JIMA Vol. 73, No. 5 (Dec. 2021) pp. 189-201.

Review

R scripts for kinship testing

Masataka TAKAMIYA

Division of Forensic Medicine, Department of Forensic Science,
School of Medicine, Iwate Medical University, Yahaba, Japan

(Received on June 30, 2021 & Accepted on July 19, 2021)

Abstract

Programming scripts were written for the statistical analysis of genetic data for kinship testing. The methods presented here involve algorithms with R for conditional probability analysis based on Bayes' theorem. Although computer programs to assess probable kinship have been published, the reports describe only minimal or generalized formulas, rendering the programs difficult for most forensic researchers to understand. Here, we present details of calculations used to evaluate probabilities of kinship and details of the R scripts used to execute these calculations. Scripts were constructed for paternity trio testing in which DNA typing data for the mother, child, and alleged father are available as well as for paternity duo, parental, grandfather, sibling, subsibling, and uncle testing. This review provides clear and orderly descriptions of calculations using R scripts, making each facet of this method easy to understand. Furthermore, access to these scripts will enable researchers to develop their own calculation systems.

Key words : *forensic mathematics, kinship testing, R, DNA typing*

I. Introduction

DNA profiling is considered one of the most important aspects of forensic science, particularly in kinship testing [1], which involves determination of the characteristics of an individual's DNA. After interpretation of DNA profiles, statistical analyses are performed to determine relatedness between individuals [2]. A number of mathematical formulas for evaluating the plausibility of kinship based on Bayes' theorem have been developed [3]. The representative case of questioned kinship is a paternity test to assess the relatedness of a trio of individuals: a mother, a child, and an alleged father. Paternity duo, parental, grandfather, sibling, subsibling, and uncle tests are also utilized in determining kinship. In addition to being tedious and time-consuming when performed manually, these statistical calculations are associated with potential errors or ambiguities [1]. A number of computer programs have therefore been written for use in determining kinship based on genetic data [4-19]. However, most of the published reports describing these programs provide only minimal and generalized calculations and are therefore

Corresponding author: Masataka Takamiya
mmmtakamiya@gmail.com

accessible only to mathematicians and researchers with an advanced mathematics background.

R, an open-source tool distributed free of charge by the R Project for Statistical Computing, is a widely used programming language for executing calculations and statistical analyses [20]. The merits of R for mathematical analyses have been described previously [21-23]. No extensive programming skills are required to exploit the advantages of R, and it is easy to modify programs written in R. As such, researchers can use R to construct calculation systems tailored to their specific purposes. Therefore, every calculation necessary for resolving a kinship question can be clearly articulated in R scripts. The aims of this review are to describe R scripts for kinship determination and show the detailed and orderly associated calculations.

II. R scripts for kinship testing

To determine kinship in cases involving DNA typing data, computer programs have been constructed based on conditional probability determination using Bayes' theorem. Each program relies on autosomal markers at a single locus and was constructed under the assumption that all DNA-based genotyping data are correct. The programs were built in R, and the calculations were based on formulas developed by Aoki et al. [6, 16]. The contents of each script include paternity trio, paternity duo, parental, grandfather, sibling, subsibling, and uncle tests. Generalized pedigree trees are presented in Figures 1 and 2. Sibling, subsibling, and uncle tests are used in cases in which DNA typing data for the parents and grandparents are not available, and the probable genotypes were constructed based on DNA typing data for one or various combinations of siblings, subsiblings, and uncles. Each program can accommodate a maximum of 5 siblings, 5 subsiblings, and 5 uncles. For each program, this review describes 1) the mathematical theory, and 2) the R scripts. The R scripts in this section are described in pseudo-codes to make them easily understandable.

1. Paternity trio test

1) Theory

This example represents the standard trio case involving a mother, a child whose paternity is in question, and an alleged father (Fig.1). Based on Bayes' theorem, Essen-Moller [24, 25] devised a formula to evaluate the probability of paternity. The probabilities calculated using the Essen-Moller formula are defined as follows:

X: probability (type of the father|the tested person is the child)

Y: probability (type of the father|the tested person is a random person)

The equation W=X/(X+Y) can be used to transform DNA types into numerical expressions representing the likelihood of paternity. In addition, the X/Y ratio was proposed to represent the paternity index (PI) [26, 27].

Komatsu [28-30] also used Bayes' theorem to develop a formula to calculate the probability of paternity. The probabilistic purpose was as follows:

probability (the tested person is the father|type of the child).

Therefore, the probabilities calculated using the Komatsu formula and Bayes' theorem are defined as follows:

P1: probability (type of the child|the tested

person is the father)

P2: probability (type of the child|the tested person is a random person)

The prior probabilities of P1 and P2 are both 0.5. In Komatsu's methodology, the equation W=P1/(P1+P2) is used to express the likelihood of paternity, and the ratio P1/P2 is used to represent the PI. Results obtained using the Essen-Moller formulas, specifically the values for the likelihood of paternity and the PI, are consistent with those obtained using the Komatsu formulas. Komatsu formulas were used in the present work because they are suitable for constructing R scripts.

