Relationship Inference with Familias and R. Statistical Methods in Forensic Genetics
X Exercises 4.12-4.22

Thore Egeland
Daniel Kling
Petter Mostad

August 24, 2016

4.5.2 X-chromosomal markers and FamLinkX

FamLinkX implements an algorithm for linked markers on the X-chromosome. In addition to linkage the software accounts for linkage disequilibrium (allelic association) and mutations. The software is intended to be user-friendly but may provide obstacles for the inexperienced user. FamLinkX provides the likelihood ratios using three different methods, M1: Exact model, considering linkage, linkage disequilibrium and mutations; M2: Cluster approach, see manual for Merlin, linkage and linkage disequilibrium is considered but not recombinations within clusters and not mutations; M3: Only linkage is considered between markers. In the following exercises we are interested in M1

	12	13
16	59	1
17	1	39

Table 4.15: Haplotype observations for Exercise 4.12 .

as this is the preferred model, specially for STR markers, but comparisons to the other models will be made. For all calculations, unless stated otherwise, we consider X-chromosomal marker data and corresponding inheritance patterns.

Exercise 4.12 (A paternity case revisited. Video)**.**

We will first revisit the paternity case (Duo) (for illustration see Figure 2.6) with hypotheses

H_1**:** The alleged father (AF) is the true father of the female child.

H_2**:** The alleged father and the child are unrelated.

a) Open `FamLinkX` and specify the frequency database. *Hint*: `File->Frequency database`. Create a new cluster and specify two diallelic systems, L1 and L2, with alleles 12, 13 and 16, 17 respectively. Let $p_{12} = 0.6$, $p_{13} = 0.4$ for L1 and $p_{16} = 0.6$, $p_{17} = 0.4$ for L2. Select the `Simple mutation model` with the mutation rate set to 0 for both systems. Set the genetic position to 10 cM for L1 and 10.1 cM for L2. Furthermore, specify haplotype observations according to Table 4.15. Why is it important that we explicitly specify the gender of all persons in calculations for X-chromosomal marker data?

b) What is the estimated recombination rate between the two loci? Use Haldane's mapping function.

c) Why do we specify the number of observations for each haplotype?

d) Use the equation below to calculate a measure of the association between the alleles

$$r^2 = \frac{(p_{12}p_{16} - p_{12,16})^2}{p_{12}p_{13}p_{16}p_{17}}. \tag{4.12}$$

e) Specify λ to 0.0001 in `Options`. We will discuss the importance of λ in Exercise 4.14 and will not dwell further on it now. In brief, setting a low λ gives large weight to the observed haplotypes in Table 4.15. Select the appropriate pedigrees using the Wizard. The alternative hypothesis depicts two unrelated girls, but the first hypothesis will override the genders. *Hint*: `File->New wizard` and specify the data for the father as 12 for L1 and 17 for L2 and the child as 12/12 for L1 and 17/17 for L2. Calculate the LR which should coincide (approximately) with the theoretical value of 100. Choose to save the file when asked. There may be a small deviation from the theoretical value, which we will return to in later exercises.

f) Change the genotypes for the child to 12/13 for L1 and 16/17 for L2. Calculate the LR.

g) Change the number of observations for each haplotype. What happens? Explain why!

h) * Discuss the high degree of LD and if this is a likely situation to occur in reality.

Exercise 4.13 (A case of sibship revisited)**.**

In the second exercise we revisit the example in Exercise 4.2 concerning disputed sibship. Two females, F1 and F2, are interested to find out whether they are siblings in some way. We specify hypotheses

H_1**:** F1 and F2 are full siblings

H_2**:** F1 and F2 are maternal half siblings

H_3**:** F1 and F2 are paternal half siblings

H_4**:** F1 and F2 are unrelated

a) Explain why we may distinguish H_2 from H_3 with X-chromosomal markers but not with autosomal markers.

b) Use the same frequency data, and haplotype observations, as in Exercise 4.12, alternatively open the file **Exercise4_13.sav**.

c) Specify $\lambda = 0.0001$ and select the relevant hypotheses. Note: if you also stored the case-related DNA data in the previous exercise, you may be asked if you wish to erase all DNA data, answer yes.

d) Enter data for both F1 and F2 as 12/12 for L1 and 17/17 for L2. Calculate the LR Scale against H_4. Comment on the importance of accounting for LD and linkage in the current case.

