Korean J Leg Med 2019;43:97-105
https://doi.org/10.7580/kjlm.2019.43.3.97

원 저

STR 자료의 데이터마이닝을 이용한 혈연관계의 분류

정수진[1] · 이효정[2] · 이숭덕[3]
이승환[4] · 박수정[4] · 김종식[4]
이재원[1]

[1]고려대학교 통계학과
[2]동아에스티 개발본부
[3]서울대학교 의과대학 법의학과
[4]대검찰청 과학수사2과

Received: May 20, 2019
Revised: July 26, 2019
Accepted: August 29, 2019

Correspondence to
Jae Won Lee
Department of Statistics, Korea University, 145 Anam-ro, Seongbuk-gu, Seoul 02841, Korea
Tel: +82-2-3290-2237
Fax: +82-2-924-9895
E-mail: jael@korea.ac.kr

Classification of Common Relationships Based on Short Tandem Repeat Profiles Using Data Mining

Su Jin Jeong[1], Hyo Jung Lee[2], Soong Deok Lee[3], Seung Hwan Lee[4], Su Jeong Park[4], Jong Sik Kim[4], Jae Won Lee[1]

[1]Department of Statistics, Korea University, Seoul, Korea, [2]Product Development HQ, Dong-A ST, Seoul, Korea, [3]Department of Forensic Medicine, Seoul National University College of Medicine, Seoul, Korea, [4]Forensic Science Division 2, Supreme Prosecutor's Office, Seoul, Korea

We reviewed past studies on the identification of familial relationships using 22 short tandem repeat markers. As a result, we can obtain a high discrimination power and a relatively accurate cut-off value in parent-child and full sibling relationships. However, in the case of pairs of uncle-nephew or cousin, we found a limit of low discrimination power of the likelihood ratio (LR) method. Therefore, we compare the LR ranking method and data mining techniques (e.g., logistic regression, linear discriminant analysis, diagonal linear discriminant analysis, diagonal quadratic discriminant analysis, K-nearest neighbor, classification and regression trees, support vector machines, random forest [RF], and penalized multivariate analysis) that can be applied to identify familial relationships, and provide a guideline for choosing the most appropriate model under a given situation. RF, one of the data mining techniques, was found to be more accurate than other methods. The accuracy of RF is 99.99% for parent-child, 99.44% for full siblings, 90.34% for uncle-nephew, and 79.69% for first cousins.

Key Words: Short tandem repeats; Kinship testing; Relationships; Likelihood ratio; Data mining

서 론

정보기술의 발달로 인해 이공계뿐 아니라 인문사회학을 비롯한 다양한 분야에서 이를 응용한 연구가 활발히 진행되고 있다. 특히 기존에 잘 알려지지 않았던 범죄수사 분야에서도 과학수사를 통해 범인을 검거하는 사례가 점점 늘고 있다. 지금까지 용의자를 가장 잘 특징지을 수 있는 현장 증거로는 지문이 대표적이었다. 하지만 과학기술이 점점 발전함에 따라 이제는 혈흔이나 담배꽁초와 같은 증거물에서 유전자 정보(DNA)를 추출할 수 있게 되었고, 이러한 정보는 기존의 지문과 마찬가지로 개개인별로 독특한 특성을 가지고 있기 때문에 용의자를 확인할 수 있게 되었다. 이러한 유전자는 개인식

pISSN 2383-5702
eISSN 2383-5710

별뿐 아니라 멘델(Mendel)의 유전 법칙에 따라 부모에게서 각각 반반씩 자식에게로 유전되기 때문에 혈육의 정보도 함께 확인할 수도 있게 되었다[1]. 따라서 현장에 떨어져 있는 작은 혈흔 자국 하나만으로도 범인이 누구인지 추적할 수도 있게 되었으며, 부모¨자식 관계가 맞는지 혈연을 확인할 수도 있으며, 더 나아가 범인이 어느 민족인지까지도 알아 낼 수 있게 되었다[2]. 또한 이러한 범죄자의 유전자 정보를 저장해 둔 '범죄자 데이터베이스 구축'을 통해 용의자의 유전자 감식 결과와 데이터베이스 내의 정보를 매칭(matching)하여 해당 유전자 정보와 일치하는 범인을 찾아내는 방법도 등장하였다[3].