2) R scripts

Genotypes of the mother, child, and alleged father were represented as [m1, m2], [c1, c2], and [f1, f2], respectively, where "m1" and "m2" represent both of the mother's alleles of one gene, "c1" and "c2" represent both of the child's alleles of that same gene, and "f1" and "f2" represent both of the alleged father's alleles of that gene. The probability that c1 was inherited from the alleged father was represented by the term fc1, which was calculated using the following R script:

fc1<-{ifelse(f1==c1,1,0)+ifelse(f2==c1,1,0)}/2

[Script 1 1]

In R script, "<-" represents the definition sign, and "ifelse(f1==c1,1,0)" directs a return of 1 if f1 equals c1 or a return of 0 if f1 differs from c1, because "==" represents the equal sign.

The probability that c2 was inherited from the alleged father was represented by the term fc2.

fc2<-{ifelse(f1==c2,1,0)+ifelse(f2==c2,1,0)}/2

[Script 1 2]

The probability that c1 was inherited from the mother was represented by the term mc1.

mc1<-{ifelse(m1==c1,1,0)+ifelse(m2==c1,1,0)}/2

[Script 1 3]

The probability that c2 was inherited from the mother was represented by the term mc2.

mc2<-{ifelse(m1==c2,1,0)+ifelse(m2==c2,1,0)}/2

[Script 1 4]

Using these probabilities, P1 and P2 were calculated as follows:

P1<- (fc1*mc2+fc2*mc1)/ifelse(c1==c2,2,1)

[Script 1 5]

P2<-([gene frequency of c1]*mc2+[gene frequency of c2]*mc1/ifelse(c1==c2,2,1)

[Script 1 6]

2. Paternity duo test

1) Theory

The genotype of the mother was unavailable, but the genotypes of the child and alleged father were available (Fig.1). Because one of the child's alleles was inherited from an unspecified mother, mc1 and mc2 were set to the gene frequencies of c1 and c2, respectively.

3. Parental test

1) Theory

This example involves a child whose parentality is in question, and the alleged mother and father (Fig.1). The probabilistic purpose is as follows:

probability (the tested persons are the father and mother|type of the child).

The probabilities calculated using the Komatsu formula and Bayes' theorem are defined as follows:

P1: probability (type of the child|the tested persons are the father and mother)

P2: probability (type of the child|the tested persons are 2 random persons)

The equation W=P1/(P1+P2) is used to express the likelihood of parentality, and the

P1/P2 ratio is used to represent the parental index.

2) R script

Genotypes of the alleged mother, child, and alleged father are represented as [m1, m2], [c1, c2], and [f1, f2], respectively, where "m1" and "m2" represent both of the mother's alleles of one gene, "c1" and "c2" represent both of the child's alleles of that same gene, and "f1" and "f2" represent both of the alleged father's alleles of that gene. The probability that c1 was inherited from the alleged father is represented by the term fc1, which is calculated using the following R script:

```
fc1<-{ifelse(f1==c1,1,0)+ifelse(f2==c1,1,0)}/2
```

[Script 3 1]

The probability that c2 was inherited from the alleged father is represented by the term fc2.

```
fc2<-{ifelse(f1==c2,1,0)+ifelse(f2==c2,1,0)}/2
```

[Script 3 2]

The probability that c1 was inherited from the mother is represented by the term mc1.

```
mc1<-{ifelse(m1==c1,1,0)+ifelse(m2==c1,1,0)}/2
```

[Script 3 3]

The probability that c2 was inherited from the mother is represented by the term mc2.

```
mc2<-{ifelse(m1==c2,1,0)+ifelse(m2==c2,1,0)}/2
```

[Script 3 4]

Using these probabilities, P1 and P2 are calculated as follows:

```
P1<- (fc1*mc2+fc2*mc1)/ifelse(c1==c2,2,1)
```

[Script 3 5]

```
P2<-([gene frequency of c1]* [gene frequency of c2]+[gene frequency of c2]*[gene frequency of c1])/ifelse(c1==c2,2,1)
```

[Script 3 6]

4. Grandfather test

1) Theory

This example involves a child whose grandpaternity is in question and the alleged grandfather (Fig.1). The probabilistic purpose is as follows:

probability (the tested person is the grandfather|type of the child).

Therefore, the probabilities calculated using the Komatsu formula and Bayes' theorem are defined as follows:

P1: probability (type of the child|the tested person is the grandfather)

P2: probability (type of the child|the tested person is a random person)

The equation W=P1/(P1+P2) is used to express the likelihood of grandpaternity, and the ratio P1/P2 is used to represent the grandpaternity index.

