The Korean Communications in Statistics **Vol. 14** No. 3, 2007, pp. 541–549

Statistical Evaluation of Sibling Relationship*

Jae Won Lee[1] Hye-Seung Lee[2] Hyo Jung Lee[3] and Juck-Joon Hwang[4]

Abstract

Testing the sibling relationship becomes more important issue in many cases such as the reunion of dispersed family members whose parents have already passed away and the discrimination between pure-bred and cross-bred dogs. Analysis of the sibling case is different from that of the paternity case. In this paper, we describe how to evaluate and determine the sibling relationship by comparing sibling pairs with unrelated pairs. We use the Korean population with 17 independent STR loci system to propose a discrimination rule.

Keywords: Sibling relationship; likelihood ratio; sensitivity; specificity; error rate.

1. Introduction

Testing the sibling relationship becomes more important issue in many cases. For example, both South and North Korea governments have recently arranged the reunion of dispersed family members who were separated 50 years ago. In many cases, one family member is searching for his or her brother/sister when

* Jae Won Lee was supported by a Korea University Grant. Hyo-Jung Lee was supported by the grant (M10640010002-06N4001-00210) from National R&D program of Ministry of Science and Technology (MOST) and Korea Science and Engineering Foundation (KOSEF).

1) Professor, Department of Statistics, Korea University, 5-1 Anam-Dong, Sungbuk-Gu, Seoul 136-701, Korea.
E-mail : jael@korea.ac.kr

2) Assistant Professor, Pediatrics Epidemiology Center, University of South Florida, 3650 Spectrum Boulevard, Suite 100, Tampa, Florida 33612, USA.
Correspondence : leeh@epi.usf.edu

3) Graduate Student, Department of Statistics, Korea University, 5-1 Anam-Dong, Sungbuk-Gu, Seoul 136-701, Korea.
E-mail : hjunglee@donga.co.kr

4) Professor, Department of Legal Medicine, Korea University, 5-1 Anam-Dong, Sungbuk-Gu, Seoul 136-701, Korea.
E-mail : jjhwang@mail.korea.ac.kr

their parents have already passed away, and the sibling relationship test plays a significant role in distinguishing the true brother/sister from the alleged brother/sister. It can also be used to discriminate between the pure-bred and the cross-bred dogs without knowing their parents' genetic information.

In addition, standard tools for linkage tests of complex genetic diseases and quantitative traits are based on affected sib pairs since the parents of affected individuals are rarely available for typing. However, since Mendelian inconsistencies don't arise with genotypes from only two individuals, false relationships go unnoticed with genotype comparison. Therefore, the accurate knowledge of sibling relationship is critical and the valid inference should be achieved. Thompson (1975) described this for unlinked loci, and Goring and Ott (1995) extended this to linked loci assuming Hardy-Weinberg equilibrium in a Bayesian framework. Boehnke and Cox (1997) discussed likelihood ratio method to infer genetic relationships on the basis of genetic marker data. They have considered the sharing of marker allele IBD considering allele frequencies, marker spacing and genotyping error rate and compared this with likelihood ratio method based on the sharing of marker allele identical by state(IBS)(Chakraborty and Jin 1993a, 1993b; Ehm and Wagner 1996; Strivers et al. 1996).

In this paper, we describe how to evaluate and determine the sibling relationship by comparing sibling pairs with unrelated pairs. We used a likelihood ratio method to evaluate sibling relationship in an unlinked loci system, and this method assesses the sharing of marker allele identical by descent (IBD) given a certain relationship. To determine sibling relationship, we considered error rate based on the sensitivity and the specificity, which is used in the diagnostic evaluation of a discrimination method. We apply this method to Korean population data from the Institute of Legal Medicine at Korea University 17 independent STR loci system and suggest a discrimination rule for sibling relationship in Korean population.

In Materials and Methods section, we describe how to compute the likelihood ratio and determine the sibling relationship based the computed likelihood ratio. In Results section, we propose a new discrimination rule for sibling relationship based on our simulation results from the data in the Institute of Legal Medicine at Korea University and illustrate this discrimination method by a real numerical example. Finally, we discuss how this rule can be affected by the genetic locus system and the relationship between two individuals in Discussion section.

2. Materials and Methods

2.1. Computation of likelihood ratio

Let's consider a testing of sibling relationship. The DNA evidence is evaluated by the setting of the two competing hypotheses, H_0 and H_1.

H_1 : they are in sibling relationship. vs. H_0 : they are unrelated.

