database and that $N \geq Q$ is the number of “possible” alleles. The model can best be described by means of a transition matrix M , where the elements m_{ij} denote the probabilities that alleles i are inherited as alleles j ($i, j = 1, \dots, N$). For this model, the probability of not mutating is for each allele $1 - R$, where R is the overall mutation rate. The probability of mutating to any of the possible other alleles is the same ($= R/(N - 1)$). This model is in fact stationary if and only if the allele probabilities are equal. So the transition matrix M is given by:

$$M = \begin{bmatrix} 1 - R & \frac{R}{N - 1} & \cdot & \cdot & \frac{R}{N - 1} \\ \frac{R}{N - 1} & 1 - R & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \frac{R}{N - 1} & \cdot & \cdot & \cdot & 1 - R \end{bmatrix}$$

Note that the “frequency” of an allele entered into **familias** is in fact interpreted as the probability of observing that allele. Thus, if the entered frequencies sum to 1, there is a zero probability of observing any other alleles, and the program requires that $N = Q$. To use $N > Q$, you need to make sure the probabilities input sum to (slightly) less than 1.

A1.4.5 An example illustrating the mutation models

This example is a paternity case with an alleged father (AF) with genotype (A, B) and a child (CH) with genotype (C, D) . This example corresponds to [Exercise 7](#) in Appendix A2. The population properties of the allele system (S1) are given in Table A1.

Table A1: *Properties of allele system S1.*

Allele label	A	B	C	D	E	F	G	H
Repeat number	14	15	16	17	18	19	20	21
Count	44	49	127	175	133	58	12	2
Proportion	0.073	0.082	0.212	0.292	0.222	0.097	0.019	0.003

We consider the following hypotheses:

- H_0 : AF is the father of CH.
- H_1 : AF and CH are unrelated.

We use a mutation rate of $R = 0.005$, and calculate likelihood ratios assuming the various mutation models.

The likelihood assuming H_0 is $p_A p_B (p_C (m_{AD} + m_{BD}) + p_D (m_{AC} + m_{BC}))$. The likelihood assuming the alternative hypothesis is $4 p_A p_B p_C p_D$. So the likelihood ratio is then

$$LR = \frac{\Pr(E | H_0)}{\Pr(E | H_1)} = \frac{p_C (m_{AD} + m_{BD}) + p_D (m_{AC} + m_{BC})}{4 p_C p_D}$$

- a) For the equal probability model (model 4) we set the number of possible alleles to 8, which leads to $m_{AC} = m_{AD} = m_{BC} = m_{BD} = 0.005/7$. The likelihood ratio then becomes

$$LR = \frac{0.212 \cdot 0.01/7 + 0.292 \cdot 0.01/7}{4 \cdot 0.212 \cdot 0.292} = 0.0029.$$

- b) For the proportional model (model 3) $m_{AD} = m_{BD} = k p_D$ and $m_{AC} = m_{BC} = k p_C$. Hence,

$$LR = \frac{2 p_C p_D k + 2 p_C p_D k}{4 p_C p_D} = k.$$

Furthermore, the constant k is equal to

$$k = \frac{R}{\sum_{i=A}^H p_i (1 - p_i)} = \frac{0.005}{0.800} = 0.0063.$$

- c) For the decreasing model (model 2) we use a mutation range $r = 0.5$. The individual mutation probabilities are

$$m_{AD} = k_1 r^3, m_{BD} = k_2 r^2, m_{AC} = k_1 r^2, m_{BC} = k_2 r,$$

where

$$k_1 = \frac{R(1-r)}{r(1-r^7)} = \frac{0.0025}{0.496} = 0.005, k_2 = \frac{R(1-r)}{r(2-r-r^6)} = \frac{0.0025}{0.742} = 0.003.$$

This leads to

$$LR = \frac{0.212 \cdot (0.005 \cdot 0.5^3 + 0.003 \cdot 0.5^2) + 0.292(0.005 \cdot 0.5^2 + 0.003 \cdot 0.5)}{4 \cdot 0.212 \cdot 0.292} = 0.0047$$

- d) We calculate $b = R(1-c)^2 / [2c(n-cn-1+c^n)] = 0.0004161$, matrices A and M as explained in Appendix 4.2 and find

$$m_{AD} = 0.00071, m_{BD} = 0.0012, m_{AC} = 0.0014, m_{BC} = 0.0025, LR = 0.0064$$

The different models lead to very small likelihood ratios as expected. However, the relative differences are considerable and the choice of model might well influence the overall LR considerably. Usually it will be a good idea to check the robustness of the conclusions by incorporating different mutation models.

A2 Solved exercises

This contains exercises for **Familias** and some suggested solutions.

In some cases, answers are given with ridiculously many digits. This is done since the exercises may also serve to demonstrate that the program calculates correctly. Mutations, kinship and silent alleles are disregarded unless otherwise stated.

Exercise 1

We would like to investigate the relationship between a female (called Female), a man (called Man) and a child (called Child). There is one locus, called S1, with alleles called A, B. The allele frequencies are all 0.05. The rest of the alleles are lumped into a 'Rest allele' with allele frequency 0.9. One possible relationship and genotypes of the individuals involved are shown in Figure 1.

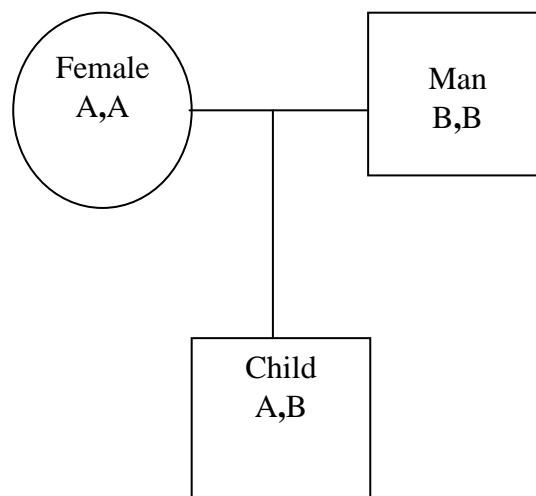


Figure 1: One possible pedigree, Exercise 1

Manual for **familias**

It is believed that the real pedigree is the above suggestion, in which the female Female and the male Man are the parents of the child Child.