2) R script

Genotypes of the child and alleged grandfather are represented as [c1, c2], and [gf1, gf2], respectively, where "c1" and "c2" represent both of the child's alleles of a specific gene, and "gf1" and "gf2" represent both of the alleged grandfather's alleles of that same gene. The probability that c1 was inherited from the alleged grandfather is represented by the term gfc1, which is calculated using the following R script:

```
gfc1<-{ifelse(gf1==c1,1,0)+ifelse(gf2==c1,1,0)}/2
/2+[gene frequency of c1]/2
```

[Script 4 1]

The probability that c2 was inherited from the alleged grandfather is represented by the term gfc2.

```
gfc2<-{ifelse(gf1==c2,1,0)+ifelse(gf2==c2,1,0)}/2
/2+[gene frequency of c2]/2
```

[Script 4 2]

Using these probabilities, P1 and P2 are calculated as follows:

P1<- (gfc1*[gene frequency of c2]+gfc2*[gene frequency of c1])/ifelse(c1==c2,2,1)

[Script 4 3]

P2<-([gene frequency of c1]* [gene frequency of c2])*ifelse(c1==c2,1,2)

[Script 4 4]

5. Sibling test

1) Theory

This example involves a person and that person's alleged siblings (Fig. 1). The probabilistic purpose is as follows:

probability (the tested persons are siblings|type of the person).

Therefore, the probabilities calculated using the Komatsu formula and Bayes' theorem are defined as follows:

P1: probability (type of the person|the tested persons are siblings)

P2: probability (type of the person|the tested persons are random persons)

The equation W=P1/(P1+P2) is used to express the likelihood of the tested persons being siblings, and the P1/P2 ratio is used to represent the sibling index.

P1 is established in two phases:

P1-1: probability (type of the parents|the tested persons are siblings)

P1-2: probability (type of the person|type of the parents)

The putative genotypes of the father and mother are constructed based on the genotype(s) of one or more siblings. Bayes' theorem is used to calculate a conditional probability for the putative genotypes of the father and mother. The genotypes of the parents and anywhere from one to five siblings are represented as [PA1, PA2], [PB1, PB2], and [S1, S2] (which represents $[S1,S2]_1$ or some sequential combination of $[S1,S2]_1$, $[S1,S2]_2$, $[S1,S2]_3$, $[S1,S2]_4$, and $[S1,S2]_5$). Given [S1, S2], the posterior probability of $[PA1, PA2]_1$ and $[PB1, PB2]_1$ is calculated as follows:

$$P([PA1,PA2]_1, [PB1,PB2]_1 | [S1,S2]) = \frac{P([PA1,PA2]_1, [PB1,PB2]_1) \prod_{k=1}^{n} P([S1,S2]_k | [PA1,PA2]_1, [PB1,PB2]_1)}{\sum_{j=1}^{m} P([PA1,PA2]_j, [PB1,PB2]_j) \prod_{k=1}^{n} P([S1,S2]_k | [PA1,PA2]_j, [PB1,PB2]_j)}$$

[Formula 5 1]

When the genotype of the person is represented as [C1, C2], P1-2 is calculated as follows:

$$\sum_{j=1}^{m} P([C1,C2] | [PA1,PA2]_j, [PB1,PB2]_j)$$

[Formula 5 2]

2) R script

Genotypes of the sibling and parents are represented as [s1, s2], [pa1, pa2], and [pb1, pb2], respectively, where "s1" and "s2" represent both of the sibling's alleles of a specific gene, and "pa1" and "pa2" and "pb1" and "pb2" represent both alleles of that gene for one of the parents. The posterior probability of $[pa1, pa2]_1$ and $[pb1,pb2]_1$ is assessed as follows: the probability that s1 was inherited from one of the parents (genotype: [pa1, pa2]) is represented by the term pas1.

pas1<-{ifelse(pa1==s1,1,0)+ifelse(pa2==s1,1,0)}/2

[Script 5 1]

The probability that s2 was inherited from one of the parents (genotype: [pa1, pa2]) is represented by the term pas2.

pas2<-{ifelse(pa1==s2,1,0)+ifelse(pa2==s2,1,0)}/2

[Script 5 2]

The probability that s1 was inherited from the remaining parent (genotype: [pb1, pb2]) is represented by the term pbs1.

pbs1<-{ifelse(pb1==s1,1,0)+ifelse(pb2=

=s1,1,0)}/2

[Script 5 3]

The probability that s2 was inherited from the remaining parent (genotype: [pb1, pb2]) is represented by the term pbs2.

pbs2<-{ifelse(pb1==s2,1,0)+ifelse(pb2==s2,1,0)}/2

[Script 5 4]

The posterior probability of $[pa1, pa2]_1$ and $[pb1, pb2]_1$ is represented by the term piba1 and calculated using the following scripts:

pabs1<-(pas1*pbs2+pas2*pbs1)/ifelse(s1==s2,2,1)

[Script 5 5]

piba1<-pab1s1*pab1s2*pab1s3*pab1s4*pab1s5*[frequency of $[pa1, pa2]_1$ and $[pb1, pb2]_1$]/

Σpabjs1*pabjs2*pabjs3*pabjs4*pabjs5*[frequency of $[pa1, pa2]_j$ and $[pb1, pb2]_j$]

[Script 5 6]

The genotype frequency of [pa1, pa2] is calculated using the following formulas:

[pa1, pa2] is homozygous:

1*[allele frequency of pa1]*[allele frequency of pa2]

[pa1, pa2] is heterozygous:

2*[allele frequency of pa1]*[allele frequency of pa2]

The following structure was identical to [Script 1 5].

