Manual for **familias**

Thore Egeland *
Ingvild Dalen
Petter F. Mostad**

*Medical Genetics
Ullevaal University Hospital
0407 Oslo, Norway
Email: thore.egeland@medisin.uio.no

**Biostatistics
National University Hospital
0027 Oslo, Norway
Email: p.f.mostad@medisin.uio.no

Version: August 12, 2005

Contents

0   Changes made for familias 1.6 and 1.7

Changes for familias 1.7

- When reading in case data on a text-table format from a file, one is asked whether one wants to change the name of the project to the name of this file. Thus one can import the project name along with the project data.

- When reading or writing data, a file name is now suggested, based on the current project name. This makes it easier to keep track of data files automatically via the project name.

- When reading or writing data, a directory is now suggested, based on what operation one wants to do, and what directory has been used for this operation previously.

- Some bugs in the user interface have been removed:
  - A bug occuring when "adding" data in an already existing system for a person.
  - A bug occuring when reading in data on text-format, and some persons already have data in the given systems.
  - A bug making the program protest that newly written files are from a newer version of the program.
  - A bug that made the program use unneccessarily long time to process allele systems with no data.
  - A bug that made the program use more time than necessary processing allele systems with zero mutation rates.
  - A bug making the program ask twice whether one wanted to store the data before doing computations.
  - A bug making it necessary to store and then open project files before the correct kinship value would be applied in computations.
  - A bug occuring in the fixed relations when removing people or fixed relations.

- Better error messages when reading in case data from file.

- There is a new button on the Pedigree box, called "Save only LR". For this function to be used, probabilities (using DNA data) must have been computed, and the "Scale" button must have been used. If more than two pedigrees are listed, the ones to be included in the report must be selected. Clicking the button makes it possible to save a file containing one line for each selected pedigree (except the one the others are scaled to), where the line will contain the relevant likelihood ratio.

- The information in the Report is now improved: More likelihoods are reported, and if the scale button has been used, then more likelihood ratios are also reported.

- A better handling of the situation when the user tries to input an allele system where the frequencies do not add to 1.

Changes for famlilias 1.6

The changes, mainly new features, made to familias 1.6 are documented below. Some bugs that have been fixed are not mentioned. This section is organised according to the five main windows of familias:

- General DNA Data.
- Persons.
- Known Relations.
- Case Related DNA Data.
- Pedigrees.

0.1. General DNA Data.

Scaling of allele frequencies

An important change is that the frequencies of allele systems are now required to sum to 1. (Previously, they could also sum to less than 1, in which case the program assumed one, or in some cases 2, extra unnamed alleles were present in the system. Furthermore, previously it was not possible to input an allele if this implied that the frequencies would sum to more than 1). This restriction will hopefully make the program more clear and easier to use. When inputting an allele system manually, or editing a system, the user will now be asked whether she would like to normalise the sum of frequencies to 1, if the sum is different from 1 (i.e., scale all the values so that they sum to 1). When reading in frequencies from a file (described below), such scaling will be done automatically, with a warning. When reading old familias files where the allele frequencies sum to less than 1, an "Extra allele" will be added.

Note that the old parameter "number of possible alleles" has been removed as a consequence of the above. To have a system with several unobserved alleles, these must be added individually.

Sorting of alleles

Alleles are now sorted according to name when they are input. This is practical 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.

Read system data from file

The file 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. 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
E	,326
F	,163
G	,056
H	,013
SYS2	
06	,056
07	,073
08	,190
09	,192
10	,253
11	,143
12	,089

Table 0.1 *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 textfile, 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.*

New mutation model

A new mutation model has been added. It corresponds to the model used by Dawid et al. (2002).

We want to generate stable mutation models. 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 vector of ones, and in the requirement that all elements of M are non-negative. Let p be the vector of allele frequencies. Then M is stable iff $p'M = p'$.

Let A be a mutation model, i.e., $A1 = 1$ and all elements non-negative. Then we will try to "stabilize" it as follows:
Define $M = DA + I – D$, where D is a diagonal matrix. Then 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 $d_{ii} >= 0$ and $d_{ii} <= 1/(1-a_{ii})$. 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 = D'p is an eigenvector of A' belonging to the eigenvalue 1.

