Exercises using Familias 3, Copenhagen 2014

The exercises are intended to provide an introduction to Familias 3. For introductory examples/exercises on the general use of Familias, see http://arken.umb.no/~theg/book/. Users are required to have basic knowledge about the dialogs/windows in. Exercises with one * is intended for intermediate users, while questions with ** are for advanced users.

Introduction, Using Familias 3

Familias 3 provides several new functionalities compared to previous versions. The users are encouraged to explore the software and provide feedback on the experience ot Daniel.Kling@fhi.no. Listed below are some of the new features,

- The user may, for each allele system, define the database size. That is, we may store in the software the number of observations.
    - When a new/unseen allele appear upon importing data for persons, we may immediately select to add the allele using the database size as reference.
- The user may change the mutation models and mutation parameters for all systems at once.
- There is a new mutation model, the 'Extended step-wise' model.
- The blind search interface may be used to scan a set of persons for pairwise relationships.
- The simulation interface allows the user to investigate the distribution of LR-s for any given case.
- The DVI (Distaster Victim Identification) module may be used to handle accidents where unidentified findings are to be compared to relatives/reference material of the missing persons.

Exercise 1, The new mutation model and its implications

The new mutation model extends the stepwise model to account for microvariants/intermediate alleles. To illustrate, consider the system S1 with alleles 9, 9.3 and 10. The old Familias would consider the mutation from 9 to 9.3 one step and similarly 9.3 to 10 one step, while the actual one step mutation 9 to 10 would be considered a two step mutation. The new model implements a new parameter, loosely defined as secondary mutation rate (Rate 2), that is the probability of observing a mutation which is non-integer, in the given example, from 9 to 9.3 or from 9.3 to 10. All non-integer transitions are considered equally probable while the integer transitions follow a pattern for stepwise mutations as described in the Familias manual. Consider a paternity case with an undisputed mother (trio). We have,

- H1: The alleged father (AF) is the true father of the child
- H2: Another man, not related to AF, is the true father of the child

Consider an allele system with 5 alleles, 9, 9.3, 10, 12 and 15. Assume uniform allele frequencies, i.e., all allele have frequency 0.2. We have parameters mutation rate, $\mu = 0.001$, mutation range, $r=0.1$ and secondary mutation rate, $\alpha=0.00001$ (the parameters are called respectively 'Rate', 'Range' and 'Rate 2' in the Familias interface).

a) **Compute the mutation matrix for the extended model, given the parameters above. (You may compare your results to the matrix calculated in the Familias core by dumping the matrix in the File -> Advanced dialog)

We next define genotypes for the mother as 9, 9, the child 9, 15 and the father as 12, 12.

b) Calculate the LR with new Familias using the non-stationary stepwise model.
c) Calculate the LR with new Familias using the extended stepwise model.
d) Discuss the difference between the results in b) and c)
e) Change the fathers genotypes to 10, 10 and compute the LR again using new Familias.
f) *Derive the theoretical formula with mutations and compare for different genotypic setups the theoretical value with the LR computed in Familias.
g) *Discuss stationary versus non-stationary mutation models.
    a. What may be the drawback with stationary models?
    b. What is the drawback with non-stationary models?

Exercise 2, Using the Blind search interface

The blind search interface is, as the name suggests, a tool to blindly search for predefined pairwise relationships in a data set. The interface currently allows the user to search for parent-child, sibling, half-sibling relations as well as direct matches. It may be used in connection with the DVI module, see Exercise S4, to search a set of unidentified remains for direct matches or relationship within the data set. Another application may be to investigate a data set for unspecified relations before conducting a medical study or prior to creating a frequency database. The computations are relatively swift and may be used to search large data set for relations. We will test the module on a smaller data set.

a) Create one allele system with four alleles, 12, 13, 14 and 15 and allele frequencies uniformly distributed as 0.25.
b) Define four males P1, P2, P3 and P4. Enter DNA data as P1=12/12; P2=12/12; P3=13/14; P4=14/15.
c) Enter the Blind search dialog and press New search. Select Direct match and Siblings as relationships and leave the remaining parameters at their default.
d) Interpret the results.
e) Confirm the LR values by manual calculations.

Exercise 3, The simulation interface and its applications

The simulation interface has many uses and we will explore one of them. The interface uses the frequency database in Familias to simulate the defined pedigrees and calculates likelihood ratio distribution for the given case. Founder alleles are sampled using allele frequencies. Non-founder alleles are subsequently sampled using the specified mutation model. We may define which persons are genotyped and those who are not.

Your lab has now received a request to whether you can reach a conclusion in a given case. We have,

- H1: The father of woman is also the alleged father of her child.
- H2: Another man, unrelated to the alleged father is the true father of the child.