- a) Generate the possible pedigrees without including any extra persons. How many pedigrees does **Familias** generate?
- b) Determine the posterior probabilities for the different pedigrees, assuming a flat (default) prior. What is the posterior probability of the above suggestion?
- c) Determine the paternity index (likelihood ratio) comparing the hypotheses
 - H_0 : The alleged father is the real father (Female is the mother)
 - H_1 : The alleged father and the child are unrelated (Female is the mother)

Exercise 2

In this exercise you are asked to evaluate the likelihood of possible family relations between a female (called Female) and a male (called Man). Female is believed to be 20 years older than Man. The analysis should be based on 13 loci and assumes for simplicity that all the allele frequencies are 0.05. The DNA-data may be summarized as in Table 1 (recall that manual typing may be avoided by loading the file AlleleSystemForExercise2.txt). The alternatives are

1. Man and Female are siblings,
2. Man and Female are half siblings,
3. Female is the aunt of Man,
4. Female is the mother of Man,
5. Man and Female are unrelated.

Table 1: The alleles of the female Female and male Man are shown for 13 loci

Locus	1	2	3	4	5	6	7	8	9	10	11	12	13
Female	ab	ac	bc	ab	ac	ab	ab	ab	ab	ac	ab	ac	aa
Man	ab	ab	ac	ab	bc	bc	ab	bc	bc	bb	ab	bb	ab

- a) Is the age difference of any use?
- b) Draw the five pedigrees (using pen and pencil) and give names to extra persons needed to define the pedigrees.
- c) Assume a flat (default) prior. Calculate the posterior probabilities of the 5 alternatives.
- d) What are the posterior probability ratios for the various alternatives when compared to Alternative 5 (the persons are unrelated)?
- e) Repeat c) with inbreeding parameter M_I equal to 0 and then 0.1. What are the probabilities now?

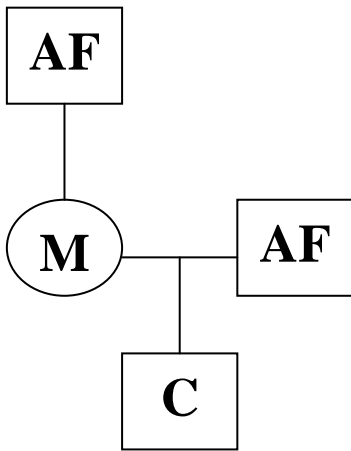
Exercise 3 (inspired by “Interpreting DNA evidence”, Evett and Weir, 1998)

A woman, M, gives birth to a child, C. The father of M, AF, is suspected to be the father of C, but there may be other explanations, incestuous or not. Assume that M, C and AF have genotypes (A_1, A_3) , (A_1, A_2) and (A_1, A_2) , respectively. The allele frequencies are 0.05. We consider five alternative pedigrees:

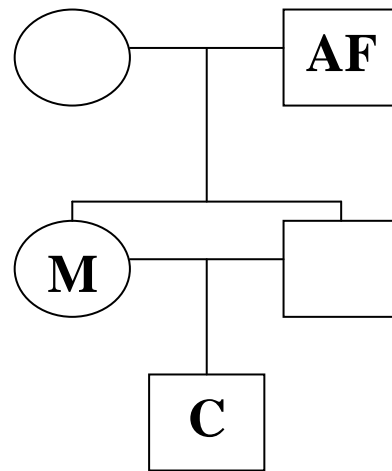
Manual for **familias**

- **Alternative 1:** AF, the father of M, is the father of C.
- **Alternative 2:** The brother of M is the father of C.
- **Alternative 3:** An unrelated man is the father of C.
- **Alternative 4:** A half-brother of M is the father of C.
- **Alternative 5:** A half-brother of C is the father of C.

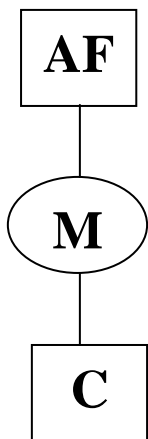
Alternative 1



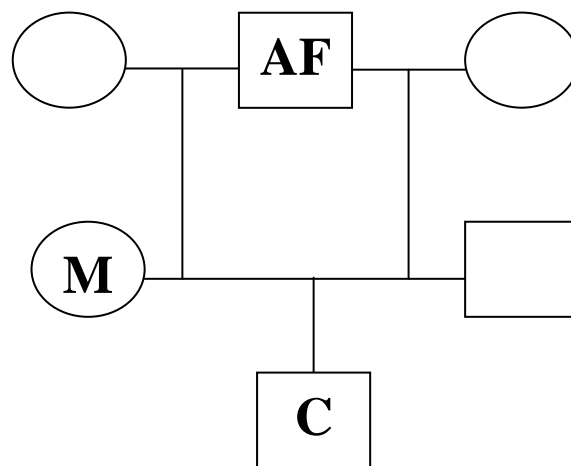
Alternative 2



Alternative 3



Alternative 4



Alternative 5

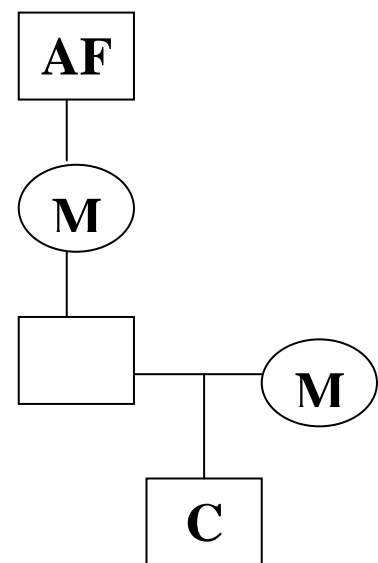


Figure 2 Five alternative pedigrees fro Exercise 2

a) What are the posterior probabilities for the five different alternatives?

- b) What are the posterior probabilities for the five different alternatives if you alter the inbreeding parameter to zero?
- c) Why is the probability in b) equal to 1 for the unrelated alternative?

Exercise 4

In this example there is an alleged father AF which is heterozygous (A, B), and a child CH which is homozygous (C, C). The allele frequencies are as listed in Table 2. The assumed pedigree is illustrated in Figure 3. The cross over the circle indicates that we have no DNA data for the child's mother. The letters in parenthesis indicate the alleles.

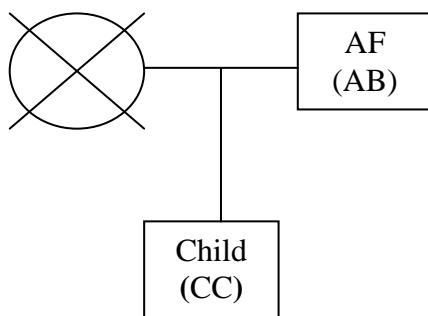


Figure 3: Assumed pedigree.