Assume A is symmetric. Then 1 is such an eigenvector, and we get a solution by defining D such that 1b = D'p, where b is some positive scalar. Note that b must be small enough so that d_ii <= 1/(1-a_ii), i.e., b <= 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.

Define A by defining a_ij = c ^|i-j| for i != j, and a_ii computed so that A1 = 1, for some constant c. Then the stabilized matrix M becomes defined by m_ij = ba_ij/p_i for i!=j and m_ii again computed so that M1=1. 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_1m_11 + p_2m_22 + … + p_n m_nn
Giving

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

With the user giving as input R and c, compute the mutation model M by first computing b as above, then compute 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 <= 2 (n-cn-1+c^n) / [(1-c)(2-c^(i-1)-c^(n-i))] * p_i

As another example, define A = 1p'. then clearly A1 = 1. To stabilize it, we choose a D such that p'DA = p'D. It is easy to see that we may choose D = kI for some constant k.
We get that we must have k<=1/(1-p_i) for all i, and, defining R as above, we get that

R <= crossum/(1-p_i) for all I, where crossum is the sum of all p_i(1-p_i) for i=1 to n.

0.2. Persons

Editing persons

It is now possible to edit the properties of persons in the "Persons" form. This is now done in the "Add" field of the form, which temporarily changes name to "Edit". The DNA data of an edited person is always kept, but if the person changed anything more than the name, the person is first removed and then added to the data structure, which means that the person will disappear from pedigrees, known relations, etc. This is necessary, as changing the sex, for example, of a person, necessarily changes its involvement in pedigrees.

0.3. Known Relations

No changes have been made.