Let E and I denote the DNA evidence and non-DNA evidence, respectively, and then the probabilities of the data set under the two hypotheses can be obtained from the odds ratio form of the Bayes' theorem as follows:

$$\frac{P(H_1|E,I)}{P(H_0|E,I)} = \frac{P(E|H_1,I)}{P(E|H_0,I)} \times \frac{P(H_1|I)}{P(H_0|I)}.$$

That is, posterior odds = likelihood ratio (LR) $\times$ prior odds. When we use DNA evidence, while it is posterior odds that the judge intends to know, it is likelihood ratio LR that the forensic scientist can know and intends to evaluate.

Thus, given H_0 and H_1, the DNA evidence, E, is summarized as a likelihood ratio, LR $= P(E|H_1)/P(E|H_0)$. In most cases, E is described by compared persons' genotypes and here, $E = (G_X, G_Y)$ and G_X is the genotype of a person X.

To calculate $p(E|H)$, let $I_{kf}(I_{km})$ be equal to 1 or 0, depending on whether the relative pair shares or fails to share their paternal (maternal) allele at locus k IBD, and let $I_k = (I_{kf}, I_{km})$ and $I = (I_1, \ldots, I_M)$. For a locus k, let's consider $P(I_k = j|H)$, $j = (0,0), (0,1), (1,0)$ and $(1,1)$. Here, $P(I_k = j|H_1) = (1/4, 1/4, 1/4, 1/4)$ and $P(I_k = j|H_0) = (1,0,0,0)$, respectively. Then

$$LR_k = \frac{\sum_{i=1}^{4} P(I_k = i|H_1)P(E_k|I_k = i)}{\sum_{i=1}^{4} P(I_k = i|H_0)P(E_k|I_k = i)}$$

where $P(E_k|I_k = i)$ is the conditional probability of the DNA evidence at locus k, given the IBD status of the pair. At a locus k, the likelihood ratio depending on the genotype combination is given in Table 2.1. When the independent assumption between loci is appropriate, $LR = \prod_{k=1}^{M} LR_k$.

Table 2.1: Likelihood ratio for sibling relationship at a locus k

G_X	G_Y	$P(E\|H_1)$	$P(E\|H_0)$	LR_k
ii	ii	$\frac{p_i^2(1+p_i)^2}{4}$	p_i^4	$\frac{(1+p0_i)^2}{4p_i^2}$
ii	ij	$p_i^2p_j(1+p_i)$	$4p_i^3p_j$	$\frac{(1+p_i)}{4p_i}$
ii	jj	$\frac{p_i^2p_j^2}{2}$	$2p_i^2p_j^2$	$\frac{1}{4}$
ii	jk	$p_i^2p_jp_k$	$4p_i^2p_jp_k$	$\frac{1}{4}$
ij	ij	$\frac{p_ip_j(2p_ip_j+p_i+p_j+1)}{2}$	$4p_i^2p_j^2$	$\frac{(2p_ip_j+p_i+p_j+1}{8p_ip_j}$
ij	ik	$p_ip_jp_k(2p_i+1)$	$8p_i^2p_jp_k$	$\frac{2p_i+1}{8}$
ij	kl	$2p_ip_jp_k$	$8p_ip_jp_kp_l$	$\frac{1}{4}$

i, j, k and l are assumed to be distinct alleles at a single locus.
p_i is a frequency of allele i.
G_X and G_Y are genotype of a person X and a person Y, respectively.
$p(E|H)$ is a probability of genetic evidence given H.
LR_k is likelihood ratio at a locus k.

2.2. Determination of sibling relationship

We considered error rate based on the sensitivity and the specificity used in the diagnostic evaluation of a discrimination method. If we know the real situation under each hypothesis, we would check whether the discrimination is correct or not. That is, if we simulate each real situation from a population data set, we can find the best discrimination method.

The probability of discriminating H_1 given H_1 is called the sensitivity, and the probability of discriminating H_0 given H_0 is called the specificity. Simply, the error rate is the sum of the probability of discriminating H_0 given H_1 and the probability of discriminating H_1 given H_0. Therefore, in this case, the sensitivity means the evaluation for how well the diagnostic method discriminates true sibling pair as sibling pair, and the specificity means the evaluation for how well the diagnostic method discriminates unrelated pair as unrelated pair. That is, if a diagnostic method is good, both the sensitivity and the specificity are high and thus the diagnostic error rate is low. However, since the sensitivity and the specificity trade off each other, two values can't be high at the same time. Therefore,

if we conclude that two persons are in sibling relationship when $\log_e LR > c$, the cut-off point c, showing the lowest error rate, would give the best discrimination. In the next section, we will apply this method to find the best discrimination rule for sibling relationship in Korean population.