Table 2: Allele frequencies

A	0.1
B	0.2
C	0.3

- a) What would the probability of the supposed pedigree be if the mutation rate were equal to zero?
- b) Calculate the paternity index, that is, the likelihood ratio between the hypothesis naming AF as the father of CH and the hypothesis assuming an unknown man is the father of CH. Use mutation rates of 0.02 and 'Equal probability' model.
- c) Do you think the above mutation model is reasonable? Repeat b) with other choices.

Exercise 5

We would like to determine whether two girls (called sister1 and sister2 below) are sisters (corresponding to hypothesis H_1) or if they are half sisters (corresponding to hypothesis H_2). We know the alleles of their mother as well as the alleles of the sisters for the five allele systems S1-S5. The pedigrees are illustrated in Figure 4 and the DNA data are summarized in Table 3. The allele frequencies are 0.1 for systems S1 and S2, and 0.05 for systems S3-S5.

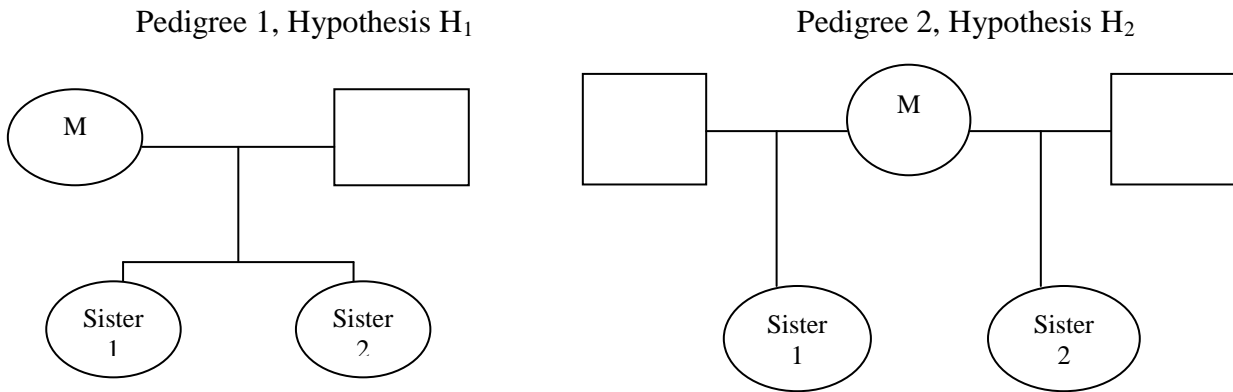


Figure 4: The pedigrees corresponding to the hypotheses of Exercise 5.

Table 3: DNA data

Person	S1	S2	S3	S4	S5
Mother (M)	A ₁ A ₂	A ₁ A ₂	B ₂ B ₃	B ₂ B ₃	B ₂ B ₃
Sister1 (CH1)	A ₂ A ₃	A ₂ A ₃	B ₃ B ₄	B ₃ B ₄	B ₃ B ₄
Sister2 (CH2)	A ₂ A ₃	A ₂ A ₃	B ₁ B ₃	B ₁ B ₃	B ₁ B ₃

- What does a flat prior mean in this case? Assume a flat prior. What are the posterior probabilities for the two hypotheses?
- What is the posterior probability ratio for the full sisters-alternative compared to the half-sisters-alternative?

Exercise 6

This is another paternity case, where M is the mother of CH, and where it's claimed that AF is the father of CH. The persons are apparently homozygous A, A and B for M, CH and AF, respectively. The allele frequencies are 0.1. Include a silent allele frequency of 0.05, and calculate the paternity index. Confirm the result by working out the formula.

Exercise 7

The purpose of this exercise is to consider the three different mutation models included in **Familias**. We have a paternity case with an alleged father AF with genotype (A, B) and a child CH with genotype (C, D). The allele system is given in Table 4 where A=14, B=15, C=16 and D=17.

Table 4: The allele system in Exercise 7.

Allele label	A	B	C	D	E	F	G	H
Repeat number	14	15	16	17	18	19	20	21
Count	44	49	127	175	133	58	12	2
Proportion	0.073	0.082	0.212	0.292	0.222	0.097	0.019	0.003

The below figure shows the mutation models of Familias. Observe that different models may be used for males and females.

The screenshot shows two panels for mutation models. The top panel is for 'Female mutations' and the bottom panel is for 'Male mutations'. Both panels have a 'Mutation rate' input field set to '0,005' and a button labeled 'As for males' (for females) or 'As for females' (for males). Under each panel, there are four radio button options: 'Prob. decreasing with range (stable)', 'Prob. decreasing with range (equal)', 'Probability proportional to frequency (stable)', and 'Equal probability (simple and fast)'. In both panels, the 'Prob. decreasing with range (equal)' option is selected, and a 'Mutation range' input field is set to '0,5'.

Calculate the paternity index assuming mutation rates of 0.005, and using the following models in turn (same model for males and females):

- 5) Prob decreasing with range (stable)
- 6) Prob decreasing with range (equal). Use mutation range 0.5
- 7) Probability proportional to frequency (stable)
- 8) Equal probability (simple and fast)
- 9) Explain the differences between the models and compare the results.

Exercise 8

Familias was not designed to handle Y-chromosomal data, so there are no "standard ways" of including such data. This exercise shows that it is possible to handle such data in simple cases. (However, it is probably easier to deal with Y data in other ways). Consider a standard paternity case, where data is available for a mother (Female), a son (Child), and a putative father (Man). We would like to determine the paternity index (likelihood ratio) comparing the hypotheses (assuming other males of the same Y-lineage are excluded)

Manual for **familias**

- H_0 : The alleged father is the real father (Female is the mother).
 - H_1 : The alleged father and the child are unrelated (Female is the mother).
- a) Initially, data is available for a Y-chromosome marker, Y1 with alleles A, B, C, D and frequencies 0.4, 0.3, 0.2 and 0.1, respectively. Assume Man and Child share A. Explain why the paternity index is 2.5
- b) Next repeat the calculation in Familias following the trick explained below. First enter data for an expanded system called, say, EY1:

Table 5 Allele data for y-chromosome exercise

Allele	Frequency	Transformed Frequency
A	0.4	0.2
B	0.3	0.15
C	0.2	0.1
D	0.1	0.05
dummy-X	0.5	0.5

Enter this marker using the transformed frequencies. Let the mother have two dummy-X alleles, and the putative father and the son one dummy-X and their actual markers, i.e., allele A, as data. Confirm the result in a).

