Netherlands Forensic Institute
Ministry of Justice

Introduction to kinship analysis

Seoul, ISFG workshop

29 August 2017

Short bio

- Studied mathematics in Leiden, The Netherlands
- PhD in mathematics at University of Amsterdam (2003)
- Work at NFI since 2008 as statistician and DNA kinship expert
- Collaboration with VU University Dept of Mathematics since 2010
- 1 day/week professor at VU since 2015
- Research interests: statistical/mathematical perspective on forensic kinship analysis and/or mixtures, properties of weight of evidence calculations in general

- Live in Leiden, The Netherlands
- www.cobblestonestories.org: Leiden studied by an American anthropologist

Contents of this block

Basic principles of kinship analysis

- Start with pairwise comparisons, without complications
- IBD concept
- Mutations
- Theta correction
- Linkage equilibrium
- Linkage
- Algorithms
- Y-STR LR calculations

Pairwise comparisons

First we focus on relatedness between two individuals A and B, on one locus.

H_p : A and B are related by pedigree P

H_d : A and B are unrelated

Bayesian framework: we must calculate the probability of the genotypes of A and B if they are related by pedigree P, and also if they are unrelated.

If they are unrelated (what is that?), then we obtain/model the likelihood of both genotypes as the product of the likelihoods of each genotype:

$P(A=g_1, B=g_2)=P(A=g_1)P(B=g_2)$

Identical by descent (IBD)

We say that allele a from person A and allele b from person B are IBD, if the alleles are inherited copies of each other within the pedigree.

E.g. Between parent and child, one allele of the child is IBD with one allele of the parent, since the child has inherited one allele from the parent.

Given a pedigree connecting two individuals A and B, it is possible to calculate the IBD probabilities

κ_0 : no IBD alleles between A and B (genetically unrelated)

κ_1 : one IBD allele between A and B (genetically parent-child)

κ_2 : two IBD alleles between persons A and B (genetically identical)

Relationship	κ_0	κ_1	κ_2
Unrelated	1	0	0
Parent-child	0	1	0
Monozygotic	0	0	1
Sibling	0.25	0.5	0.25
Half-siblings	0.5	0.5	0
Grandparent-grandchild	0.5	0.5	0
Uncle-nephew	0.5	0.5	0
First cousins	0.75	0.25	0
Double first cousins	9/16	6/16	1/16

Resulting LRs for pairwise kinship

- H_p states relatedness according to pedigree P
- From pedigree we get the IBD coefficients κ_0 , κ_1 , κ_2 (cf. exerc.)
- $P(B=g_2 \mid$ connected by pedigree P, $A=g_1)=$

  $\kappa_0 * P(B=g_2 \mid B$ unrelated to $A)+$

  $\kappa_1 * P(B=g_2 \mid B$ parent-child of A, $A=g_1)$

  $\kappa_2 * P(B=g_2 \mid B=A, A=g_1)$

- Kinship Index: $KI_\kappa = \kappa_0 * UN + \kappa_1 * PI(A,B) + \kappa_2 * ID(A,B)$,

where UN=1; PI=parent-child likelihood ratio; ID=LR for being genetically same person (i.e. zero if different or 1/match probability if equal genotype).

Therefore all pairwise kinship calculations reduce to the computations of the IBD coefficients, the parent-child case and match probability.

Example: obtain SI (sibling index) and HSI (half-sibling index)

G1	G2	UN	PI	ID	SI=UN/4+PI/2+ID/4	HSI=UN/2+PI/2
(aa)	(aa)	1	$1/p_a$	$1/p_a^2$	$¼+1/(2p_a)+1/(4p_a^2)$	$½ +1/(2p_a)$
(aa)	(aX)	1	$1/(2p_a)$	0	$¼+1/(4p_a)$	$½+1/(4p_a)$
(aa)	(XX)	1	0	0	¼	½
(ab)	(ab)	1	$(p_a+p_b)/(4p_ap_b)$	$1/(2p_ap_b)$	$¼+(p_a+p_b)/(8p_ap_b)$ $+1/(8p_ap_b)$	$½ +$ $(p_a+p_b)/(8p_ap_b)$
(ab)	(aX)	1	$1/(4p_a)$	0	$¼+1/(8p_a)$	$½ +1/(8p_a)$
(ab)	(XX)	1	0	0	¼	½