Exercise 4.14 (On the importance of λ)**.**

This exercise is intended to provide some insight into how the haplotype frequencies are estimated and the importance of the parameter λ. Our model for haplotype frequency estimation is described by

$$F_i = \frac{c_i + p_i\lambda}{C + \lambda} \tag{4.13}$$

where F_i is the haplotype frequency for haplotype i, c_i is the number of observations for the haplotype, p_i is the expected haplotype frequency (assuming linkage equilibrium) calculated using the unconditional allele frequencies, C is the total number of observations for all haplotypes and λ is a parameter giving weight to the expected haplotype frequencies. This model allows for unobserved haplotypes to be accounted for, in contrast to models which estimate the haplotype frequency solely based on the counts. The difficulty lies in the choice of a good λ. Our recommendation is to compute the LR for a number of different values and select the least extreme value, i.e., the value closest to 1.

To start, we specify a case with an aunt of a female child

H_1 : The female is the aunt of the child.

H_2 : The two females are unrelated.

a) Again use the same frequency data as in Exercise 4.12. Select the relevant hypotheses. *Hint*: Select the *Aunt/Uncle* as the main hypothesis.

b) Enter data for the child as 12/12 for L1 and 16/17 for L2 and for the aunt as 12/13 for L1 and 17/17 for L2.

c) Calculate the LR for d $\lambda = 0.0001$, 0.01, 1, 100 and 10000.

d) What happens for large and small values of λ? *Hint*: use the equation for haplotype frequency estimation above.

e) Change the data for the child to 13/13 for L1. Repeat c) and discuss the results.

Exercise 4.15 (Extended example. Combining `Familias`, `FamLinkX`).

This exercise provides a challenge where the user needs to combine the results from `Familias` and `FamLinkX` to obtain a final result. The data is extracted from a real case (anonymized) where three females provided DNA samples. The hypotheses are

H_1 : The three females are all full siblings.

H_2 : Any other pedigree constellations.

Obviously H_2 cannot be used in the current setting, in a simple way, and we need to refine possibly alternative hypotheses. a), b) and c) involve the use of `Familias`, however you may also skip to d) for `FamLinkX`.

a) * Open the file **Exercise4_15.fam** in `Familias 3`. We may assume that all females are children with no children of their own. Specify that the three females are children. Also define two parents, a mother and a father and specify that they are both born 1970.

b) * Use `Familias` (Generate pedigrees) to find the pedigrees with a posterior probability above 0.001. Which are the most probable relationships according to the results? Interpret the results.

c) * Discuss the constraints specified in a) and their impact on the results in b)

d) Open the database **Exercise4_15.sav** in `FamLinkX`, which contains frequency and haplotype data for the Argus X12 kit from QIAGEN based on a Swedish population sample. Explore the haplotype frequency database.

e) Based on the results in b), we specify the hypotheses

H_1 : The three females are all full siblings

H_2 : Two females are full siblings and the third (named F3) is a paternal half sibling

H_3 : Two females are full siblings and the third (named F3) is a maternal half sibling

f) Import the DNA data, available in **Exerecise4_15.txt**. Make sure to import the data in the file to the correct corresponding persons, the person denoted F3 should be imported to 3.

g) Calculate the LR and interpret the results. Be patient, the computation may require some time >20 min. *Hint*: To speed the computations up, go into `File->Advanced`, select and edit the hypothesis we have selected to investigate. For each pedigree set the `Threshold` value to 0.001 and the `Steps` value to 0.

h) Discuss if the LR in g) may be combined with the results in b)? What is your final conclusion on the case?

Exercise 4.16 (Further discussion of λ).

We will provide an example of how the value of λ may crucial to the conclusion in a case. We use anonymized data from a real case with two typed females (F1 and F2) and consider hypotheses

H_1 : F1 and F2 are paternal half siblings, with different mothers.

H_2 : The two females are unrelated.

a) Open the file **Exercise4_16.sav**, containing the frequency database.

b) Select appropriate pedigrees and import genotype data from the file **Exercise4_16.txt**.

c) Compute the LR for a number of different values on λ, e.g., 0.001, 1, 100 and 1000.

d) Discuss the results in c). What conclusion can be drawn?

e) * Explore the genotype data and see if you can find an answer to the results. Use your knowledge about haplotype phases under the different hypotheses. *Hint*: Use the frequency estimation tool in the `Edit cluster` dialog.

f) Discuss what value on λ should be chosen. What is most conservative?

Exercise 4.17 (ESWG 2013 paper challenge).

This exercise covers the calculation of the LR for the X-chromosomal data included in the ESWG paper challenge 2013 [50]. The question involved a paternity duo with the following hypotheses,

H_1 : AF is the biological father of a female child.

H_2 : AF and the child are unrelated.

a) Open the file **Exercise4 1̲7.sav** containing the frequency database. Explore the specification of the haplotypes. Make sure λ is set to 1.

b) Select appropriate pedigrees and import genotype data from the file **Exercise4 1̲7.txt**.

c) Calculate the LR.

d) Change the value of λ to 212 and see how this affects the results.

e) Compare the results calculated with the equilibrium (LE) model with the Exact .

f) ** Try deriving the algebraic formula for the two markers in the first cluster. For simplicity assume mutations can be disregarded.