- c) We next assume data is also available for a conventional marker called S1. The alleles in system S1 are respectively BB, AA and AB, for Man, Female and Child, respectively, and the allele frequencies are 0.05. Explain why the combined paternity index for markers Y1 and S1 is 50. Verify the result using Familias.
- d) Explain why the above approach works. Does the frequency assigned dummy-X matter?
- e) How far can this approach be generalized?

Exercise 9

This exercise deals with kinship and follows the approach suggested by Balding and Nichols (1995) and endorsed by NRCII. We will expand on Exercise 1. There is a man Man, a female Female and a child Child. The alleles in system S1 are respectively BB, AA and AB, and the allele frequencies are $p_1 = p_2 = 0.05$ for A and B (we consider only one locus).

- a) Determine the paternity index (PI) comparing the hypotheses
- H_0 : The alleged father is the real father (Female is the mother),
 - H_1 : The alleged father and the child are unrelated (Female is the mother).

Perform the calculations in Familias with F_{ST} ('Kinship parameter') 0, 0.001, 0.01 and 0.1

- b) Verify the results using the formula

$$PI = \frac{1 + 3F_{ST}}{2F_{ST} + (1 - F_{ST})p_B}$$

- c) We will next derive the above formula using the Dirichlet distribution. The crucial formula, which is more general than presently needed, is the recursion

$$E(\tilde{p}_A^{r+1} \tilde{p}_B^s \tilde{p}_C^t \tilde{p}_D^u) = E(\tilde{p}_A^r \tilde{p}_B^s \tilde{p}_C^t \tilde{p}_D^u) \frac{rF_{ST} + p_A(1 - F_{ST})}{1 + (r + s + t + u - 1)F_{ST}} \quad (1)$$

where A, B, C and D are arbitrary indices and r, s, t and u non-negative integers.

We need to calculate two likelihoods, the first assuming H_0 . The version without kinship, i.e., $p_1^2 p_2^2$ should now be replaced by

$$P(\text{data} | H_0) = E(\tilde{p}_1^2 \tilde{p}_2^2) \quad (2)$$

The likelihood under H_1 becomes

$$P(\text{data} | H_1) = E(\tilde{p}_1^2 \tilde{p}_2^3) \quad (3)$$

To be able to use equation (1), we make the identifications $p_1 = p_B$, $p_2 = p_A$, $r=2, s=2, t=u=0$. Show that Eq. (3) becomes

$$P(\text{data} | H_1) = E(\tilde{p}_1^2 \tilde{p}_2^2) \frac{2F_{ST} + (1 - F_{ST})p_B}{1 + 3F_{ST}}$$

and verify the expression for PI.

Exercise 10.

This exercise discusses the Romanov case documented in Gill et al. (1994) and subsequent papers. The Romanovs, i.e., Tsar Nicholay II, Tsarina and three of their five children as well three servants and a doctor were executed in 1918. A grave was found in Ekaterinburg in 1991 and various DNA analyses were performed to verify that the Royal Family had been found.

- a) Use Familias to show that
 - there are 4536 possible family relations between one male (called Tsar) and the four females (called Tsarina, Child1, Child2 and Child3),
 - there are 2020 possible family relations if incestuous alternatives are discarded,
 - there are 192 possible family relations if we assume Child1, Child2 and Child3 have no children of their one. How many are there if the incestuous alternatives are discarded?
- b) Consider the last alternative, i.e., the one with 192 alternatives. What is the prior flat distribution on this set of pedigrees?
- c) Read the romanov data into Familias. Verify the below table from Egeland et al. (2002) given below, using Familias. Verify also that the accepted solution, i.e., Tsarina and Tsar are parents of the children comes out as the most likely in all cases.

Table 6 Results of Exercise 10

Alternative models	Number of pedigrees	Ratio of two most likely alternatives
I “paternity case”	2	2692
II “Child1-3 have no children”	192	63
III “All except incestuous”	2020	63
IV “All (inbreeding par 0.1)”	4536	20
V “All”	4536	1.2

Exercise 11 (Database size. Familias not used)

We will base calculations below on the following data:

ALLELE	No	Frequency
14	44	0,073
15	49	0,082
16	127	0,212
17	175	0,292
18	133	0,222
19	58	0,097
20	12	0,02
21	2	0,0033

- A random person is typed. What are the probabilities of a homozygous {14,14} and heterozygous {16,17}?
- A random sample of 1000 is typed. How many are expected to be {14,14} and {16,17}?
- Assume the table above is based on a foreign database and we would like to make a database from our country assuming initially the same distribution applies. How many persons need to be typed before we can expect to observe the first {21,21}? How many persons need to be typed to be 95% certain to include a {21,21}?
- Discuss assumptions underlying the above calculations.

Exercise 12 (requires some knowledge of probability)

Parts of this exercise is no longer consistent with current Familias terminology and is therefore obsolete

In this exercise we will fill in some details related to the mutation models. Throughout we consider data from the system VWA:

Allele	A	B	C	D	E	F	G	H
Rep	14	15	16	17	18	19	20	21
No	44	49	127	175	133	58	12	2
Freq	.073	.082	.212	.292	.222	.097	.020	.003

The case data and hypotheses are:

- Alleged father : A,B
Child : C,D
- H₁: Alleged father true father
H₂: Alleged father and Child unrelated

Manual for **familias**

Assume mutation rate $R=0.005$.

a) Show that

$$LR = \frac{P(\text{data} | H_1)}{P(\text{data} | H_2)} = \frac{p_C(m_{AD} + m_{BD}) + p_D(m_{AC} + m_{BC})}{4p_Cp_D}$$

b) Consider the uniform mutation model with possible number of alleles, $M=10$. Explain why all mutation rates equal $R/(M-1)$

c) Calculate LR (answer 0.0023)

d) Consider next the proportional model. We introduce some notation

$$p_i, i = 1, \dots, n.$$

$$\sum_{i=1}^n p_i = 1$$

m_{ij} = probability of mutation i to j

Explain why

$$R = kp_j$$

$$k = R / \sum_{i=1}^n p_i(1 - p_j)$$

e) Show that $LR=k=0.0063$ for the proportional model.

f) Consider lastly the “range” mutation model. Explain why the following is a reasonable model assumption

$$m_{ij} = k_i r^{i-j}$$

Show that

$$k_i = \frac{R(1-r)}{r(2 - r^{i-1} - r^{n-i})}$$

g) Assume $r=0.5$ and calculate LR (answer 0.0047).

Exercise 13

Consider one locus S1 and a Man, born 1940, with genotype A,B; a Female, born 1960, with genotype B,C; and a Child with genotype A,C. All allele frequencies are 0.05.