0.4. Case Related DNA Data

Read case data

Data for specific samples can now also be read in from files. The data for specific samples should be given as a table, also in a format that can be output from Excel, using tabs as separators. 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. 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. An example of an input file is given below.

```
name   amel1 amel2 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 0.2 *Example of case data that can be read into familias from the Case Related DNA Data window. The system called SYS1 and D13 must be given on beforehand, for example by reading the data of Table 0.1 above. The names (na1, na2 and Jakob) may or may not be given. The loading of the data is explained previously.*

0.5. Pedigrees

Report

The output from the "Report" function has been improved. The likelihood for each locus is now included.

Active loci

It is now possible to select the loci to be used for calculations. By default, all loci are used.

1   Introduction

There is a section preceding this one, Changes made in familias 1.6. The idea is that experienced familias users need not read more than the mentioned section. What follows is an update of the manual for familias. The intention with the **familias** program is to enable the user to determine the most probable familial relations within a group of persons. The program determines the relationship based on DNA-data provided by the user. It is also possible to specify other information, like gender, age and population data, if available. Pedigrees are created either automatically or manually, and then posterior probabilities are calculated based on a combination of DNA data and prior probabilities based on non-DNA data. Alternatively, one can choose to calculate likelihoods, ignoring the latter source of information. In either case, the program provides the possibility of including and comparing several different alternative pedigrees, thereby extending for example the standard paternity case with only two hypotheses. In fact, **familias** may even be used for other species than humans.

There are several applications for this program, 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 present document is a manual for **familias**. The program is obtainable as freeware from http://www.nr.no/familias. Several example data files, including those for the exercises of this manual, are available from the site. In addition to this online help, a short tutorial is available directly from the help-function of the program.

The contents of this document are as follows. Section 2 gives a brief introduction to the theory and methods on which the program is based. Next, Section 3 provides an overview of the options available in the program, along with suggestions for typical values for the various parameters. There is a tutorial in Section 4. Finally, some more theory is included in the Appendix.

2 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.

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.

2.1 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

$$M_I^{b_I} M_P^{b_P} M_G^{b_G} \quad , \tag{1}$$

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: *H₁: M1 is the father of M2 who is the father of M3* and *H₂: 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.

2.2   Posterior model

The DNA for each locus for all persons having such data available as well as mutation rates for each system is used to compute the likelihood of all pedigrees considered in the prior model. These likelihoods are multiplied with the priors to obtain the posterior probability. All details related to the computation of the likelihoods are provided in [4].