데이터베이스를 통해 용의자 개인을 식별할 수 있으며, 이보다 또 더 의미 있는 부분은 만약 유전자 증거와 일치하는 유전자 프로필(DNA profile)을 찾지 못했을 때, 혈연관계에 있는 가족들을 조사하는 이른바 '혈통 유전자 추적(familial searching)' 기법을 응용하여 용의자를 식별할 수 있게 되었다. 이러한 최첨단 유전자 수사기법을 적용한 대표적인 사건으로는 25년 간 미국 LA를 공포로 몰아넣었던 '음침한 수면자(Grim Sleeper)' 사건이 있었다[3]. 이 사건은 유전자 데이터베이스 검색을 통해 다른 혐의로 체포된 젊은 남성 용의자의 유전자와 연쇄살인 사건의 범죄 현장에서 채취된 유전자 증거에 대한 혈연관계를 정밀 분석함으로써 이들끼리의 혈연관계가 있다는 사실을 통해 용의자의 아버지가 연쇄살인범으로 밝혀진 사건이다[3]. 이와 같이 혈통 유전자 추적 기법은 지금까지 다양한 사건에 적용되어 성공적으로 용의자를 검거하면서부터 이에 대한 효용성이 입증되고[4] 많은 국가에서 좀 더 많은 수사에 이를 적용하기 위한 연구가 늘고 있지만, 아직까지는 그 식별 범위에 있어서 한계점이 있으며 또한 계량화하는 기준이 모호하여 그 적용이 어려운 실정이다[5,6].

따라서 이를 해결하기 위해 본 연구에서는 혈연관계 식별을 위해 적용할 수 있는 데이터마이닝 기법들을 알아보고, 1촌 - 4촌까지의 혈연식별에 적당한 모델을 적용하여 특히 3촌 이상의 혈연관계 식별에서 해결하지 못한 부분에 대해 식별력을 높일 수 있는 통계적 모델을 확립하고자 한다.

재료 및 방법

1. 자료 설명

본 연구를 위해 서울대학교 법의학교실과 질병관리본부 국립중앙인체자원은행으로부터 분양받은 896명(118가구)을 대상으로 혈연관계 분석을 실시하였다. 해당 자료의 가족 구성원의 경우, 1촌(parents-child) 혈연관계는 778쌍, 2촌(full siblings) 혈연관계는 522쌍, 3촌(uncle-nephew) 혈연관계는 859쌍, 4촌(first cousins) 혈연관계는 468쌍으로 구성되었다. 또한 혈연관계가 없는 경우로는 수집된 가계도 가운데 서로 혈연관계가 없는 사람들의 조합으로 총 1,000쌍을 짝지어 구성하였다. 해당 유전자 분석을 위한 실험은 서울대학교 법의학과에서 실시하였으며, 유전자 자료에는 대립형질(allele)과 유전자형(genotype) 빈도에 관한 정보가 모두 포함되어 있다.

분석에 사용된 short tandem repeat (STR) 마커는 미국 연방수사국에서 사용하고 있는 CODIS 13 마커(D3S1358, D5S818, D7S820, D8S1179, D13S317, D16S539, D18S51, D21S11, CSF1PO, FGA, TH01, TPOX, vWA)에 법유전학 분석에서 흔히 사용하는 9개 마커(D2S1338, D19S433, Penta_E, Penta_D, D1S1656, D2S441, D10S1248, D12S391, D22S1045)를 포함하여 총 22개였다[7-10].