- a) Verify that there are 8 possible pedigrees and that the flat prior produced by Familias is 0.125. (Hint: pedigree>Options and tick no ‘Case specific DNA data’)
- b) Are there any incestuous alternatives?
- c) What are the prior and posterior probabilities of the case where Man is the father of Female who is in turn mother of Child in the following cases given in the table below. (M_I , M_G and M_P denote respectively ‘Generations parameter’, ‘Inbreeding parameter’ and ‘Promiscuity parameter’ in Familias terminology).

	$M_I=1, M_G=1, M_P=1$	$M_I=1, M_G=0.5, M_P=1$	$M_I=1, M_G=2, M_P=1$
prior probability			

posterior probability			
--------------------------	--	--	--

d) Add a locus S2 with similar data and repeat item c)

Answers

Exercise 1

Answers:

- Familias** generates 12 different pedigrees (provided Child is entered as a child; otherwise, 22 pedigrees are generated).
- The probability of the suggestion in Figure 1 is 0.90497738.
- LR = 20.

How to do it:

- **General DNA data:** Click **Add** to enter an allele system. In the new window, enter allele system name S1 and the two alleles A and B, both with frequencies 0.05 and 'Rest allele' with frequency 0.9 . Press **OK**.
- **Persons:** Enter the persons Female, Man and Child and their gender. For Child also select **Is Child**.
- **Case-Related DNA data:** Double-click each person to enter his or her DNA data. In the new window, enter the appropriate allele system (use the pull-down menu) and the observed alleles for this person, then press **Add** and **OK**. Choosing the same allele twice enters homozygous loci. Repeat the procedure with the other persons.
- **Pedigrees:** Press **Generate**, and do not include any extra persons (just press **OK**). (If you generate pedigree several times, previous lists should first be removed clicking **Remove All**. Press **Probabilities** to calculate posterior probabilities for the alternative pedigrees. To identify which pedigree is which, double-click them. Press **LR** to obtain likelihoods instead of probabilities. This LR is scaled wrt to the first pedigree. Find and mark the pedigree that indicates only Female as mother of Man and press **Relative or absolute values**. Now the likelihood ratio of interest is wrt to the marked pedigree.

The analytical solution is derived in the manual and gives a paternity index of

$$PI = \frac{1}{p_B} = \frac{1}{0.05} = 20,$$

which is in concordance with the result obtained by **Familias**.

Exercise 2

Answers:

- The age difference is mostly useful when **Familias** generates pedigrees; here, it only prevents users from entering relationships that contradict the given ages.

Manual for **familias**

b)

c) The posterior probabilities of the 5 alternatives are listed below

- 1) Man and Female are siblings: 0.9986147
- 2) Man and Female are half siblings: 0.00139706
- 3) Female is the aunt of Man: 0.00125187
- 4) Female is the mother of Man: 0
- 5) Man and Female are unrelated: 0.00000001

d) The posterior probability ratios are

- 1) Man and Female are siblings: 1322405398
- 2) Man and Female are half siblings: 917236
- 3) Female is the aunt of Man: 917236
- 4) Female is the mother of Man: 0
- 5) Man and Female are unrelated: 1

e) With M_I equal to 0 and 0,1 the probabilities do not change. This is because there are no incestuous alternatives.

How to do it:

- **General DNA data:** The allele systems are all equal, and are constructed by entering alleles a, b, and in some cases c, all with frequencies 0.05. This must be repeated 13 times, that is, for loci 1, 2, ..., 13. Alternatively, read the database AlleleSystemForEx2.txt clicking **Read from File**.
- **Persons:** Enter the two defined persons, and choose the woman's year-of-birth 20 years prior to the man's (use for example 1940 and 1960).
- **Case-Related DNA data:** Enter the data given in Table 1. It is most practical to enter all information for one person first, before moving on to the next person. Alternatively read case data from the file CaseDataForEx2.txt clicking **Read from File**.
- **Pedigrees:** The pedigrees must be entered manually. Click **Add**. Assuming persons Mother and Father have been defined, alternative 1 is given in Figure 2 below and the parent child-relations can be entered. Press **Probabilities** to calculate posterior probabilities for the various pedigrees. Do not alter the settings for the priors. To obtain posterior probability ratios, mark the alternative pedigree and press **Relative or absolute values**. To calculate probabilities with altered inbreeding parameter, press **Options** and change the inbreeding parameter in the appearing window.

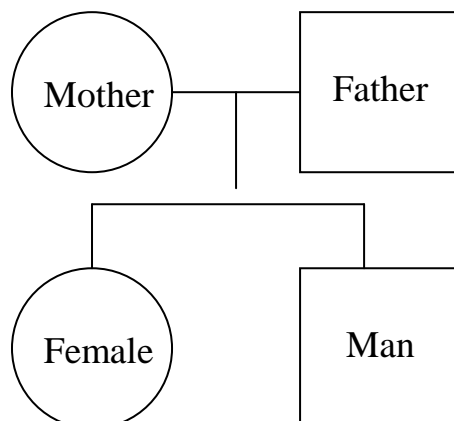


Figure 5: Entering the pedigree of alternative 1 in Exercise 2.

Exercise 3

Answers:

- a) The probabilities for the five different alternatives are listed below
 - **Alternative 1:** 0.4494382
 - **Alternative 2:** 0.23595506
 - **Alternative 3:** 0.04494382
 - **Alternative 4:** 0.24719101
 - **Alternative 5:** 0.02247191
- b) All the alternatives have posterior probabilities equal to zero except Alternative 3, which has probability 1.
- c) Since Alternative 3 is the only non-incestuous alternative, it gets a posterior probability of 1 when all the others get zero prior probabilities.

How to do it:

Enter the allele system with mutation rates zero, and the persons without information of birth dates. The child's gender is arbitrary. Under **Known relations**, enter AF as the father of M, and M as the mother of C. Under **Case-Related DNA data**, enter the data given. Construct the alternative pedigrees manually, using Figure 2 as guidance (you could alternatively practice making such figures yourself. Note that it is unnecessary to enter the mothers of M and of the extra male as extra persons: Everybody has a mother, and entering one with a single child in the pedigree is thus the same as not entering any. The other pedigrees are entered similarly.

Exercise 4

Answers:

- a) If the mutation rates were zero the supposed pedigree would not be possible, as the child could not have inherited any of his alleles from the alleged father.
- b) $PI = 0.022$.
- c) Answer omitted

How to do it:

Enter the allele system, the known persons (enter CH as a child) and their DNA data as usual. The pedigrees can be constructed either manually or automatically. The alternative pedigree involves no defined relations. Press **Probability** to obtain posterior probabilities, then **Likelihood** for the corresponding likelihoods. The likelihood ratio is then constructed by selecting the alternative (unrelated) pedigree and pressing **Scale**.