The alleged father is dead and we may only receive samples from the mother and the child. In both cases the alleged father is the undisputed father of ‘mother’.

a) Open the file database_S3.fam, which contains allele frequencies for 23 autosomal markers. These are all markers you have access to.
b) Define the necessary persons. (Note we do not define any genotypes as we do not yet have any data)
c) Create the pedigrees according to the hypotheses above.
d) In the pedigree window enter the Simulation interface.
e) Select the persons which we will genotype.
f) Enter 1000 simulations, Specify seed to 1234 and Data for all markers. Press Simulate.
g) Try interpreting the results.
    a. What is the Median and the 5% and 95% values.
    b. What is the difference between the median and the mean?
h) Enter the Limit dialog. Select the threshold you (your lab) consider as a sufficient LR and press update. What is the false positive rate given your threshold? What is the chance we will be able to give a conclusion in the given case? Try using LR=100 as threshold, what is your decision given the results?
i) *Try performing a new simulation with only a subset of the markers, by selecting markers in Included systems dialog in the pedigree window and the deselect Data for all markers prior to simulating. Do we need all markers in the given case?

Exercise 4, A small accident – using the DVI module to identify the remains

In the last exercise we consider a small scale accident. Consider the crash of a small aeroplane with 10 passengers. We obtain reference data from 5 different families.

a) Open the database_4.fam file, which contains frequency data for some 23 autosomal markers.
b) Enter the first step in the DVI module, Add unidentified persons.
    a. We may define individuals manually, similar to normal Familias procedure, though we prefer importing data from file to skip as much manual input as possible. Import the file pmdata_4.txt.
    b. The file only contain 8 unidentified remains. Discuss why this may be a realistic scenario. How may this effect the calculations?
c) Deselect use list and enter 10 in the Size box.
    a. This is used to define the priors
    b. * Discuss the use of priors and how priors should be addressed in a DVI operation. (See also h)).
d) Press next to define reference families.
    a. We may now either define families manually or we may import them from file.
    b. Define the first family manually by selecting Add.
        i. Enter a name for the family, Family 1
        ii. Import data for the persons in the family (A father). Import file amdata_family1_S4.txt
        iii. Define other persons included in the family, in the current family none. Note, this may also be untyped persons necessary to define the relations between the reference persons and the missing person(s). We will return to an example of this later

iv. We continue be defining the relation between the defined person(s) and the missing person. (Note, simply naming the person father/mother/brother etc. does not define the relationships). Select Add in the pedigree window to add a new pedigree. Name the pedigree appropriately, e.g. Father and add necessary relation between the reference person(s) and the missing person.

c. Define also a second family manually by selecting Add.
   i. Enter name, Family 2
   ii. Import reference data from file amdata_family2_S4.txt
   iii. Add necessary persons and the define the reference person as brother to the missing person

d. Add the rest of the reference families by selecting the import option Simple and select files amdata_family3_S4.txt, amdata_family4_S4.txt and amdata_family5_S4.txt. Change the names of the families to Family 3, Family 4 and Family 5. Also check the persons and pedigrees in each imported family.

e) Press next to perform the search.
   a. Select the threshold for a match to be reported. Now enter 1.0, as we would rather obtain more matches at this stage and later remove matches which may be spurious.

f) Interpret the results.
   a. Were all remains identified?
   b. Select a match and press View match to investigate the individual LR:s for each system.

g) * We suspect there might be relatives among the unidentified persons. Enter the first step, Add unidentified persons and select Blind search. Use your knowledge from exercise S2 to perform a blind search for parent-child relations. How may the results be used in the DVI operation? (Use 1.0 as match limit, leave all other options at default)

h) * Change the size of the accident in step c) to 100 and try explain how this effect the priors in the current case. How does this effect the posteriors?

i) ** New information is added to the case. The first family, defined manually in d) a. also contains a second missing person. The brother of the reference father is also missing. Try finding out how this could be solved using the means available in the DVI module.
   a. Perform a new search (Use the same threshold/limit as in e))
   b. Discuss the solution and other ways to improve the algorithm

Now that we are familiar with the new features of Familias we are ready to accept more challenging tasks.

*Exercise 5, Further use of the simulation interface

Use Familias' simulation interface to find how many persons we should genotype in order to get sufficient results in the current case:

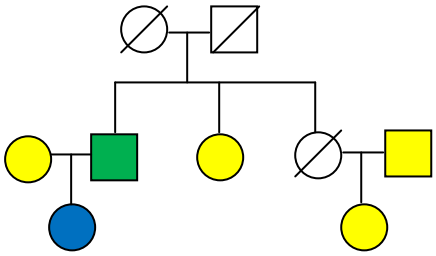

Figure 1. Pedigree where the individual marked with blue indicates the child interested to know whether the individual marked with green is her father. Individuals marked with yellow are available for genotyping.