2.3   Terminology

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

Posterior probability ratio $=$ Likelihood ratio $\times$ Prior probability ratio

In a more mathematical terminology

$$\frac{\Pr(H_p \mid E, I)}{\Pr(H_d \mid E, I)} = \frac{\Pr(E \mid H_p, I)}{\Pr(E \mid H_d, I)} \times \frac{\Pr(H_p \mid I)}{\Pr(H_d \mid I)}, \qquad (2)$$

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.

2.4   Calculating probabilities (Hardy-Weinberg's law)

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 $A_i A_i$ or homozygote $A_i A_j$ is given by

$$\Pr(A_i A_j) = \begin{cases} P_{ii} = p_i^2 & \text{if } i = j \\ P_{ij} = 2 p_i p_j & \text{if } i \neq j \end{cases}, \qquad (3)$$

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. For a more detailed explanation of the mutation models included in **familias**, see Appendix A.1. For more extensive theory the reader is referred to [5].

2.5   Structured populations

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.

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

$$\Pr(A_i A_j) = \begin{cases} F_{ST} p_i + (1 - F_{ST}) p_i^2 & \text{if } i = j \\ 2(1 - F_{ST}) p_i p_j & \text{if } i < j \end{cases}, \qquad (4)$$

which are the actual formulas used by **familias**.

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 (see [6] for further details).

2.6   Example

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

- H_p: *M1 is the father of M2*
- H_d: *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.
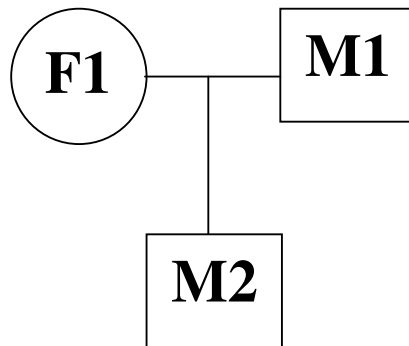
Figure 2.1: *The pedigree of hypothesis* H_p.

Given the following genotypes: F1 = {A, A}, M1 = {B, B} and M2 = {A, B}. Then it is certain that the child has inherited the allele A from his mother and therefore the allele B must be inherited from the father. Assuming Hardy-Weinberg equilibrium, the likelihood of the data given H_p is given by

$$\Pr(F1=\{A,A\}, M1=\{B,B\}, M2=\{A,B\}\,|\,H_p) = p_A^2 p_B^2,$$

where p_A and p_B are the frequencies of the alleles A and B, respectively. The likelihood of the data given hypothesis H_d is given by

$$\Pr(F1=\{A,A\}, M2=\{B,B\}, M2=\{A,B\}\,|\,H_d) = p_A^2 p_B^3.$$

Manual for **familias**

The likelihood ratio is then given by

$$LR = \frac{p_A^2 p_B^2}{p_B p_A^2 p_B^2} = \frac{1}{p_B}.$$

If the hypotheses are considered equally probable a priori, then the prior probability ratio equals 1, and so the posterior probability ratio equals the likelihood ratio.

3   User's guide

In this chapter we explain how to use the **familias** program. After completing this section and possibly the tutorial of Section 4, you should be well equipped to attack problems in your own work. The main menu of **familias** is illustrated in Figure 3.1 below. The first four buttons are common to most windows programs. They stand for 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. 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, see for example under the links provided in Appendix A.2. 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), adjust priors, and perform calculations of probabilities and likelihoods. A shorter introduction may be obtained directly from the help function of the program.

Figure 3.1: *The main menu of familias.*

Manual for **familias**
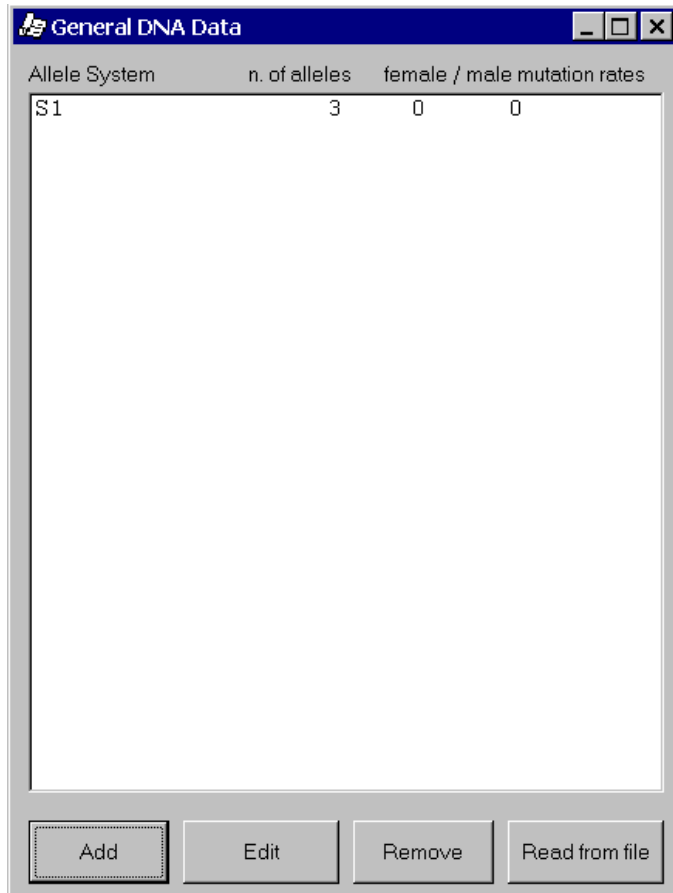
Figure 3.2 *A system may be added, edited or removed manually. It is possible to read data from a file*.

3.1   General DNA data