Exercise 4.18 (* Creating pedigrees)**.**

`FamLinkX` includes some predefined pedigrees where the calculations are performed using methods M2 and M3. In addition, you may wish to create new pedigrees. The implementation currently does not allow the exact method to be used on user-defined pedigrees, therefore only calculation with the other methods will be done. Two females F1 and F2 are related as maternal cousins. In addition, we are asked to find out whether they also share the same father,

H_1 : F1 and F2 are maternal cousins as well as paternal half siblings.

H_2 : F1 and F2 are maternal cousins with different fathers.

a) Open the file **Exercise4_18.sav** containing the frequency database. We will consider different λ, start by setting the parameter to 1.

b) Continue by creating the pedigree for H_1. You need do define several extra persons to specify the necessary relations.
Hint: `Tools->Select pedigree->Create/Edit pedigree`.

1. Rename the pedigree to H1
2. Add extra persons *Hint*: `Persons` button within the edit pedigree dialog. Add persons named F1, F2, grandmother, grandfather, mother1, mother2 and father. Make sure the genders are correct for all individuals.
3. Specify the relations in H1.
4. Close the edit pedigree dialog. The created pedigree will appear last in the list of pedigrees. Select the created pedigree and `Next`.
5. Create the alternative hypothesis, H_2, using the same procedure as for H1. Note: you need not define any new persons, use the same as create for H1.

c) Import genotype observations from the file **Exercise4_18.txt**. Make sure to import the person named *cousin1* to F1 and *cousin2* to F2.

d) Calculate the LR.

e) Repeat the calculations for different values of λ. What is your conclusion regarding the relationship between the two women.

Exercise 4.19 (* Theoretical considerations)**.**

For verification purposes, and to validate calculations/implementations, it may be interesting to derive theoretical formulas. Generally this is infeasible, but for some cases, using simplifications, algebraic formulas can be derived. We will start by consider one diallelic SNP marker, S1, with alleles 1 and 2. Frequencies are $p_1 = 0.4$ and $p_2 = 0.6$ and mutation rates are zero. Furthermore, we specify,

$H_1:$ A women (AM), is the alleged mother of the girl.

$H_2:$ The alleged mother and the girl are unrelated.

a) The AM has genotype 1/1 while the child has genotype 1/2. Compute the LR by hand and verify in `FamLinkX`.

The AM is discarded as the true mother for reasons other than the DNA. Instead an alleged father is presented. The hypotheses become,

$H_1:$ A man (AF), is the alleged father of a female child.

	1	2
3	1	59
4	39	1

Table 4.16: Haplotype observations for Exercise 4.19.

H_2 : The alleged father and the child are unrelated.

b) The AF has genotype 1 while the child still has genotype 1/2. Compute the LR by hand and verify in `FamLinkX` (Note: the alternative hypothesis displays two unrelated females, while as pointed out earlier, the genders in the main hypothesis overrides this information).

c) A second SNP, S2, is introduced with alleles 3 and 4. Frequencies are $p_3 = 0.6$ and $p_4 = 0.4$. The AF has genotype 3 while the child has genotype 3/3. Compute the LR by hand and verify in `FamLinkX`. You may assume that the two SNP:s are located far from each other and thus the calculations independent. This is in practice impossible in FamLinkX, since we only have one chromosome, a sufficient approximation is to define the genetic position well apart, e.g. 500 cM.

d) S1 and S2 are in fact closely located, separated by only 0.1 cM. Specify this in `FamLinkX`. Calculate the LR again.

e) There are further haplotype observations according to Table 4.16. Specify this in `FamLinkX`.

f) Use three different values for λ, 0.01, 1, 100, and compute the LR in `FamLinkX` for each.

g) *Derive the theoretical formula to confirm the calculations in g) for λ=100.

h) ** Plot the LR as a function of λ. *Hint*: Use the formula in Exercise 4.14)

i) ** Try deriving the theoretical formula for the paternity case in Exercise 4.12 using the specifications in e) and f) in the exercise.

Exercise 4.20 (** Exploring the algorithm).

The exercise will explore the algorithm implemented in `FamLinkX`. The paper by Kling et al [67] provides further details that may be needed to solve some of the questions. We first introduce one genetic marker with alleles 12,13,14,15 and 16. Consider hypotheses

$H_1:$ Two girls, F1 and F2, are maternal half siblings.