The solution is also available [here](#).

Exercise 5

Answers:

- a) A flat prior implies that the alternatives are equally likely a priori and so the prior distribution gives 0.5 to both alternatives. The posterior probabilities are 0.791 for them being full sisters, and 0.209 for the half-sister alternative.
- b) LR = 3.78.

How to do it:

Enter the allele systems or use the provided file. Enter the persons M, CH1 and CH2, and their DNA data as usual. Optionally, enter the known relations, that is, M as mother of both girls. Construct the alternative pedigree. Note that for the half-sister alternative, no extra males need to be entered. Press **Probabilities** to obtain posterior probabilities. Then select the pedigree where they are half-sisters and press **Relative or absolute values** to obtain LR.

Exercise 6

Answers:

The paternity index is 1.36.

How to do it:

Enter the allele system setting the silent allele frequency to 0.05. Enter the persons and their DNA data as usual. Construct the pedigrees manually. Obtain the paternity index (likelihood ratio) as explained in previous solutions.

Let $p_A = p_B = 0.1$ and $p_S = 0.05$. According to <http://dna-view.com/patform.htm>, the answer equivalent is given by

$$PI = \frac{p_s(p_A + p_s)}{(p_A + p_s)^2(p_B + 2p_s) + p_s p_A(p_B + 2p_s)} = \frac{0.05 \cdot 0.15}{0.15^2 \cdot 0.2 + 0.05 \cdot 0.1 \cdot 0.2} = 1.36,$$

which is in exact concordance with the answer obtained by **Familias**.

Exercise 7

Answers:

- 0.0064
- 0.0047
- 0.0063
- 0.0029

How to do it:

Manual for **familias**

The mutation models are edited by making changes in the window reproduced in the exercise

Exercise 8

Answers given in problem text.

Exercise 9

a and b)

Kinship	PI
0	20.0
0.001	19.3
0.01	14.8
0.1	5.3

c) Answer mostly given in text

How to do it:

Compared to previous exercises, the kinship parameter need to be provided in under options of the pedigree window.

Exercise 10-12

Answers in problem text or not given.

Exercise 13

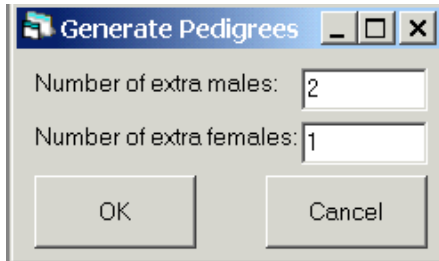
c)

	$M_I=1, M_G=1, M_P=1$	$M_I=1, M_G=0.5, M_P=1$	$M_I=1, M_G=2, M_P=1$
prior probability	0,125	0,0833	0,1667
posterior probability	0,128	0,0845	0,171

A3 Generating pedigrees automatically

The **Generate** button of the **Pedigrees** window can be used to generate pedigrees automatically. All possible pedigrees involving parent child relationships are generated. Keep in mind that as more persons are introduced, the number of generated pedigrees increases almost explosively. Often, as in the cases where only two pedigrees are to be compared, it is preferable to construct them manually. So far the largest number of pedigrees generated in a case is about 10000 (in test examples). There is no limit to the number of pedigrees produced, however, extreme cases may cause the program to “hang” [2].

When generating pedigrees, the program uses the information that some persons are designated as children (i.e., having no children) and the Year of Birth information. No pedigrees will be generated that imply a generation length of less than 12 years. The generated pedigrees are named Ped1, Ped2, ... etc. To view the details of a pedigree, double-click it; and the window in the figure below appears.



This is the same window that appears when pressing Add for manual construction of pedigrees. The pedigree is defined by the list of parent-child relations, and is thus altered by adding or removing these relations. You use the **Persons**-button to add the extra men and women that are necessary to define the wanted pedigree.

Figure A2. Adding extra persons.

As an alternative to adding anonymous extra persons here, the extra persons could have been defined in the **Persons** window described above. This is especially useful if one wants to put constraints on the number and types of possible pedigrees generated automatically, by introducing, e.g., extra persons that are of a certain age. Note that this may influence the computation of the **Generations** parameter .

A4 Implementation of prior distribution

After having entered the interesting pedigrees, one can calculate posterior probabilities for the various alternatives. By pressing **Probability** (see Figure 3.8), the window shown in Figure 3.11 appears. Here you are supposed to specify parameters that are used in the calculations of posterior probabilities, including the parameters defining the prior. The default corresponds to a non-informative prior, that is, where all the pedigrees get the same prior probability. After a possible change in the parameters the pedigrees' posterior probabilities appear. The pedigrees are now listed by decreasing probability.

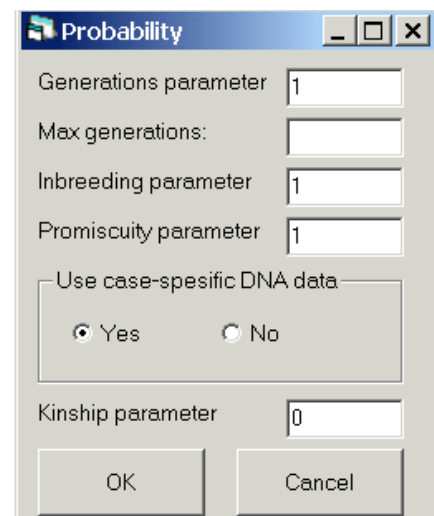


Figure A3. Calculating probabilities

The **Generations** parameter gives you the opportunity to modify the likelihood of pedigrees extending over several generations. More precisely, the calculated number is the number of persons in the longest chain of generations starting with a named person (not an “extra” person) and ending in an adult (not a “child”). For example, a pedigree consisting of a father and an adult son has generations value $b_G = 2$, while if the son is marked as “child”, the generations value is $b_G = 1$. By setting the generations parameter to a number between 0 and 1, short pedigrees are emphasized, and by using a number larger than 1, long pedigrees are emphasized. In addition, it is possible to define a cut-off length for the generation chain. By

specifying Max generations to, e.g., 2, you give a prior probability of zero to all pedigrees extending over more than two generations.

The Inbreeding parameter is used to alter the prior probability of incestuous pedigrees. More precisely, the value b_I computed is the number of persons in the pedigree such that its parents have a common ancestor. Thus, for example, a pedigree where cousins have tree children together gets a value $b_I = 3$, while one where siblings have one child gets a value $b_I = 1$. Setting the inbreeding parameter to zero is equivalent to giving a zero prior probability to all incestuous constellations. A value between 0 and 1 decreases the prior probability of incestuous alternatives relative to the non-incestuous ones, while a value exceeding 1 increases the probability of incestuous constellations.