6. Subsibling test

1) Theory

This example involves a person and that person's alleged subsiblings (Fig. 2). The probabilistic purpose is as follows:

probability (the tested persons are subsiblings| type of the person).

Therefore, the probabilities calculated using the Komatsu formula and Bayes' theorem are defined as follows:

P1: probability (type of the person|the tested persons are the subsiblings)

P2: probability (type of the person|the tested persons are random persons)

The equation W=P1/(P1+P2) is used to express the likelihood of the tested persons being subsiblings, and the P1/P2 ratio is used to represent the subsibling index.

P1 is established in two phases:

P1-1: probability (type of the parents|the tested persons are subsiblings)

P1-2: probability (type of the person|type of the parents)

The putative genotypes of the parents are constructed based on the genotype(s) of one or more subsiblings. Bayes' theorem is used to calculate a conditional probability for the putative genotypes of the parents. The genotypes of the parents and anywhere from one to five subsiblings are represented as [PA1, PA2], [PB1, PB2], and [S1, S2] (which represents $[S1,S2]_1$ or some sequential combination of $[S1,S2]_1$, $[S1,S2]_2$, $[S1,S2]_3$, $[S1,S2]_4$, and $[S1,S2]_5$). Given [S1, S2], the posterior probability of $[PA1, PA2]_1$ and $[PB1, PB2]_1$ is calculated as follows:

$$P([PA1,PA2]_i, [PB1,PB2]_1 | [S1,S2]) = \frac{P([PA1,PA2]_i, [PB1,PB2]_i) \prod_{k=1}^{n} P([S1,S2]_k | [PA1,PA2]_i, [PB1,PB2]_i)}{\sum_{j=1}^{m} P([PA1,PA2]_j, [PB1,PB2]_j) \prod_{k=1}^{n} P([S1,S2]_k | [PA1,PA2]_j, [PB1,PB2]_j)}$$

[Formula 6 1]

When genotypes of the person are represented as [C1, C2], P1-2 is calculated as follows:

$$\sum_{j=1}^{m} P([C1,C2] | [PA1,PA2]_j, [PB1,PB2]_j)$$

[Formula 6 2]

2) R script

Genotypes of the subsiblings, parents, person

are represented as [s1, s2], [pa1, pa2] and [pb1, pb2], and [c1, c2], respectively, where "s1" and "s2" represent both of the subsibling's alleles of a specific gene, and "pa1" and "pa2", "pb1" and "pb2", "c1" and "c2", represent both of the alleles of that gene for the parents and person. The posterior probability of $[pa1, pa2]_l$ and $[pb1, pb2]_l$ is assessed as follows: the probability that s1 was inherited from one of the parents (genotype: [pa1, pa2]) is represented by the term pas1.

```
pas1<-{ifelse(pa1==s1,1,0)+ifelse(pa2==s1,1,0)}/2
```

[Script 6 1]

The probability that s2 was inherited from one of the parents (genotype: pa1, pa2]) is represented by the term pas2.

```
pas2<-{ifelse(pa1==s2,1,0)+ifelse(pa2==s2,1,0)}/2
```

[Script 6 2]

The probability that s1 was inherited from the remaining parent (genotype: [pb1, pb2]) is represented by the term pbs1.

```
pbs1<-{ifelse(pb1==s1,1,0)+ifelse(pb2==s1,1,0)}/2
```

[Script 6 3]

The probability that s2 was inherited from the remaining parent (genotype: [pb1, pb2]) is represented by the term pbs2.

```
pbs2<-{ifelse(pb1==s2,1,0)+ifelse(pb2==s2,1,0)}/2
```

[Script 6 4]

The posterior probability of $[pa1, pa2]_l$ and $[pb1, pb2]_l$ is represented by the term pibal and calculated using the following scripts:

```
pabs1<-(pas1*pbs2+pas2*pbs1)/ifelse(s1==s2,2,1)
```

[Script 6 5]

```
pibal<-pabls1*pabls2*pabls3*pabls4*pabls5*[frequency of [pa1, pa2]l and [pb1, pb2]l]/
Σ pabjs1*pabjs2*pabjs3*pabjs4*pabjs5*[frequency of [pa1, pa2]j and [pb1, pb2]j]
```

[Script 6 6]

The probability that c1 was inherited from the parents (genotype: [pa1, pa2], [pb1, pb2]) is represented by the term pc1, which is calculated using the following R script:

```
pc1<-
{ifelse(pa1==c1,1,0)+ifelse(pa2==c1,1,0)}/2/2
+{ifelse(pb1==c1,1,0)+ifelse(pb2==c1,1,0)}/2/2
```

[Script 6 7]

The probability that c2 was inherited from the parents is represented by the term pc2.

```
pc2<-
{ifelse(pa1==c2,1,0)+ifelse(pa2==c2,1,0)}/2/2
+{ifelse(pb1==c2,1,0)+ifelse(pb2==c2,1,0)}/2/2
```

[Script 6 8]

Using these probabilities, P1 and P2 are calculated as follows:

```
P1<- (pc1*[gene frequency of c2]+pc2*[gene frequency of c1])/ifelse(c1==c2,2,1)
```

[Script 6 9]

```
P2<-([gene frequency of c1]* [gene frequency of c2])*ifelse(c1==c2,1,2)
```

[Script 6 10]

7. Uncle test

1) Theory

This example involves a child and alleged uncles (Fig. 2). The probabilistic purpose is as follows:

probability (the tested persons are uncles|type of the child).