This window provides options for adding, editing, removing and reading 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 maunually, press Add. Then the Allele system window, illustrated in Figure 3.3, appears. Here you enter the system name and the belonging existing alleles and their respective frequencies.

Sorting

Alleles are now sorted according to name when they are input. This is practical 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 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 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. 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
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 textfile, 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*

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.

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 differ between male and female mutation rates. The reason for this is that paternal alleles tend to mutate more often than maternal alleles. There are 3 different mutation models to choose from: *Decreasing* (there are two options for this model called stable and equal), *Proportional* and the *Equal probability* model. These are explained in Appendix A.1, 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:

- Probability decreasing with range is your choice for the decreasing model. The stable option gives a stationary model (see Dawid et al., 2002) whereas the equal model is the one called decreasing in Familias 1.5. Under the decreasing model the probability of a mutation decreases with the range between the parent allele and the offspring allele. 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. A typical value of the mutation range is 0.1, which 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.
- Probability proportional to frequency applies to the proportional model. 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.
- Possible alleles with equal probability applies to the equal probability model. In this model you assume that the probability of mutating from one allele to another allele is the same independently of the frequency and the range of the alleles.
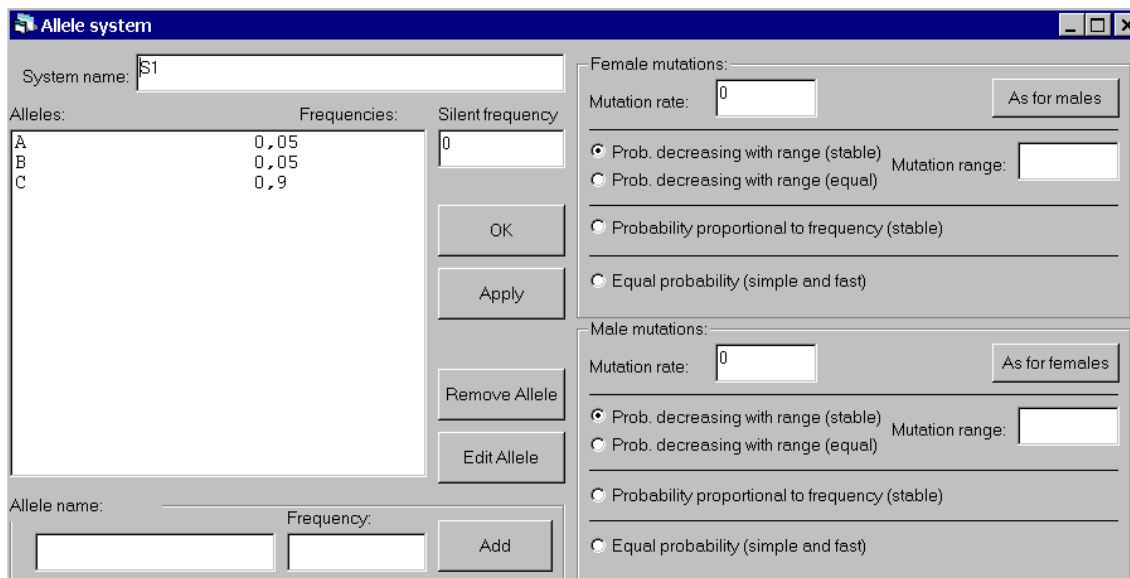
Figure 3.3: *The window for entering an allele system.*

Once you have entered the allele systems, it is a good idea to save your data. The saved file will then become a database for these systems, and can be used as a starting point for other cases using the same systems.
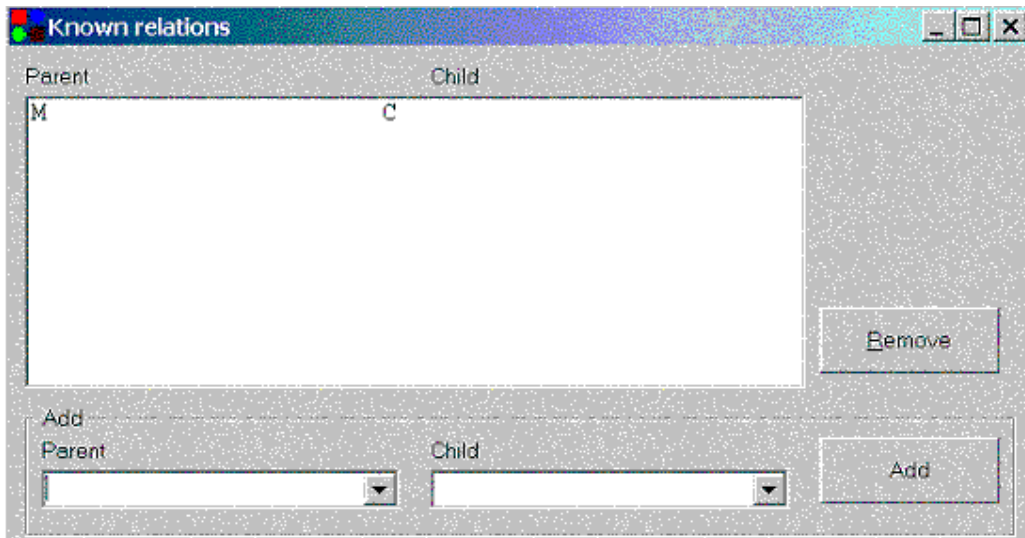
3.2   Persons

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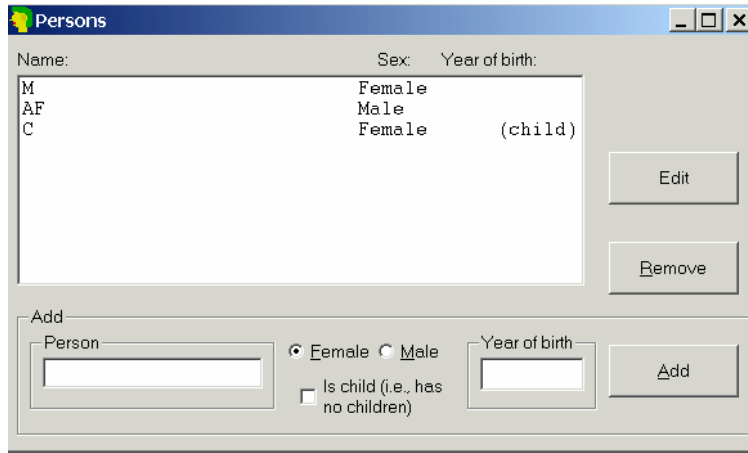
Figure 3.4: *The window for entering the persons involved in a case.*

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 it should 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 though their relations with the common parents. The window is illustrated in Figure 3.5.

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

This means specifying the alleles for the various systems for the persons involved. 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 present (see Section 3.1).

Figure 3.6: *The form for entering case-related DNA data.*
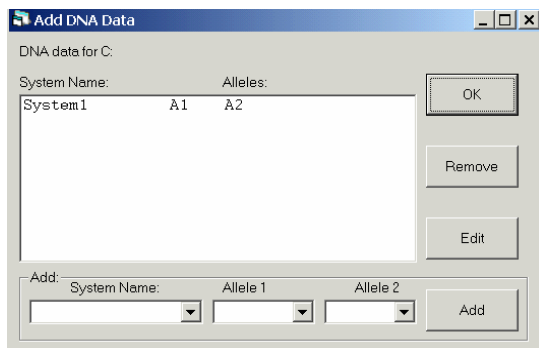
Figure 3.7: *Adding DNA data for a selected person is done in this window***.**

Read case data

Data for specific samples can now also be read in from files. The data for specific samples should be given as a table, also in a format that can be output from Excel, using tabs as separators. 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. 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. An example of an input file is given below.