The Promiscuity parameter is used to alter the prior probability of pedigrees involving “promiscuous” behaviour. More precisely, the value b_P computed is the number of pairs of half-siblings. Again: a value between 0 and 1 suppresses such pedigrees, while a value superseding 1 enhances them. Setting the parameter to zero gives a zero prior probability to all pedigrees where any person has children with more than one partner.

Example: Given the pedigree in Figure 3.12 and parameters provided by the user $M_G = M_I = M_P = 0.5$. The program will evaluate the pedigree and calculate the corresponding b – factors, which in this case are $b_G = 3$, $b_I = 3$ and $b_P = 3$, giving a value $M_G^{b_G} M_I^{b_I} M_P^{b_P} = 0.5^3 \cdot 0.5^3 \cdot 0.5^3 = 0.00195$. (Note that each of the children are half-siblings with their mother, giving $b_P = 3$). This means that this pedigree will be weighted down in comparison to a pedigree where the generations-, inbreeding- and promiscuity-factors are smaller, giving the pedigree in Figure 3.12 a smaller prior.

You may choose to calculate the posterior probabilities *without* the use of case-specific DNA data, by selecting “No” for the relevant variable. This will show you the prior probability distribution on the pedigrees, and is useful whenever you use a non-flat (non-default) prior.

The Kinship parameter takes values between 0 and 1.

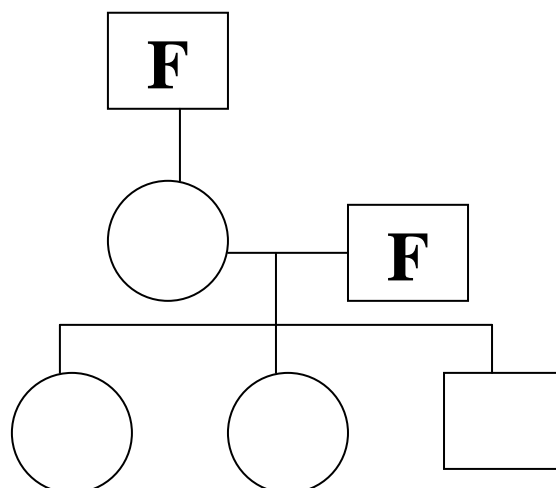


Figure A4: Pedigree

By pressing OK in the Probability-window, the posterior probabilities for all pedigrees appear, and the list is also sorted from the most probable to the least probable pedigree.

Note that, when computing probabilities, the program will remove all pedigrees found to be equivalent to a previous pedigree in the list. For example, if the user has entered the pedigree without any relationships several times, only the first of these will remain.

A5 Description of general input files for familias

The purpose of this document is to enable programmers to write code producing input files for the familias program, on the “familias format”. This format was designed to store all information contained in the user interface between runs of familias. Thus it was never designed to be readable or practical to use as a data exchange format. Nevertheless, it is released now as possibly the simplest way to exchange data between familias and other computer applications.

Versions

The familias format can often change for each release. However, backwards compatibility is ensured in that any version of familias is able to read the format output from all previous versions. The key here is that the version number of the program producing the file is always given in the third line of the file. The format described in this document is the one output from version 1.81. Thus it will not be readable by previous versions of familias. But it will be readable by future versions.

Simplified format

The “familias format” stores some information which is irrelevant when importing data from other programs in order to do computations with familias. For example, it is possible to store already computed probabilities, or the order in which pedigrees are listed in the pedigree list. In the table given below, lines specifying such information have been simplified and set to default values. However, the completely general use of these lines is explained in the notes below the table.

Format verification

As the format was not really intended to be produced by other programs, very little verification is done at the input stage that the input file really conforms to the correct format. For example, it is not checked that indices of persons and alleles are within the bounds given by the sizes of the lists, that pedigrees are legal as pedigrees, and so on. Run-time errors, or even faulty results, are likely to occur if the input files are incorrect.

How to interpret the table

The table below has four columns. The first contains line numbers: These are included for reference only; the actual format does not contain such numbers. The second column describes the information each line should contain. Information inside comparison signs,

Manual for **familias**

< like this >

is meta-information, and should be substituted with the corresponding content. Information without such signs should be included as written.

The third and fourth columns of the table indicate which lines should be repeated, and how they should be repeated.

File format

1	<Any text, in quotes; may describe the file >		
2	<Any text, in quotes; may describe the file >		
3	1.81		
4	< Number of persons involved in pedigrees >		
5	<Name of person >	}	
6	#FALSE#	}	
7	-1	}	Repeated
8	#FALSE#	}	for each
9	<#TRUE# if male, #FALSE# if female >	}	person
10	< Number of allele systems with data for this person >	}	
11	< Index of first allele, starting at zero >	}	Repeated
12	< Index of second allele, starting at zero >	}	for each
13	< Index of system, starting at zero >	}	allele system
14	< Any text , in quotes>		
15	0		
16	0		
17	0		
18	< Number of pedigrees >		
19	< Index of pedigree, starting at 0 >	}	
20	< Name of pedigree >	}	
21	0	}	Repeated
22	0	}	for each
23	< Number of relations in pedigree >	}	pedigree
24	< Index of parent in relation, starting at zero >	}	Repeated for
25	< Index of child in relation, starting at zero >	}	each relation
26	#FALSE#		
27	< Number of allele systems >		
28	#FALSE#		
29	< Name of allele system, in quotes >	}	
30	< Female mutation rate >	}	
31	< Male mutation rate >	}	
32	< Female mutation model >	}	
33	< Male mutation model >	}	
34	< Number of possibilities >	}	Repeated
35	< Female mutation range >	}	for each
36	< Male mutation range >	}	allele
37	< #TRUE# or #FALSE#: system has a silent allele >	}	system
38	< Silent allele frequency >	}	
39	< Number of alleles in the system >	}	
40	< Allele name >	}	Repeated for