2. 혈연관계 평가 방법

(1) 공유 대립형질 수에 기초하여 우선순위를 부여하는 방법

공유 대립형질 수에 기초하여 우선순위를 부여하는 방법은 Evett와 Weir (1998) [7]가 제시한 가장 기본적인 혈통유전자 추정 방법에 해당하며, 유전자 증거의 프로파일과 데이터베이스의 각 유전자 프로파일을 비교하여 서로 공유하는 대립유전자의 수를 계산하고, 이 수에 기초하여 혈연관계의 가능성이 있는 후보 가족에 대해 우선순위를 부여하는 방법(matching allele counting ranking method)이다[11]. 이 방법은 적용하기 쉽고 계산이 간단하지만, 동일 순위의 많은 가족들이 존재하게 되어 추가적인 조사를 필요로 하게 되고, 또한 비교적 적은 대립유전자를 공유하는 실제의 혈연 가족을 배제시킬 가능성이 있다[11].

(2) 우도비에 기초한 우선순위 부여 방법(likelihood ratio ranking method)

우도비(likelihood ratio, LR)에 기초한 우선순위 부여 방법(LR ranking method)은 현재까지 가장 많이 사용되는 방법으로 이전의 공유 대립형질의 수에 기초한 방법의 제한점을 해결할 수 있으며, Bieber 등(2006) [2]와 Cowen과 Thomson (2008) [11]의 연구에서 다른 방법들보다 효과적인 식별 방법으로 알려져 있는 방법이다[2,11]. 친자확인과 같은 유전자 감식에 사용되는 유전자는 전체 유전서열 중 일부 반복되는 염기서열 부위이다. 보통 이러한 염기서열을 유전자 감식에서는 마커라고 부른다. 이는 데이터베이스의 각 유전자 프로파일에 대해 유전자 증거(E)와 정황증거(I)를 고려하였을 때, 혈연관계인 경우 유전자 프로파일이 관측될 확률 $P(H_1 \mid E, \mathrm{I})$와 혈연관계가 없는 경우 유전자 프로파일이 관측될 확률 $P(H_0 \mid E, \mathrm{I})$의 비인 LR $\frac{P(E \mid H_1, I)}{P(E \mid H_0, I)}$ 를 계산하고, 각

https://doi.org/10.7580/kjlm.2019.43.3.97

LR 값에 대해 정렬하고 순위를 부여하는 방법이다[2,12].

(3) 로지스틱 회귀모형(logistic regression)

로지스틱 회귀모형은 일반적인 선형 모델(generalized linear model)의 특수한 경우로 독립 변수의 선형 결합을 이용하여 사건의 발생 가능성(odds ratio)을 예측하는 목적으로 사용되는 통계 방법이다[13]. 이 때, 오즈(odds)는 $p(y = 1 \mid x)$을 $p(y = 0 \mid x)$으로 나눠준 값이며, 로지스틱 함수는 $p(y = 1 \mid x) = \frac{e^{\beta x}}{1+e^{\beta x}}$ 이다[13,14]. 자료의 정규성이나 두 그룹의 공분산구조가 동일하다는 가정이 필요하지 않으며, 전통적인 선형판별분석(linear discriminant analysis, LDA)에 비해 보다 일반적인 경우에 적용할 수 있다.

(4) 데이터마이닝 방법

본 연구에서 수행될 혈연관계 평가 분석을 통해, 3촌 이상에서의 혈연관계 평가는 비혈연관계 집단과 혈연관계 집단과의 분포와 서로 겹쳐 기존에 대다수에서 사용하고 있는 분석방법으로 비교하는 것은 쉽게 혈연관계를 구분하기가 어려운 것이 현실이다. 따라서 일반적으로 이러한 3촌 이상의 혈연관계 평가를 위해서는 분류분석을 위한 다양한 통계학적 방법을 적용하여야 한다.