```
name   amel1 amel2 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.*

Manual for **familias**

3.5   Pedigrees

In this form you may add your own pedigrees or you may generate pedigrees. 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]. After having generated the pedigrees, one can calculate probabilities and likelihoods for the various constellations. In the following we will go through the entire set of buttons and options of the window shown in Figure 3.8.
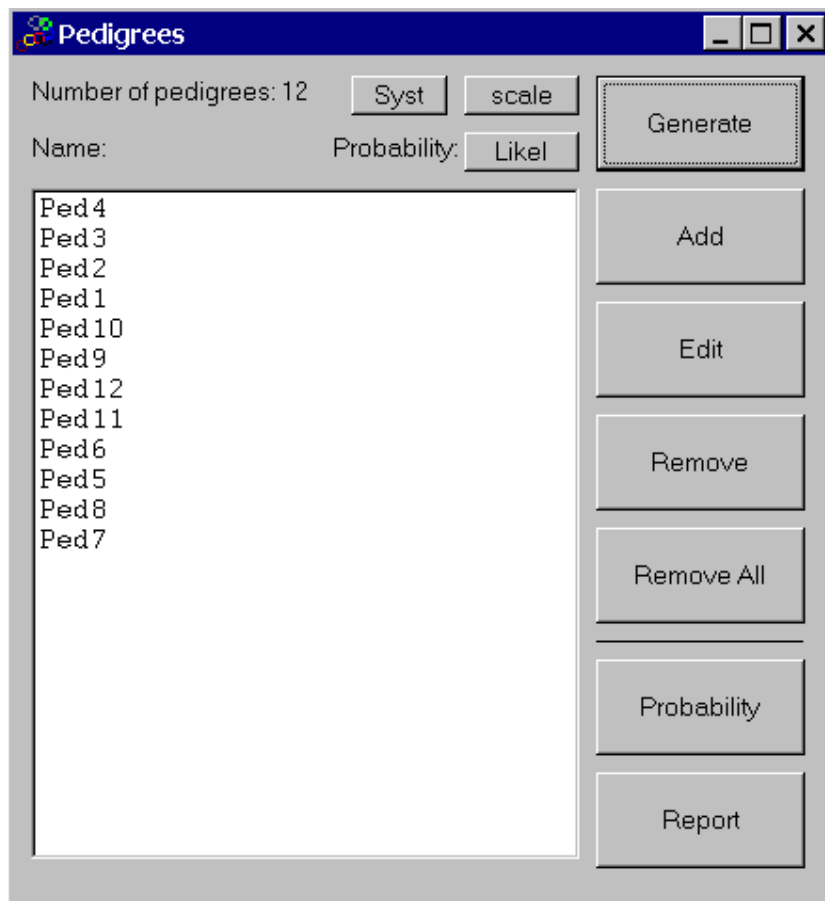

Figure 3.8: *The Pedigrees window with 12 generated pedigrees.*

3.5.1   Creating and editing pedigrees

Usually, the first thing to do here is to create a set of pedigrees, either automatically or manually. By pressing Generate, **familias** will generate all possible pedigrees made up from the defined persons plus a specified number of extra persons (see Figure 3.9). Remember that **familias** defines pedigrees only through parent-child relations; therefore it is often necessary to include additional persons when constructing pedigrees.

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 Figure 3.10 appears.
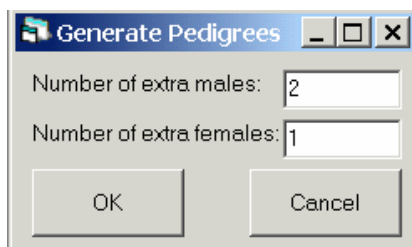
This is the same window that appears when pressing Add for manual construction of pedigrees. The pedigree is defined by the list of parent-child relations, and is thus altered by adding or removing these relations. You use the Persons-button to add the extra men and women that are necessary to define the wanted pedigree.