Therefore, the probabilities calculated using the Komatsu formula and Bayes' theorem are defined as follows:

P1: probability (type of the child|the tested persons are uncles)

P2: probability (type of the child|the tested

persons are random persons)

The equation W=P1/(P1+P2) is used to express the likelihood of the tested person being an uncle of the child, and the P1/P2 ratio is used to represent the uncle index.

P1 is established in two phases:

P1-1: probability (type of the grand-parents| the tested persons are uncles)

P1-2: probability (type of the child|type of the grandparents)

The putative genotypes of the grandparents are constructed based on the genotype(s) of one or more uncles. Bayes' theorem is used to calculate a conditional probability for the putative genotypes of the grandparents. The genotypes of the grandparents and anywhere from one to five uncles are represented as [GA1, GA2], [GB1, GB2], and [S1, S2] (which represents $[S1,S2]_1$ or some sequential combination of $[S1,S2]_1$, $[S1,S2]_2$, $[S1,S2]_3$, $[S1,S2]_4$, and $[S1,S2]_5$). Given [S1, S2], the posterior probability of $[GA1, GA2]_1$ and $[GB1, GB2]_1$ is calculated as follows:

[Formula 7 1]

$$P([GA1,GA2]_p,\ [GB1,GB2]_1 \mid [S1,S2]) = \frac{P([GA1,GA2]_p,\ [GB1,GB2]_1)\ \prod_{k=1}^{n} P([S1,S2]_k \mid [GA1,GA2]_p,\ [GB1,GB2]_1)}{\sum_{p=1}^{m} P([GA1,GA2]_p,\ [GB1,GB2])\ \prod_{k=1}^{n} P([S1,S2]_k \mid [GA1,GA2]_p,\ [GB1,GB2])}$$

When the genotype of the child is represented as [C1, C2], P1-2 is calculated as follows:

$$\sum_{p=1}^{m} P([C1,C2] \mid [GA1,GA2]_p,\ [GB1,GB2])$$

[Formula 7 2]

2) R script

The genotypes of the uncle, grandparents, and child are represented as [s1, s2], [ga1, ga2] and [gb1, gb2], and [c1, c2], respectively, where "s1" and "s2" represent both of the uncle's alleles of a specific gene, and "ga1" and "ga2", "gb1" and "gb2" , and "c1" and "c2" represent both of the alleles of that gene for the grandparents and child. The posterior probability of $[ga1, ga2]_1$ and $[gb1, gb2]_1$ is assessed as follows: the probability that s1 was inherited from one of the grandparents (genotype: [ga1, ga2]) is represented by the term gas1.

```
gas1<-{ifelse(ga1==s1,1,0)+ifelse(ga2==s1,1,0)}/2
```

[Script 7 1]

The probability that s2 was inherited from one of the grandparents (genotype: [ga1, ga2]) is represented by the term gas2.

```
gas2<-{ifelse(ga1==s2,1,0)+ifelse(ga2==s2,1,0)}/2
```

[Script 7 2]

The probability that s1 was inherited from the remaining grandparent (genotype: [gb1, gb2]) is represented by the term gbs1.

```
gbs1<-{ifelse(gb1==s1,1,0)+ifelse(gb2==s1,1,0)}/2
```

[Script 7 3]

The probability that s2 was inherited from this remaining grandparent (genotype: [gb1, gb2]) is represented by the term gbs2.

```
gbs2<-{ifelse(gb1==s2,1,0)+ifelse(gb2==s2,1,0)}/2
```

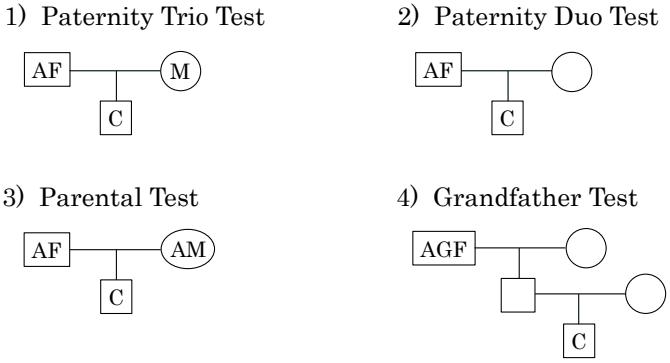
[Script 7 4]

The posterior probability of $[ga1, ga2]_1$ and $[gb1, gb2]_1$ is represented by the term gpiba1 and calculated using the following scripts:

```
gabs1<-(gas1*gbs2+gas2*gbs1)/ifelse(s1==s2,2,1)
```