IBD versus IBS

- IBD alleles are inherited copies of the same ancestral allele, and therefore (ignoring mutation) are by definition identical
- IBS (identical by state) alleles are simply alleles that are indistinguishable

- IBD is relative to a pedigree, IBS is relative to a technology:
  - Alleles that are IBS when STR lengths are measured (as by CE) need not have the same internal structure and may be no longer IBS when sequenced
  - Alleles that are IBD are always identical since (ignoring mutations) they are copies of each other

Several loci

If the loci are independent and inherit independently (often the case for forensic loci):

- can multiply the LR over all loci
- but cannot distinguish between relationships with the same IBD coefficients (e.g. uncle-nephew and half-brothers)

If the loci are linked (close to each other on chromosome):

- then inheritance is not independent
- LRs cannot be multiplied: the LR on two linked loci is not the product of the LR's per locus
- But can distinguish between relationships with the same IBD coefficients, e.g. between uncle-nephew and half-brother; however power is in practice low unless many loci are typed.

Short overview of standard complications

- Allele frequencies
- Mutations
- Theta correction
- Lineage marker data

Allele frequencies

- Usually estimated from a case-independent reference database
- If that database is from another population than the individuals in the case at hand, tendency to overestimate the LR
- Reason: persons tend to have alleles that are common in their own population. If these alleles are rarer in the reference database, then the LR is higher when calculated from the reference database then in their own population
- I.e., from the point of view of the reference database, people in another population look more related than they actually are
- Possible ad hoc fix: theta correction
- New alleles: usually added with some minimum frequency; there isn’t a best/standard method.

Mutations

- A parent may transmit his/her allele *a* as allele *b*
- On usual forensic STR's, higher probability for father-child (order of 1/1000 per locus) than for mother-child (order of 1/10000 per locus)
- Probability depends on the distance between *a* and *b*
- Locus-specific: higher for more polymorphic loci
- Most mutations yield *b=a+1* or *b=a-1,* but longer differences are not impossible
- Ignoring the possibility of mutations can lead to wrongly excluding a parent-child relationship, and also other relationships (e.g. look for a third sibling given the first two)

Observing mutations

- Measuring mutation rates, especially allele specific, is not so easy: they are rare events
- Not all mutations are observable: e.g. if a father 16/17 transmits allele 16 as allele 17, this will be unnoticed
- The length of the mutation can not be observed, e.g. if a father 16/17 transmits allele 18, this could be 16 ->18 or 17 -> 18.
- It is in practice impossible to estimate for each locus the whole mutation matrix $(M_{a,b})_{a,b}$ from data

- Therefore some kind of model is needed

- The mutation rate $\mu = \sum_a p_a \sum_{b \neq a} M_{a,b}$ is the probability that a random allele will be transmitted as another allele and can be reasonably measured.

Mutation models: stationarity

- Suppose the allele frequencies in the population are p_a
- Now consider a child of an unknown father issued from that population
- The child will inherit allele a with probability $q_a = \sum_b p_b M_{b,a}$: the father has allele *b* with probability p_b and then this allele needs to be transmitted as allele *a*
- In general $q_a \neq p_a$
- If $q_a = p_a$ we say the model is *stationary*

Effect of stationarity

If a model is not stationary, then the LR for

- *Hp: C is child of AF and a mother (unrelated to AF) that has not been genotyped*
- *Hd: C and AF are random unrelated members of the population*

is different from the LR for

- *Hp: C is child of AF and a mother that has not been genotyped*
- *Hd': C is the child of parents (unrelated to AF) that have not been genotyped.*

because for *Hd*, the genotype probabilities for *C* are given by the p_a and for *Hd'* by the q_a

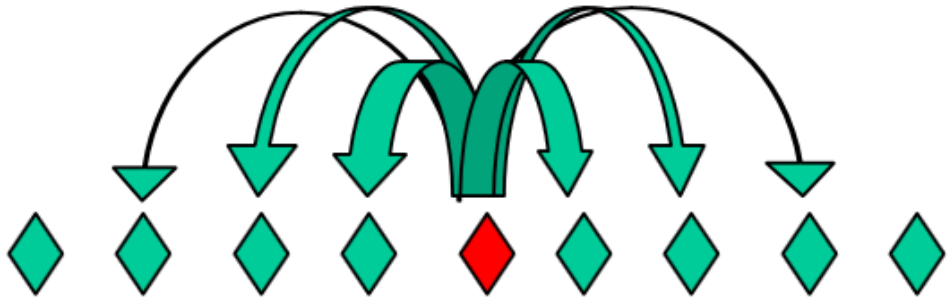
Different mutation models: uniform model

Consider a locus with L alleles and mutation rate μ.