Figure 3.9. *Adding persons.*

As an alternative to adding anonymous extra persons here, the extra persons could have been defined in the Persons window described above (Section 3.2). 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 (see Section 3.5.2).

Besides adding new pedigrees to the list in the window of Figure 3.8 by means of Add, one uses the buttons Edit (or simply double-click), Remove and Remove All to manage the list of pedigrees. Be aware, though, that the posterior probabilities are dependent on the alternative pedigrees that are being considered, and must always be evaluated in that context.
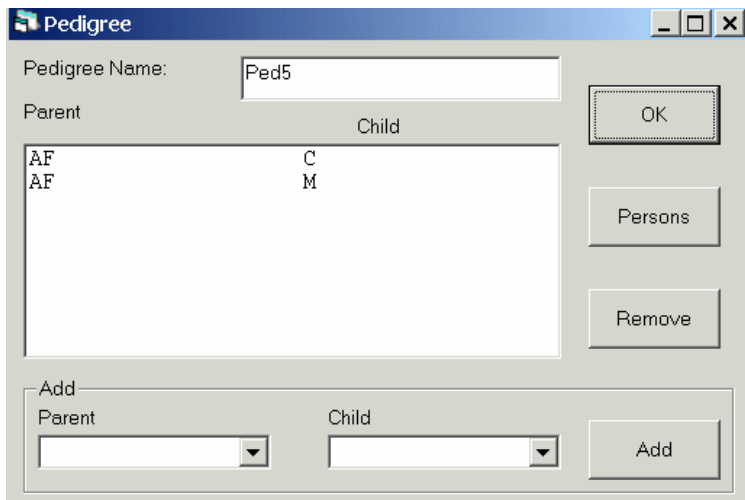
Figure 3.10: *The window for editing a given pedigree.*

3.5.2   Calculating probabilities

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
given in Section 2.1. 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.
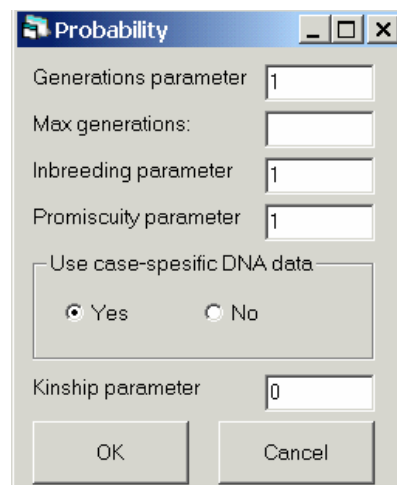
Figure 3.11. *Calculating probabilities*

The parameters relating to generations, inbreeding and promiscuity are explained quite
thoroughly in Section 2.1.

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. The biological foundation for the parameter is introduced in Section 2.5, along with suggestions for reasonable values.
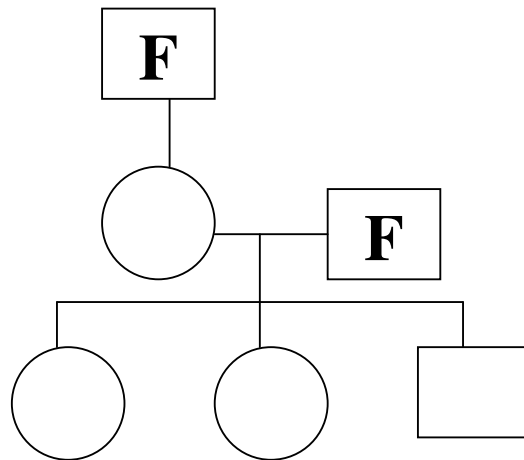
Figure 3.12: *Pedigree*

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.

3.5.3 Alternative measures (likelihoods and ratios)

The Pedigree window gives two options to the usual calculation of posterior probabilities, namely Likelihood and Scale. Pressing Likelihood makes the program display likelihoods instead of probabilities, that is, the likelihood of the data given the various hypothesized pedigrees, without incorporating the prior. See Section 2.3 for an introduction to the relevant terminology. Selecting a pedigree and pressing Scale causes the program to calculate ratios between the likelihood or probability measures of each pedigree and the selected one, thus creating likelihood ratios (LRs) or posterior probability ratios (PPRs), depending on the measure currently in use. Pressing either button a second time takes you back to the original measures.