$H_2:$ F1 and F2 are unrelated

a) * Given H_1, specify the founder alleles patterns, i.e., possible set of alleles for the founders, if the genotypes are 12/13 for F1 and 12/15 for F2. Assume mutation rates are zero. *Hint*: Plot the pedigree on a paper and specify the founders.

b) * Given H_1, specify the founder alleles patterns if the genotypes are 12/13 for F1 and 14/15 for F2. Now assume one-step mutations are possible.

c) * Set up the inheritance patterns, i.e. which are the meioses required to be accounted for?

d) * Using the results in a) and c), compute the number of different combinations we have to consider. In other words, the number of founder allele patterns times the number of inheritance patterns.

e) ** Now consider a second marker with the same data, i.e., the results in d) apply also for this marker. The two markers are linked such that for each combination for the second marker we also have to consider the different inheritance patterns in c). Using this information, compute the updated number of combination we have to consider for the second marker.

f) ** We also need to account for linkage disequilibrium using founder allele patterns, i.e., for each combination at the second marker we have to consider all possible states at the first marker. Using this information, compute the updated number of combination we have to consider for the second marker.

g) ** Finally, we consider a third marker, tightly linked with the second marker. Given that we have to consider linkage disequilibrium across all three markers what are the number of combinations we have to consider for the third marker? *Hint*: use that linkage only stretches one marker, i.e., given a specific inheritance state for the second marker, the third marker is independent of the different marker states at the first marker)

h) ** Consider instead H_2 and compute the number of combinations for the third marker. What effect does linkage have on the number of combinations we have to consider?

Exercise 4.21 (A case with mutation)**.**

We will consider a case where we introduce a possible mutation. Consider two females, F1 and F2, with hypotheses about relationship given as

H_1 : Maternal half siblings.

H_2 : Paternal half siblings.

which is not distinguishable using autosomal markers.

a) Import frequency data from the file **Exercise4_21.sav**.

b) Select maternal half siblings as the main hypothesis and select paternal half siblings as the alternative.

c) Import genotype data from the file **Exercise4_21.txt**

d) Compute the LR. Use the default $\lambda = 1$.

e) Change the data for the first individual (F1) to 19/25.1 for the marker DXS10148.

f) Compute the LR. Use default values for mutation parameters, i.e the simple model where each mutation is equally likely.

g) Change the data for the first individual (F1) to 19/19 and compute the LR.

h) Change the mutation model to `Extended model` with the `Range` parameter equal to 0.1 and the `Rate 2` parameter equal to 0.000001.

i) Compute the LR for the data as indicated in e) and g).

j) * Discuss the results and the importance of good models for mutations. Why is this decisive in the current case?

Exercise 4.22 (* Understanding complex parameters)**.** We will now have a closer look at some parameters that may further affect the results in `FamLinkX`, the *Threshold* and *Step*. The former, here denoted t, is between 0 and 1 and specifies the minimum value required for the overall pedigree likelihood to be included in the calculations. In other words, this limit apply to the $\Pr(D_i \mid V_i, F_i)$ in Equation 4.9.

The latter parameter, here denoted s, is a non-negative integer, and specifies the number of steps we consider for founder alleles. For instance, in a case of paternal half sisters, we need to sum over the possible alleles for the common father, given by the shared alleles of the sisters. If we consider mutations, all alleles are possible, while we can use the *Step* parameter do define the number of steps away from the sisters' alleles that we wish to include in the summation. In reference to Equation 4.9, this limits the founder alleles space F_i.

We will use an example to illustrate the effect of tuning the different parameters, consider

H_1 : Two females, F1 and F2, are paternal half sisters.

H_2 : The two females are unrelated.

a) Import frequency database from the file **Exercise4_22.sav**. The file contains a database for a Somali population [66]. The extended step wise mutation model is specified for all markers.

b) Specify the hypotheses according to H_1 and H_2. Import case data from the file **Exercise4_22.txt**.

c) Calculate the LR. Make sure λ is specified to 1.

d) Explore the genotype data to find an explanation for the results.

e) In the Advanced settings (`File->Advanced`), change the *Threshold* parameter for H_1 to 0.00001 leaving ***** other parameters unchanged? ***** the *Step* parameter unchanged for both H_1 and H_2. Save.

f) Compute the LR

g) Change the *Threshold* parameter for H_1 to 0.0001. Compute the LR.

h) * What can be said about the effect of the threshold parameter?

i) Change the *Threshold* parameter to 0.0001 and the *Step* parameter to 1 for H_1. Compute the LR.

j) Change the *Step* parameter to 2 for H_1 and compute the LR.

k) Change the genotypes for F1 for the locus DXS10101

l) * What can be said about the effect of the *Step* parameter?

We recommend keeping the value of *Threshold* high (e.g. 0.01-0.0001) and the value of *Step* low (e.g. 0 or 1) as this reduces the computation effort considerably compared to when we have lower values on the *Threshold* parameter and higher values on the *Step* parameter. If the LR is still zero for all computation models, a lower value on the Threshold parameter may be considered.