Then $M_{a,b} = \mu/(L-1)$ for $a \neq b$ and $M_{a,a} = 1-\mu$: if an allele mutates, it becomes another allele with equal probability regardless of the allele frequencies

- Advantages
  - Simple and fast, helpful for algorithms
  - No transmissions are excluded: never a likelihood equal to zero
  - Can also be used to counter clerical errors and silent alleles
- Disadvantages
  - Unrealistic as model
  - Not stationary

Mutation models: stepwise model



$$M_{a,b} = k_a r^{|a-b|} \; for \; a \neq b$$
$$M_{a,a} = 1 - \mu$$

One can solve k_a such that we obtain a mutation matrix.

Advantages: more realistic model

Disadvantage: computationally more demanding; microvariants should be considered separately; needs estimation of parameter *r* that measures how much less likely longer mutations become. Not stationary

Stationary stepwise

- Making a matrix stationary can be done in many ways
- In Familias, this is done by adjusting the allele-specific mutation probabilties such that the conditional probabilities (conditional on a mutation having happened) are proportional to each other in the original and the stationary variant
- However, consider a very rare allele *a*
- Its neighbours *a-1* and *a+1* may mutate into it, thereby increasing its frequency in the next generation
- In order for a not to become too frequent, allele a itself has to mutate very often
- In the extreme case, a child is more likely to get allele a from a paternal *a+1* or *a-1* than from a paternal allele a
- This is a disadvantage of the stationary variant.

Theta correction

- Several finite subpopulations descending from a common ancestral population with allele frequencies p_a
- Over time each population will develop its own allele frequencies
- In expectation no tendency for allele frequencies to increase or decrease

- If we sample allele a from one of the subpopulations, this is more likely to have happened if the allele frequency of that allele is relatively high
- Thus, alleles become predictive for subpopulations and therefore dependent
- In the joint population no HWE or Linkage Equilibrium (for an example in the Dutch DNA database see [1])

Theta correction

The theta correction was designed to deal with the situation where we have a sample from all subpopulations together, without knowing how the subpopulations are represented in the sample.

It then gives the probabilities to see samples of alleles sampled from the same subpopulation.

$$P(a|n_a, n) = \frac{(1-\theta)p_a + n_a\theta}{1+(n-1)\theta}$$

However, in many cases we do the opposite: we have a sample from a subpopulation (e.g., local population) and we want to provide a conservative estimate valid for the global population.

There is no reason why the θ correction would work in this case; however (cf [2]) θ=0.03 empirically does this.

General kinship analysis

	Victim	Mother missing person M	Brother missing Person S1	Brother missing Person S2	Father (missing)
VWA	16 16	14 16	16 16	14 16	

Father	P (father)	$P(S_1, S_2, UI \mid H_p)$	$P(S_1, S_2, UI \mid H_d)$
16 16	P_{16}^2	1x½x1x½x1x½=1/8	1x½x1x½ p_{16}^2
14 16	$2p_{14}p_{16}$	1/32	1/8 p_{16}^2
16X (x ≠ 14, 16)	$2p_{16}(1-p_{14}-p_{16})$	1/64	1/16 p_{16}^2

LR for victim=missing father versus victim=random

$$LR = \frac{\frac{1}{8}p_{16}^2 + 1/32\ 2p_{14}p_{16} + 1/64\ 2p_{16}(1-p_{14}-p_{16})}{\frac{1}{4}p_{16}^4 + 1/8\ 2p_{14}p_{16}^3 + 1/16\ 2p_{16}^3(1-p_{14}-p_{16})} = \frac{1+p_{14}+3p_{16}}{p_{16}^2(4+4p_{14}+4p_{16})}$$

General principle

- A pedigree with genotypes for some of the individuals
- We want to know whether a certain person X is equal to Y, a untyped member of the pedigree
- The available genotypes predict the genotypes of the untyped individuals, in particular we get a probability distribution on the genotypes of Y given the pedigree and population allele frequencies
- For independent loci this can be done locus per locus, several algorithms and implementations available

Linkage Disequilibrium

- Locus 1 with allele frequencies p_a
- Locus 2 with allele frequencies q_a

- Haplotype frequencies H_{ab}

- If $H_{ab} - p_a q_b = 0$: "linkage equilibrium" (LE). Otherwise Linkage Disequilibrium (LD).