3. Results

3.1. Discrimination rule for Korean population

We obtained the distributions of $\log_e LR$ for both true sibling pairs and unrelated pairs, respectively. To do this, we simulated 410 sibling pairs and 11764 unrelated pairs from 232 families of 1164 people in the Institute of Legal Medicine at Korea University. We used the 17 loci system consisted of THO1, TPOX, CSF1PO, FES/FPS, F13A1, ACTBP2, D12S391, GABA plus the 9 loci system (D3S1358, vWA, FGA, D8S1179, D21S11, D18S51, D5S818, D13S317 and D7S820) of Amplifier profiler plus kit (Perkin-Elmer). In this system, the independence assumption between the loci is appropriate (p-value > 0.05) and the cumulative mean exclusion chance(CMEC) is 99.99997%. As shown in Figure 3.1,

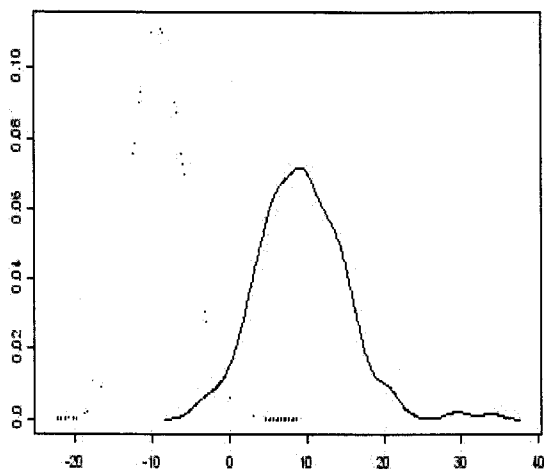

Figure 3.1: $\log_e LR$ for sibling pairs(solid line) and for unrelated pairs(dotted line)

the cut-off point lies in the area where two distributions overlap. We calculated the error rate, sensitivity and specificity at all possible points, and these values are shown in Table 3.1. As shown in the table, all the cut-off values between 1.92 and 2.02 give the lowest error rate. Among these values, we choose the value 1.92, which gives the highest sensitivity. Hence, based on our data, we can declare the sibling relationship if $\log_e LR.1.92$ with total error rate 0.38%(46 out of 12174 total cases), sensitivity 92.68%(380 out of 410 sib pairs) and specificity 99.86%(11748 out of 11764 unrelated pairs).

Table 3.1: Total error rate, sensitivity and specificity for evaluating diagnostic method

CUT	Sensitivity	Specificity	Error
1.70	0.9317073	0.9982149	0.0040250
1.72	0.9317073	0.9982999	0.0039428
1.74	0.9317073	0.9983849	0.0038607
1.76	0.9317073	0.9983849	0.0038607
1.78	0.9317073	0.9983849	0.0038607
1.80	0.9292683	0.9983849	0.0039428
1.82	0.9268293	0.9983849	0.0040250
1.84	0.9268293	0.9983849	0.0040250
1.86	0.9268293	0.9984699	0.0039428
1.88	0.9268293	0.9984699	0.0039428
1.90	0.9268293	0.9985549	0.0038607
1.92	0.9268293	0.9986399	0.0037785
1.94	0.9243902	0.9987249	0.0037785
1.96	0.9243902	0.9987249	0.0037785
1.98	0.9243902	0.9987249	0.0037785
2.00	0.9243902	0.9987249	0.0037785
2.02	0.9243902	0.9987249	0.0037785
2.04	0.9219512	0.9987249	0.0038607
2.06	0.9195122	0.9987249	0.0039428
2.08	0.9195122	0.9987249	0.0039428
2.10	0.9195122	0.9987249	0.0039428
2.12	0.9170732	0.9987249	0.0040250
2.14	0.9146341	0.9987249	0.0041071
2.16	0.9146341	0.9987249	0.0041071
2.18	0.9146341	0.9987249	0.0041071
2.20	0.9146341	0.9988949	0.0039428
2.22	0.9146341	0.9989799	0.0038607
2.24	0.9146341	0.9989799	0.0038607
2.26	0.9146341	0.9989799	0.0038607
2.28	0.9121951	0.9989799	0.0039428

3.2. A numerical example

Now we apply our proposed discrimination rule to a real example. Let us consider a real case in the Institute of Legal medicine at Korea University. This is a case where person 2 claims that person 1 is his brother, and the parents of person 1 died long time ago and left a fortune. To test this brother relationship, the forensic scientist examined the genotypes of the two persons, and the results are shown in Table 3.2. At a locus k, the likelihood ratio LR_k depending on the genotype combination is computed by the formula described in Table 2.1, and we call it brother index. When the independent assumption between loci is appropriate, the cumulative brother index is $LR = \prod_{k=1}^{M} LR_k$. We calculated the cumulative brother index in this way with 17 loci system. In this example, the cumulative brother index is 847354.32, and log (cumulative brother index) = 13.65 which is much greater than 1.92. Therefore, we discriminate two persons are true brother when the total error rate 0.38%, sensitivity 92.68% and specificity 99.86%, are considered.