[Script 7 5]

```
gpiba1<-gab1s1*gab1s2*gab1s3*gab1s4*gab1s5*[frequency of [ga1, ga2]1 and [gb1, gb2]1]/
Σgabjs1*gabjs2*gabjs3*gabjs4*gabjs5*[frequen
```



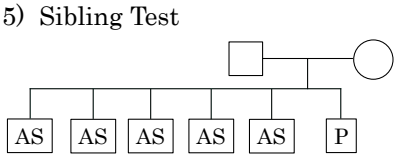
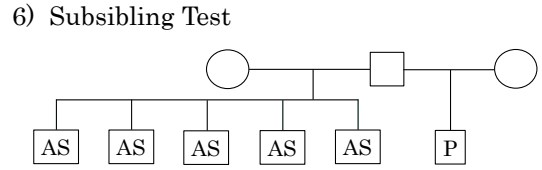


Fig. 1. Pedigrees for a 1) paternity trio test, 2) paternity duo test, 3) parental test, 4) grandfather test, and 5) sibling test. AF, alleged father; M, mother; C, child; AM, alleged mother; AGF, alleged grandfather; AS, alleged sibling; P, person; Blank, No DNA genotyping available.



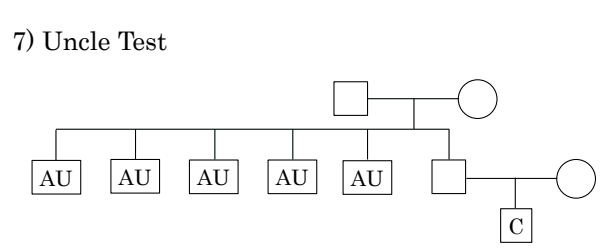


Fig. 2. Pedigrees for a 6) subsibling test and 7) uncle test. AU, alleged uncle; Other explanatory notes are the same as those for Figure 1.

cy of [ga1, ga2]$_j$ and [gb1, gb2]$_j$]

[Script 7 6]

The probability that c1 was inherited from the grandparents (genotype: [ga1, ga2], [gb1, gb2]) is represented by the term gc1, which is calculated using the following R script:

```
gc1<-
{ifelse(ga1==c1,1,0)+ifelse(ga2==c1,1,0)}/2/2
+{ifelse(gb1==c1,1,0)+ifelse(gb2==c1,1,0)}/2/2
```

[Script 7 7]

The probability that c2 was inherited from the grandparents is represented by the term gc2.

```
gc2<-
{ifelse(ga1==c2,1,0)+ifelse(ga2==c2,1,0)}/2/2
+{ifelse(gb1==c2,1,0)+ifelse(gb2==c2,1,0)}/2/2
```

[Script 7 8]

Using these probabilities, P1 and P2 are calculated as follows:

```
P1<- (gc1*[gene frequency of c2]+gc2*[gene
frequency of c1])/ifelse(c1==c2,2,1)
```

[Script 7 9]

```
P2<-([gene frequency of c1]* [gene frequency
of c2])*ifelse(c1==c2,1,2)
```

[Script 7 10]

```
x<-c(0,1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20)
y<-letters[1:21]
z<-c(1,0.001,0.002,0.003,0.004,0.005,0.006,0.007,0.008,0.009,0.010,0.011,0.012,0.013,0.014,0.015,0.016,0.017,0.018,0.019,0.81)
frequency<-data.frame(x,y,z)

#prior probability    tpp<-0.5    fpp<-0.5

#father    f1<-2    f2<-4
#mother    m1<-2    m2<-4
#child    c1<-2    c2<-4

fc1<-({ifelse(f1==c1,1,0)}+{ifelse(f2==c1,1,0)})/2
fc2<-({ifelse(f1==c2,1,0)}+{ifelse(f2==c2,1,0)})/2
mc1<-({ifelse(m1==c1,1,0)}+{ifelse(m2==c1,1,0)})/2
mc2<-({ifelse(m1==c2,1,0)}+{ifelse(m2==c2,1,0)})/2

ba1<-{fc1*mc2+fc2*mc1}/{ifelse(c1==c2,2,1)}

selectionc1<-{subset(frequency,subset=x==c1,select=c(z))}
selectionc2<-{subset(frequency,subset=x==c2,select=c(z))}

ba2<-{(selectionc1*mc2+selectionc2*mc1)/ifelse(c1==c2,2,1)}

probability<-{(tpp*ba1)/(tpp*ba1+fpp*ba2)}
ratio<-{(ba1)/(ba2)}
```

Fig. 3. Script for the paternity trio test. The genotypes of the mother, child, and alleged father are represented as [2, 4], [2, 4], and [2, 4], respectively. The function of each code is as follows; <-, definition; ==, equal; ifelse, conditional element selection; data.frame, store of data tables; subset, conditional element selection. In addition, the definition of each formula is shown in subsections "1. Paternity trio test" and "8. Instructions for users" in section II. R scripts for kinship testing.