분석에 적용될 수 있는 통계학적 방법으로는 LDA [15], 대각선형판별분석(diagonal linear discriminant analysis, DLDA) [16], 대각이차형판별분석(diagonal quadratic discriminant analysis, DQDA) [16] 등의 모수적 방법과 서포트벡터머신(support vector machines, SVM) [17], 분류회귀나무(classification and regression trees, CART) [18]의 비모수적 방법, 그리고 마지막으로 K-최근접 이웃(K-nearest neighbor, KNN) [19], 랜덤 포레스트(random forest, RF) [20], 벌점다변량분석(penalized multivariate analysis, PMA) [21] 등의 앙상블 방법이 있다.

1) 선형판별분석(linear discriminant analysis, LDA)

가능한 클래스간의 정보를 최대한 유지하면서 클래스 분리를 최대화하며 차원을 축소하는 것이 목적이다. P 차원의 표본 데이터 집합 $X=\{x^{(1)}, x^{(2)},..., x^{(N)}\}$이 주어졌을 때, $\grave{u}_1$ 클래스에 속하는 것이 N_1개이고, $\grave{u}_2$ 클래스에 속하는 것이 N_2개일 때, x를 임의의 선을 따라서 사영하여 스칼라 $y = W^T x$가 되며, 가능한 모든 선들 가운데 스칼라 값들의 분리를 최대화시키는 것으로 선택되어진다[15]. 이 때, 이러한 분리는 클래스간 분산(between-class scatter)과 클래스 내 분산(within-class scatter)의 비율을 최대화시키는 것으로 선별한다[15]. 즉,

$|\tilde{\mu}_1-\tilde{\mu}_2| = |W^T(\mu_1 - \mu_2)|$, $\tilde{S}_j^2 = \sum_{y\in w_j}(y - \tilde{\mu}_j)^2 = \sum_{y\in w_j}(y - \tilde{\mu}_j)(y - \tilde{\mu}_j)^T$

$J(W) = \frac{|\tilde{\mu}_1-\tilde{\mu}_2|^2}{\tilde{S}_1^2-\tilde{S}_2^2}$ 을 최대화시키는 W를 찾는 것이다[15].

2) 서포트벡터머신(support vector machine, SVM)

SVM은 러시아 과학자 Vapnik (1979)에 제안하였으며, 2000년대 들어 분류와 회귀 분석을 적용한 패턴 인식과 기계학습 분야에서 크게 활용되었다[17]. 선형 SVM의 경우, 주어진 데이터의 집합이 있는 상황에서 새로운 데이터가 어느 그룹에 속할지 판단하는 비확률적 이진 선형 분류 모델에 해당한다[17].

$D = \{(x_j, y_j) \mid X_j \in R^P, y_j \in \{-1, 1\}\}_{j=1}^n$

$\arg \| \omega \|, y_j(\omega \cdot x_j - b) \geq 1, \text{for all } 1 \leq j \leq n$

x_j는 p차원의 실수 벡터, y_j는 x_j가 포함되어 있는 클래스 결과(−1, 1)이며, 이 때, · 은 내적 연선자이며, w는 초평면의 정규벡터(normal vector)이며, 마진을 최대값으로 하는 조건으로 분석한다[17].

3) 분류회귀나무(classification and regression trees, CART)

Breiman 등(1984)은 이진 결정트리 구성 기법인 CART를 제안하였다[18]. CART의 구축과정은 분할기준의 선택과 현재 노드의 분할여부를 판단하고, 터미널 노드로 자료를 재대입하여 오차를 추정하는 것으로 이루어져있다. 현재의 노드에서 어떤 분리변수가 선택될 것인지가 결정되며 선택된 노드에서 불순도를 정의하여 현재 노드와 자식노드간의 불순도의 값을 최대폭으로 감소시키는 분리변수를 선택한다. 불순도 함수로 엔트로피나 지니 계수(gini index)를 이용하며 주어진 자료를 가장 잘 구분할 수 있는 연산을 통해 결정 트리를 반환한다[18].