- This is a *statistical* property
- It does not depend on the loci themselves, e.g., loci may be in LE in a single population but not in a composed population
- Is a property similar to Hardy-Weinberg equilibrium: a statistical property, following from Mendelian segregation. LE is asymptotically reached (LD diminishes per generation) in a homogeneous infinite population if recombination is possible.

Linkage

- Not a statistical, but a *biological* concept
- Two loci are located on the same chromosome
- Person has a_1b_1 on chromosome 1, and a_2b_2 on the same chromosome
- Suppose on locus 1, allele a_1 is passed on to an offspring
- If locus 2 would be independent of locus 1, then allele b_1 or b_2 would be passed on with probability 0.5
- If the loci are on the same chromosome, then allele b_1 will be passed on if there is an even number of recombination events between the loci, and otherwise b_2.
- For small distances, 0 recombination events may have a large probability
- Let ρ be the probability of having a recombined offspring (a_1b_2 or a_2b_1), then $0 \leq \rho \leq 0.5$;
- ρ=0.5 corresponds to independence and ρ=0 to full linkage

Relevance of linkage to calculations

If a person has genotype (a_1,a_2) on locus 1 and (b_1,b_2) on locus 2, then he or she may have as chromosomes a_1b_1 / a_2b_2 or a_1b_2 / a_2b_1; we can't see which.

LE implies that these two combinations are equally likely (exercise).

Offspring is therefore equally likely to get either one of those four.

Observing an offspring gives information and makes one combination more likely than the other one, therefore also altering the probabilities for what the next offspring will get.

In general, linkage is therefore influential on the genotype likelihoods as soon as there are individuals with *more than one descendant.*

So, (with LE) not altering parent-child comparisons.

Effect of linkage

Consider two loci that are fully linked ($\rho=0$) and a pairwise kinship comparison for Hp: κ_0 , κ_1 , κ_2 versus Hd: unrelated.

- Suppose person 1 has genotype (a_1,a_2), (b_1,b_2) on the two loci
- Person 2 has (a_3,a_4), (b_3,b_4)
- If there are no shared alleles between (a_1,a_2) and (a_3,a_4) or between (b_1,b_2) and (b_3,b_4) then there can be no IBD alleles on that, and therefore on neither, locus. Thus, LR= κ_0
- If Hp is true, then the average LR in favour of Hp is more than for independent loci
- If Hd is true, LR= κ_0 is the smallest possible
- With many fully linked markers:
  - If Hp true: get either LR= κ_0 (with probability slightly less than κ_0) or a very large LR
  - If Hd is true: almost always LR=κ_0 and sometimes large false positive

Many linked versus many unlinked markers

- Independent: LR's multiply over the loci
- The weight of evidence (Log(LR)) is additive, and the distribution will become approximately normal

- Linked loci: LR's do not multiply, depend on all loci together
- Log(LR) distribution for a set of linked loci is nowhere near normally distributed

- As seen in previous slide: get either weak evidence for H_d or strong evidence for H_p. If H_d is true the first usually is obtained, if H_p is true the latter.

LR computation

Hp: PoI is person X in pedigree

Hd: PoI is a random person unrelated to the pedigree

LR computation corresponds to

- From the available pedigree data, predict the genotype of person X
- Match the genotype of the PoI with the inferred genotypes for X weighted with their probabilities, i.e.,
- LR for a PoI with genotype g is $P(G_X=g)/f_g$

Thus, algorithms are needed to predict the genotype of X.

- Elston-Stewart (Familias): inbreeding and mutations fine, linkage problematic
- Lander-Green (FamLink): linkage fine, mutations and inbreeding not.

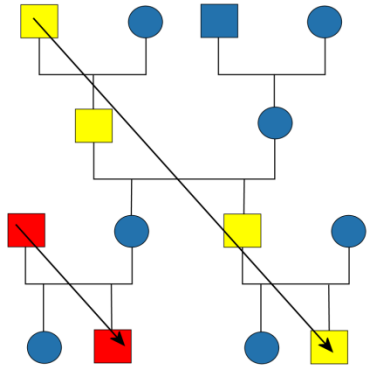
Algorithms

For this reason, linkage has not been taken into account into Familias but is implemented separately into FamLink

Note however, FamLink cannot deal with mutations, nor with theta.