Active loci

It is now possible to select the loci to be used for calculations. This is done by pressing the syst button of Figure 3.8. By default, all loci are used.

3.5.4 Reporting the results

By selecting the pedigrees of interest and pressing Report, a text file is generated, containing details on those pedigrees, that is, the persons involved and the relations between them, and also the probability of the particular pedigrees. The report can be saved in a file; the format of the file is rtf.

References

[1]   Balding, D.J. and Nichols, R.A. (1995). *A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity*. Genetica 96, pp. 3-12.

[2]   Egeland, T. et al (2000). *Beyond traditional paternity and identification cases: selecting the most probable pedigree*. Forensic Sci Int 110, pp. 47-59.

[3]   Egeland, T. and Mostad, P.F. (2000). *Statistical Genetics and Genetical Statistics: a Forensic Perspective*. Scand J Statist 29, pp. 297-307.

[4]   Egeland, T., Mostad, P.F. and Olaisen, B. (1996). *Computerised probability assessments of family relations*. Sci Justice 37, 269-275.

[5]   Evett, I.W. and Weir, B.S. (1998). *Interpreting DNA evidence*. Sinauer, Sutherland MA.

[6]   From Shaun Purcell's speech on "Genetic background and population stratification" at Wellcome Trust Centre for Human Genetics: http://www.well.ox.ac.uk/~valdar/biostag/purcell_1may2002.pdf

[7]   From the homepage of Ph.D. Charles Brenner: http://dna-view.com/patform.htm

[8]   Dawid AP, Mortera J, Pascali VL, Van Boxel D (2002) . Probabilistic expert systems for forensic inference from genetic markers. Scand J Statistics, 29:4, pp. 577-596.

A   Appendix

NB This chapter not updated for familias 1.6 ***********

A.1   Mutation models

There are three 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) Equal probability
2) Proportional
3) Decreasing

The equal probability and the decreasing model are non-stationary and the proportional is stationary. Being stationary means that the allele distribution will not change in time. An unpleasant consequence of using a non-stationary model is that the likelihood ratio will change by including extra irrelevant persons in the calculation. We shall take a closer look at the three models.

A.1.1   The Equal probability model

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 $j\ (i, j = 1,…,N)$. 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:

$$
M = \begin{bmatrix}
1-R & \dfrac{R}{N-1} & . & . & \dfrac{R}{N-1} \\
\dfrac{R}{N-1} & 1-R & . & . & . \\
. & . & . & . & . \\
. & . & . & . & . \\
\dfrac{R}{N-1} & . & . & . & 1-R
\end{bmatrix}
$$

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.

A.1.2   The Proportional model

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:

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

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.

A.1.3 The Decreasing model

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:

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

where R is the overall mutation rate, r is a constant between 0 and 1 (0<r<1), and k_i is chosen such that $\sum_{j=1}^{N} m_{ij} = 1$. A simple calculation gives $k_i = \frac{R(1-r)}{r(2-r^{i-1}-r^{N-i})}$.
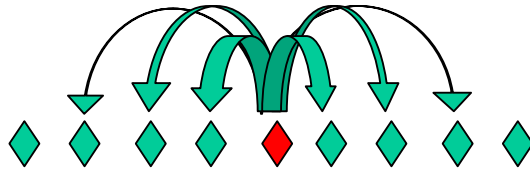
Figure A.1: *Mutation model*

A.1.4   An example illustrating the three different 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). 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	0.02	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

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

a) For the equal probability model 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 $m_{AD} = m_{BD} = kp_D$ and $m_{AC} = m_{BC} = kp_C$. 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 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

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

The three 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 [3]. Usually it will be a good idea to check the robustness of the conclusions by incorporating different mutation models.

A.2 Allele databases

Several allele databases have been constructed and are available to the public. As the allele frequencies vary between populations, one should always try to use the appropriate database in each case. For allele frequencies databases you may want to check out the following websites:

- http://alfred.med.yale.edu/alfred/index.asp.
- http://www.contexo.info/DNA%20Main%20Page.htm
- http://www.csfs.ca/databases/index.htm.