Table 3.2: Genotypes for the examined persons and brother index

LOCUS	Person1	Person2	Brother Index
D3S1358	15–15	15–15	3.415
VWA	16–17	14–16	0.961
FGA	21–24	22–24	1.033
D8S1179	10–10	10–15	2.427
D21S11	30–32	30–32	23.120
D18S51	12–14	12–15	3.110
D5S818	9–10	9–10	11.092
D13S317	10–11	8–10	1.234
D7S820	12–12	11–12	1.171
THO1	7–9	7–9	2.059
TPOX	8–11	8–11	1.628
CSF1PO	11–11	11–11	8.410
ACTBP2	25.2–29.2	18–29.2	1.982
F13A1	3.2–4	3.2–4	4.422
FES/FPS	11–11	11–12	0.789
D12S391	19–20	15–19	0.814
GABA	13–14	13–15	0.562
Cumulative Brother Index			847354.32
log(Cumulative Brother Index)			13.650

4. Discussion

Testing the sibling relationship would be the only way to determine the paternity without parents' genetic information. Using the likelihood ratio index seems to be fairly reasonable, but there is no unified way to decide the paternity or sibling relationship. In this paper, we used a likelihood ratio method to evaluate sibling relationship in an unlinked loci system, and suggested a discrimination rule for determining the sibling relationship through our simulation study based on the Korean population.

The analysis of a sibling case is quite different from either the paternity case with complete evidence like trio case or the motherless case. It is also different from the deficiency paternity case where the alleged father's genotypes are unavailable. In our previous works, we discussed the various situations for paternity testing and we applied this discrimination method for paternity determination when the alleged father's genetic information is not available (Lee et al., 2001).

It is important that a numeric expression in the DNA identification should be based on the genetic locus system and the relationship of the individuals. We suggested a discrimination rule entirely based on our simulation results from 232 families of 1164 people in the Institute of Legal Medicine at Korea University. Note that, in our simulation, we used 17 loci system when the two persons are not related, and the suggested cut-off point was based on the lowest total error rate and the highest sensitivity. Therefore, the suggested cut-off points in our rule may change when either the different genetic locus system is used or the two persons belong to the same subpopulation. We can easily conjecture that the cut-off point is getting a little bit higher if we use more than 17 loci in our system. We are conducting our simulation studies to investigate how the suggested rule can be affected by the changes in genetic locus system. Therefore, we remind that the readers should be careful to use our discrimination rule if their real case is quite different from our simulation scheme.

References

Boehnke, M. and Cox, N. J. (1997). Accurate inference of relationships in sib-pair linkage studies. *American Journal of Human Genetics,* **61,** 423–429.

Chakraborty, R. and Jin, L. (1993a). Determination of relatedness between individuals using DNA fingerprinting. *Human Biology,* **65,** 875–895.

Chakraborty, R. and Jin, L. (1993b). A unified approach to study hypervariable polymorphisms: statistical considerations of determining relatedness and population distances. In: Pena SDJ, Chakraborty R, Epplin JT, Jeffreys A (eds) DNA fingerprinting: state of the science. Birkhaeuser Verlag, Basel, 153–175.

Ehm, M. G. and Wagner, M. (1996). Test statistic to detect errors in sib-pair relationships. *American Journal of Human Genetics (supplement),* **59,** A217.

Goring, HHH and Ott, J. (1995). Verification of sib relationship without knowledge of parental genotypes. *American Journal of Human Genetics,* **57,** A192.

Lee, J. W., Lee, H. S., Park, M. and Hwang, J. J. (2001). Paternity determination when the alleged father's genotypes are unavailable. *Forensic Science International,* **123,** 202–210.

Strivers, D. N., Zhong, Y., Hanis, C. L., and Chakraborty, R. (1996). RELTYPE: a computer program for determining biological relatedness between individuals based on allele sharing at microsatellite loci. *American Journal of Human Genetics (supplement),* **59,** A190.

Thompson, E. A. (1975). The estimation of pairwise relationships. *Annals of Human Genetics,* **39,** 173–188.

[Received March 2007, Accepted August 2007]