8. Instructions for users

Figure 3 shows the script for the case of paternity trio testing. The term x represents an individual allele and is expressed as an Arabic numeral between 0 and 20. The term 'z' represented the frequency of a particular allele, which corresponds to the number of x by turns. Genotypes are then expressed by Arabic numerals for term x. As R is a mathematical software application, alleles must be expressed in Arabic numerals. For example, it is not possible to analyze a genotype represented as [gene^{A}, gene^{a}]. In such cases, "gene^{A} " is replaced with "1" in term x, and " gene^{a}" is replaced with "2" . Moreover, genotypes must be listed in ascending numerical order: for example, [2, 4] is suitable, but [4, 2] is not. For any unavailable allele, "0" is used in term x: for example, if the genotype of sibling 5 is unavailable, this genotype is expressed as [0, 0]. In addition, "0" in term x corresponds to "1" in term z. The likelihood of kinship and the kinship index are then calculated when the terms "probability" and "ratio" , respectively, are entered.

III. Discussion

Statistical calculations using allele frequencies of tested loci are important for establishing links among DNA profiles of individuals [1]. Computer-aided approaches are increasingly popular for performing DNA-related statistical analyses [1]. A number of programs have been proposed for kinship analyses using DNA information [4-19], and these computer programs can be classified into two categories.

The programs in the first category employ distinctive pull-down menus. Those in the second category use common software applications, and the end users must perform manual operations, such as general computer manipulations and spreadsheet selections.

Programs with pull-down menus make operations easy for most investigators because only inputs and click selections of DNA types and gene frequencies are required for calculating the kinship likelihood and index. However, published papers describing the theories behind the calculations used in programs with pull-down menus generally provide only minimal formulas. In actuality, the pedigrees and cells into which DNA typing information is entered are fixed in these programs. Therefore, modifications to the system and flexible extensions cannot be easily implemented by researchers. Moreover, some programs are commercial and extremely expensive, which restricts their use in examining pedigrees to researchers at only a small number of institutions.

Software applications that require more manual operation by the user are more flexible and can be used for more pedigrees; these programs also provide valuable information about the calculations. Viewed in this light, the Aoki spreadsheet [6, 16] remains one of the most innovative and promising. The Aoki spreadsheet was constructed in Excel (Microsoft, Redmond, WA); consequently, the mathematical and genetic hypotheses necessary to analyze pedigrees are easily accessible to researchers. Increasingly, however, R is becoming the standard programming language for biostatistics [21-23], and it is distributed free of charge. Given these circumstances, I created kinship testing scripts using R. I also confirmed that these scripts generated the same values for the likelihood of kinship and the kinship index as were previously reported [6, 16, 31, 32]. In this review, I have described the formulas in an orderly fashion so that investigators who lack a programming background should be able to understand the calculations by reading the scripts. Modifications to the R scripts, such as changes to the number of relatives, should be relatively easy to implement by copying and pasting existing formulas. Each of the R scripts presented here is an open-source tool, and I am happy to share them with any researcher (http://mtakamiya.starfree.jp/kinshiptest/indexkinshiptest.html).

This review describes fundamental R formulas that can be used for kinship testing. By learning these scripts, any interested researcher can develop their own calculation systems.

Conflict of interest: The author has no conflict of interest to declare.

References

1) **Rasool N** and **Hussain W**: ForeStatistics: A windows-based feature-rich software program for performing statistics in forensic DNA analysis, paternity and relationship testing. Forensic Sci Int **307**, 110142, 2020.
2) **McDonald J** and **Lehman DC**: Forensic DNA analysis. Clin Lab Sci **25**, 109-113, 2012.
3) **Gjertson DW, Brenner CH, Baur MP, et al.**: ISFG: Recommendations on biostatistics in paternity testing. Forensic Sci Int Genet **1**, 223-