Y-STR markers

- Y-chromosomal data: transmitted from father to son
- No recombination, basically one big allele



Weight of evidence from Y-STR markers

Various methods are available for the calculation of the LR in case of matching Y-profiles.

E.g., child C and alleged father AF:

- H_p: AF is the father of C
- H_d: AF is unrelated to C
- E: haplotypes G_C and G_{AF}, $G_C = G_{AF}$

However, note that

- The notion of unrelatedness is more difficult for Y-STR's: when are two persons unrelated?
- The LR can be written as approximately $1/P(G_C \mid G_{AF}, H_d)$, where C is considered as a random person from the population

Methods

- Population genetic models that assign a frequency to all possible haplotypes, e.g. Discrete Laplace method
- Combinatorial methods based on the distribution of haplotypes in the reference database e.g. kappa method

- These methods are conceptually quite different. The first type aims to predict population frequencies for all profiles, such as is done for autosomal profiles. The second type does not do that, it only applies to the case at hand with these two profiles

Discrete Laplace

- Population is a disjoint union of subpopulations
- In each subpopulation, every locus has a central allele and the allele frequencies are descending, dependent only on the distance to the central allele (a discrete Laplace distribution)
- In each subpopulation, haplotype frequency obtained as product of the allele frequencies

- On every locus, two parameters: the position of the central allele and the spread of the allele frequency distribution. These are estimated

- Theoretical motivation: alleles on a single locus in an evolving population tend to 'stay together' and the allele frequency distribution can be approximated by a discrete Laplace
- See [3]

Kappa method

- Designed to deal with new haplotypes (i.e., not observed in database)
- If the database is large, then the probability that a haplotype sampled from the population is a new one, is about the same as the probability that the last profile added to the database was a new one, which is the fraction of single-copy haplotypes in the database. Call this fraction κ.
- Under Hd we need to calculate the probability that C has the same profile as AF, which is a profile that is a singleton in the extended database (=database + AF).
- The probability that C's profile is not new to the (extended) database is 1- κ and the profile frequency is then estimated at 1/n (it being a singleton).
- The LR is then (1- κ)/n where *n* is the size of the database
- See [4]

Generalizations (Good-Turing)

Under H_p, a single profile has been seen and it is unobserved in the database.

Under H_d, two profiles have been seen, they are identical and did not occur in the database.

Let C_i be the number of profiles in the database that are in a cluster of exactly i matching profiles.

The probability for E, is $\kappa=C_1/n$ under H_p, and is $2C_2/(n(n-1))$ under H_d

Thus LR=$\frac{C_1}{n}\frac{n(n-1)}{2C_2} \approx n\frac{C_1}{2C_2}$

(see also [5])

Difference

The latter methods were not set up to give a frequency to all profiles.

They model purely how likely it is to have the given observations in terms of the combinatorics of the already observed profiles.

The haplotype itself is not used, only its count.

Y-STR profile frequency has strong local variation

- Reference databases are worldwide
- Does not always contain the relevant population
- Y-STR profiles behave like family names: may be very common in a specific area and hardly exist outside that
- Therefore, the second hypothesis (random man) should specify where that random man is from, but then it is in many cases not so clear how representative available databases still are

- Same comments apply also to mitochondrial DNA.

References

[1] M.V. Kruijver, Characterizing the genetic structure of a forensic DNA database using a latent variable approach, FSI:G 23, 2016

[2] C. Steele et. al., Worldwide math formula Estimates Relative to Five Continental-Scale Populations, Annals of Human Genetics, 2014

[3] M.M. Andersen, P.S. Eriksen, N. Morling, The discrete Laplace exponential family and estimation of Y-STR haplotype frequencies., J. Theor. Biol 2014

[4] Brenner CH (2010) Fundamental problem of forensic mathematics – The evidential value of a rare haplotype

Forensic Sci. Int. Genet. 4 281–291

[5] G. Cereda, Impact of model choice on LR assessment in case of rare haplotype match (frequentist approach), https://arxiv.org/pdf/1502.04083.pdf