41 < Allele frequency >

} each allele }

Explanation by line numbers

LINE NUMBERS	EXPLANATION
1-2	We encourage that the first two lines are used to describe the program producing the file. They are not read by the familias program.
3	The version number for the format: The file will be readable by familias with a version number higher than or equal to this number.
4	In the simplified format, we assume that all persons involved in the pedigrees have a name, and are listed in the list of persons further down. However, in general in familias , some pedigrees could be described using “extra persons” without names. These persons would then not be included in the total number of persons given on this line.
5	The name should be given in quotes, like “Mother”.
6	This line should be #TRUE# if year of birth data is included for this person, otherwise it should be #FALSE#. Note that familias uses the year of birth for a person only when automatically generating new pedigrees, so it should rarely be necessary to input such information.
7	When the previous line is #FALSE#, this line should contain -1. Otherwise, it should contain the year of birth for the person.
8	This line should be #TRUE# if the person is specified as a “child”, otherwise, it should be #FALSE#. The default value is #FALSE#. Note that familias only uses this information when automatically generating new pedigrees, so it should rarely be necessary to use anything but the default setting here.
10	Persons included just to describe the pedigrees properly, and for which no DNA data is available, should have 0 here. Note that it is not necessary that all the persons who do have DNA data have data from exactly the same allele systems.
11	The index of the first allele for this person in this allele system should be given: It is the index (when counting from zero) of the allele in the list of alleles for this system given further down in the file.
12	The index of the second allele; see line 11. Note that even homozygote persons must be input with two (equal) alleles.
13	The index of the allele system in which the two alleles above are contained: It is the index (when counting from zero) of the allele system in the list of such systems given further down in the file.
14	This line will generally contain the text “Known relations” in files produced by familias : The line is not used when reading in data.
15 - 17	In general, these lines are used to specify fixed relations, i.e., relations which occur in all pedigrees. In our simplified format, we assume that all relations are specified directly in the pedigrees. Thus these lines should just contain zeroes. In general, the format is as follows: Line 15 contains the number of “extra females” used to specify the fixed relations. Line 16 contains the number of “extra males”. Line 17 contains the number of fixed relations. Then follow pairs of lines, with one pair for each relation. The first line of each pair specifies the index of the parent of the

	relationship, and the second line the index of the child. The indices start at zero in the list of available persons: This list starts with the list of named persons, continues with the extra females, and ends with the extra males.
18	The number of pedigrees. For most applications, this number should be 2.
19	In our simplified format, this should just be 0 for the first pedigree, 1 for the second pedigree, and so on. In general, these numbers are used for the following purpose: The order in which the pedigrees are listed in the pedigree window can change after probabilities are computed; then they are listed by decreasing probability. When a familias file containing already computed probabilities is stored, the line indicates the place (starting at zero) of this pedigree in that ordered list.
20	The name of the pedigree. The name should be included in quotes, like "Ped1". The actual name can be chosen freely, as long as two pedigrees do not have identical names.
21	This indicates the number of "extra females" in the pedigree, i.e., females that are included the pedigree in order to describe it correctly, but for whom we have no DNA data. For most applications, we would recommend using named persons instead, and include these in the list of persons above.
22	The number of "extra males" in the pedigree: See line 21.
23	The number of relations in the pedigree.
24	The index (when counting from zero) of the parent in the relation, in the list of persons given above. When "extra persons" are included, the list is extended, first with the extra females, and then with the extra males.
25	The index (when counting from zero) of the child in the relation, in the list of persons given above. When "extra persons" are included, the list is extended, first with the extra females, and then with the extra males.
26	This line should be #TRUE# when already computed probabilities are included in the file. The default is #FALSE#. If computed probabilities are included, a number of lines are inserted between lines 27 and 28: For each pedigree, the following is listed: First the probability for the pedigree, and then, with one line for each allele system, the likelihood for the data in this system with this pedigree. Then follows a line with #TRUE# or #FALSE# according to whether DNA data was used to compute the probabilities (i.e., whether they are prior or posterior probabilities), a line with the kinship parameter used in computations, and finally lines specifying each of the following parameters used in the computations: The generations parameter, the max generations parameter, the inbreeding parameter, and the promiscuity parameter.
27	The number of allele systems
28	This should be #TRUE# if "database information" is included in the data. The default is #FALSE#. (This line is NOT repeated for each allele system.) If database information is included, an extra line is inserted below this line: It contains, in quotes, the string specifying the database information. (This is the string specified in the top field of the "General DNA data" box of the user interface).
29	The name of the allele system, in quotes, like "Allele system". Note that the name can be anything, but that different allele systems must have different names.
30-31	These are the female and male mutation rates; default is 0. The numbers must be non-negative and less than 1. If the mutation model is 1 or 3, there are also

	some (high) theoretical limits to how high the mutation rate can be; consult the manual for details.
32-33	Indices in the list of possible mutation models: 0 means "Equal probability (simple and fast)" 1 means "Probability proportional to frequency (stable)" 2 means "Prob. decreasing with range (equal)" 3 means "Prob. decreasing with range (stable)" Default is 3. See the allele system window in the user interface, and the manual, for further explanation.
34	The total number of alleles, including the silent allele.
35-36	The "range" of the mutation model: This is the same number that is input in the corresponding box in the user interface. It must be a positive number less than 1. For more detailed information, consult the manual. Default is .1.
37	Use #TRUE# if the system has a silent allele, use #FALSE# otherwise.
38	When there is no silent allele, use 0 as a default; otherwise use the frequency of the silent allele.
39	The number of alleles in the system, NOT including a possible silent allele. There must be at least 2 alleles in each system.
40	The name of the allele, in quotes, like "A1". Note that the names can be anything, but that different allele systems must have different names. Note also that for mutation models 2 and 3 (see lines 32-33) the order of the alleles matter. Thus the alleles are always assumed to be ordered alphabetically, and should be input in alphabetical order.
41	The frequency of the allele. Note that the frequencies of alleles in a system, including any silent allele, must add up to exactly 1. All frequencies must be positive numbers.

References

- Balding, D. J. and R. A. Nichols (1994). "DNA profile match probability calculation: how to allow for population stratification, relatedness, database selection and single bands." Forensic Sci Int **64**(2-3): 125-140.
- Buckleton, J. S., C. M. Triggs, et al. (2005). Forensic DNA evidence interpretation, CRC Press.
- Dawid, P. A., J. Mortera, et al. (2002). "Probabilistic Expert Systems for Forensic Inference from Genetic Markers." Scand J of Statistics **29**(4): 577-595.
- Drabek, J. (2009). "Validation of software for calculating the likelihood ratio for parentage and kinship." Forensic Sci Int Genet **3**(2): 112-118.
- Egeland, T., P. F. Mostad, et al. (2000). "Beyond traditional paternity and identification cases. Selecting the most probable pedigree." Forensic Sci Int **110**(1): 47-59.
- Elston, R. C. and J. Stewart (1971). "A general model for the genetic analysis of pedigree data." Hum Hered **21**(6): 523-542.