231, 2007.
4) **Akane A, Matsubara K** and **Shiono H**: Investigation of algorithm for the calculation of probability of paternity likelihood using personal computer program, including the application to parentage testing in the decreased party. Jpn J Legal Med **46**, 254-265, 1992.
5) **Krawczak M** and **Bockel B**: A genetic factor model for the statistical analysis of multilocus DNA fingerprints. Electrophoresis **13**, 10-17, 1992.
6) **Aoki Y, Hashiyada M, Morioka A, et al.**: Spreadsheets of a conventional application software for calculation of plausibility of paternity: Application to parentage testing with highly polymorphic markers in deceased party. Jpn J Legal Med **51**, 196-204, 1997.
7) **Brenner CH**: Symbolic kinship program. [Erratum appears in Genetics 1997;147:398] Genetics **145**, 535-542, 1997.
8) **Egeland T, Mostad PF** and **Olaisen B**: A computerised method for calculating the probability of pedigrees from genetic data. Sci Justice **37**, 269-274, 1997.
9) **Egeland T, Mostad PF, Mevag B, et al.**: Beyond traditional paternity and identification cases. Selecting the most probable pedigree. Forensic Sci Int **110**, 47-59, 2000.
10) **Dawid AP, Mortera J, Pascali VL, et al.**: Probabilistic expert systems for forensic inference from genetic markers. Scand J Statist **29**, 577-595, 2002.
11) **Fung WK**: User-friendly programs for easy calculations in paternity testing and kinship determinations. Forensic Sci Int **136**, 22-34, 2003.
12) **Riancho JA** and **Zarrabeitia MT**: A windows-based software for common paternity and sibling analyses. Forensic Sci Int **135**, 232-234, 2003.
13) **Drabek J**: Validation of software for calculating the likelihood ratio for parentage and kinship. Forensic Sci Int Genet **3**, 112-118, 2009.
14) **Gomes RR, Campos SVA** and **Pena SDJ**: PedExpert: a computer program for the application of Bayesian networks to human paternity testing. Genet Mol Res **8**, 273-283, 2009.
15) **Berent J**: DNAStat, version 2.1--a computer program for processing genetic profile databases and biostatistical calculations. Arch Med Sadowej Kryminol **60**, 118-126, 2010.
16) **Aoki Y, Kato H, Maeno Y, et al.**: Conventional application for calculation of likelihood ratio for identification of human remains by comparison with the DNA of relatives. Res Pract Forens Med **54**, 271-274, 2011.
17) **Haned H**: Forensim: an open-source initiative for the evaluation of statistical methods in forensic genetics. Forensic Sci Int Genet **5**, 265-268, 2011.
18) **Kling D, Egeland T** and **Tillmar AO**: FamLink--a user friendly software for linkage calculations in family genetics. Forensic Sci Int Genet **6**, 616-620, 2012.
19) **Egeland T, Pinto N** and **Vigeland MD**: A general approach to power calculation for relationship testing. Forensic Sci Int Genet **9**, 186-190, 2014.
20) R core team: R: A language and environment for statistical computing. R foundation for statistical computing, Vienna, Austria, 2014. http://www.r-project.org/
21) **Crawley MJ**: Statistics: An introduction using R, Wiley, Hoboken, 2005.
22) **Matloff N**: The art of R programming, No Starch Press, San Francisco, 2011.
23) **Teetor P**: R Cookbook, O'Reilly, Sebastopol, 2011.
24) **Essen-Moller E**: Die Beweiskraft der ähnlichkeit im Vaterschaftsnachweis. Theoretische Glundlagen. Mitt Anth Ges Wien **68**, 9-53, 1938.
25) **Essen-Moller E** and **Quensel CE**: Zur theorie des vateschaftsnachweises auf grund von ähnlichkeitsbefunden. Zeitschr f d ges gerichtl Med **31**, 70-96, 1939.
26) **Gurtler H**: Principles of blood group statistical evaluation of paternity cases at the University Institute of Forensic Medicine, Copenhagen. Acta Med Leg Soc **9**, 83-93, 1956.
27) **Hummel K**: On the theory and practice of Essen-Möller's W value and Gürtler's paternity index (PI). Forensic Sci Int **25**, 1-17, 1984.
28) **Komatsu Y**: Probability of blood type heredity. Acta Crim Japon **10**, 594-600, 1936.
29) **Komatsu Y**: Correction: Probability of blood type heredity. Acta Crim Japon **12**, 890-893, 1938.
30) **Komatsu Y**: Paternity testing with blood types. Acta Crim Japon **13**, 485-494, 1939.
31) **Katsumata Y, Katsumata R, Yamamoto T, et al.**: Estimating probabilities and dealing with mutations in paternity testing--verification of DNA testing with commercially available STR kits. Jpn J Legal Med **55**, 205-216, 2001.
32) **Suzuki K**: Kinship testing. In "Handbook for Dead Body Inspection", 4th edition, ed. by Kondo T and Kinoshita H, pp. 270-280, Kinpodo, Kyoto, 2020.

岩手医誌 73 巻，5 号（令和 3 年 12 月）189-201 頁.

血縁鑑定における R を用いた肯定確率・尤度比算定

髙宮正隆

岩手医科大学医学部，法科学講座法医学分野

(Received on June 30, 2021 & Accepted on July 19, 2021)

要旨

DNA 鑑定は血縁関係の確認，身元の確認を目的に行われ，PCR, 電気泳動などの手法を用いて該当者の各座位の DNA 型を検出する．さらに該当者の DNA 型から肯定確率または尤度比を算定することにより血縁，身元の尤もらしさを評価することがあるが，これら肯定確率・尤度比の算定は鑑定内容によっては計算量が膨大であり，手計算では作業が煩雑になる．一方，近年は計算機の個人使用が普及しており，各人がプログラムを記述することにより大規模な計算を行うことが可能となっている．本綜説では父子鑑定（父 - 子 - 母），父子鑑定（父 - 子），両親鑑定，祖父鑑定，同胞鑑定，半同胞鑑定，叔父鑑定における肯定確率および尤度比の算定を R を用いて記述した．また本綜説で提示しているプログラムは改変が可能で，鑑定への関与人数の増加など鑑定内容の拡張にも容易に対応できると考えられる．