Try including different sets of the available persons and find out which ones are necessary to obtain sufficient LR values. Find a LR threshold where the false positive rate is low enough while we still have a high probability of reaching a conclusion. (Use you own database or use database_S5.fam)

**Exercise 6, Simulation with multiple hypotheses

We consider a case with three hypotheses and are interested in finding the distribution of LR:s. Consider hypotheses,

- H1: Two persons are full siblings
- H2: Two persons are half siblings
- H3: Two persons are unrelated

Use the same database as in exercise S5 (or your own database) and try interpreting the results.

Exercise 7, A case with low quality profiles (dropouts)

We consider a paternity case with a mother and her daughter and an alleged father,

- H1: The alleged father is the true father of child
- H2: Some other man, unrelated to the alleged father is the true father of the child

a) Use the same frequency data as in exercise 2. Enter a dropout probability of 0.1 for marker L1. (Option dialog when editing allele system)
b) Add the alleged father, the mother and the child.
c) Enter DNA data as alleged father=15/15; mother=12/12; child=12/13
d) Select that you wish to model dropouts for the father. (Consider dropouts when editing allele data for the person)
e) Setup the pedigrees according to H1 and H2
f) Calculate the LR
g) Try changing the dropout probability and see how this affects the LR. Do you see a pattern? (*Hint, use the Step dropout function in the advanced settings)
h) **Now let us consider something more advanced. Go to Advanced setting and select Logistic dropouts (Use) and leave the regression parameters at their default values.

    a. How can we estimate the dropout parameters? (How do we setup the experiment?)
    b. Edit the profile of the alleged father and change the peak height for 15 allele to 100.
    c. Calculate the LR and see how this changes if you change the peak height.

i) What is the difference between silent alleles and dropouts? Could a silent allele explain the data, try in Familias.
    a. Set the dropout probability to 0.0
    b. Add a silent allele in the Options dialog. For simplicity use the same frequency as for the dropout probability, 0.1. Scale the allele frequencies
    c. Calculate the LR again. What do you see?

Exercise 8, Searching a database for relatives (Familial searching)

This exercise is meant to provide a brief overview of the new upcoming database searching feature in Familias 3.1. (The feature is still largely untested, but should work for the purpose of this exercise) We will consider a fictive database with >50.000 individuals, see Database8.txt. All individuals are genotyped for one STR marker S1. Again, use the same frequency data as in Exercise 2.

a) Open the Familial searching feature under Tools -> Database searching (Alpha)
b) In the first window we import the database. (Usually containing traces from crime scenes and convicted offenders). In this case it contains only individuals, no mixtures, although the module can handle mixtures. You will receive a message that the file do not contain any gender information, press OK/Yes.
c) Press Next.
d) The next task is to import or define the profiles we wish to search for. (Usually traces from crime scenes, which can be single person or mixture profiles). We may search several profiles at once.
e) Press import and select to display All files. Import mixture.xml and press yes when Familias asks if the file is in CODIS xml format.
f) The files contain a trace from a crime scene with three alleles.
g) Next define the search criteria in the right side of the window.
h) Select 1.0 as LR threshold. Select Direct match and Siblings to search for in the Relationships window. Leave the other settings at the default values.
i) Press Next.
j) Press Search to perform a database search. This may take some time as we have >50000 individuals in the database and we have a mixture to search against.
k) View the matches.
l) Go back and remove the mixture profile. Import the file named no_mixture.xml
m) Use the same parameter settings and perform a new search. Can you see the difference in the weight of evidence when we have a mixture compared to when we have a simple single individual evidence.
    a. Try to calculate the matches by hand (Select view match to see the allele setups)

Exercise 9, Using the create database feature

The current exercise will guide you through the steps of creating a frequency database from scratch using the new functionality of Familias 3. (As with the Familias searching feature, this function is also not completely tested, be aware of bugs) Also how to remove duplicates and related individuals. The feature is found in the File menu -> Create database. The basis for this exercise is that we have some raw output from the genotyping of 25 individuals for the creation of a new population frequency database. (In reality, this is of course a to small number of samples)

a) Enter the create database feature as described above
b) Import the samples from the file Genotypes9.txt, by pressing the Import button
c) View the samples and the systems that have been created
d) Press Check data to enter the blind search interface
e) Press New search to define a new search
f) Select 10 as the Match limit and specify that you wish to search for Direct matches, Parent-childs and Siblings. Further, select 2^{nd} cousins as the alternative hypothesis.
    a. What does this mean?
    b. Why is this sometimes necessary?
g) Press search and view the matches.
    a. Are there any duplicates?
    b. Are there any related individuals that should be removed?
h) Close the blind search interface
i) Now press Statistics to get an output of some forensic parameters together with allele frequencies. Select file name and view the output.
j) Last, press Create to create the database. View the resulting database and allele frequencies