4) K−최근접 이웃(K−nearest neighbor, KNN)

기계 학습 알고리즘에 있어서 KNN은 분류나 회귀에 사용되는 비모수 방식의 일종이다[19]. 특징 공간 내의 k개의 가장 가까운 경우에 클래스를 선택할 수 있게 훈련받는 것이다[19]. KNN은 압축 근접 이웃(condensed nearest neighbor, CNN)을 이용하여 데이터 집합을 축소한다[19]. 이는 훈련 데이터(training data)를 통해 프로토 타입의 집합을 선택하고 이를 통해 첫번째 최근접 이웃(one-nearest neighbor, 1NN)을 분류하며, 이러한 과정의 CNN을 계속 반복해서 수행한다[19].

$a(x) = \| x' - y \| / \| x - y \|$

이 때, $\| x - y \|$는 x와 가장 가까운 y까지의 거리를 나타내며, $\| x' - y \|$는 y와 가장 가까운 x'까지의 거리를 나타낸다[19].

5) 랜덤 포레스트(random forest, RF)

기계학습에 있어서 앙상블 학습 방법의 일종인 RF는 훈련 과정에서 다수의 결정트리로부터 분류 또는 회귀분석 결과를 통해 정의되는 학습 방법이다[20]. 이는 훈련단계에 있어서 트리의 노드로 들어오는 데이터들이 최적의 포인트로 분리되기 위한 정보 획득량(information gain)을 측정 기준으로 삼아 각각의 신뢰도(confidence)를 최대화하는 것을 의미한다[20].

$$I = H(S) - \sum_{j\in\{L,\,R\}} \frac{|S^j|}{S} H(S^j),\ H(S) = -\sum_{c\in C} p\,(c)\log\,(p\,(c))$$

여기서 C는 전체 클래스 집합을 나타내는, p(c)는 각 부류에 대한 확률 질량 함수이다[20].

6) 벌점다변량분석(penalized multivariate analysis, PMA)

PMA는 특이값 분해(singular value decomposition)를 통해 원자료 근사 시에 라소 벌점(Lasso penalty)을 부과하는 벌점 행렬 분해(penalized matrix decomposition) 방법을 구현할 수 있는 방법을 의미한다[21].

$$\hat{X} = \sum_{a=1}^{a} d_a u_a v_a^T$$

이 때, u_a, v_a는 R^n, R^p의 단위 벡터(unit vector)이며, d_a는 음수가 아닌 상수(non-negative constant)이며, X는 벌점 행렬 분해이다[21].

3. 혈연관계 평가 기준

분석 결과는 민감도(sensitivity)와 특이도(specificity)를 이용하여 평가하였다. 민감도는 실제로 혈연관계에 있는 두 사람에 대해 분석 결과에서도 혈연관계가 있는 것으로 판단할 확률을 의미하며, 특이도는 실제로 혈연관계에 없는 두 사람에 대해 분석 결과에서도 혈연관계가 없다고 판단할 확률을 의미한다. 또한 혈연관계 평가에서 정확도(accuracy)는 전체 중에 정확하게 분류된 개수로 나눈 값으로 계산되었다. 따라서 민감도와 특이도가 가장 높은 결과를 나타내는 지점을 절단값(cut-off value)으로 정의하였다[5].

또한 데이터 마이닝 기법에 사용된 분석은 기본적으로 무작위배정(randomization)을 통해 훈련집단(training set) 및 평가집단(testing set)에 사용될 데이터를 1:1로 임의 추출하여 훈련집단으로 해당 분석을 통해 모델링을 하고 나머지 평가집단을 평가하는 목적으로 수행하였다. 해당 과정은 100번 반복하였고, 모든 결과 평가는 100번 반복된 결과의 평균으로 제시하였다.

4. 분석 프로그램

해당 분석은 SAS 9.4 (SAS Institute Inc., Cary, NC, USA)과 R3.5.3 (https://cran.r-project.org) 통계 프로그램으로 분석하였으며, 해당 통계 분석을 위한 R package로는 다음과 같다. 로지스틱 회귀분석은 'glm', LDA는 'MASS'와 'lda', DLDA, DQDA는 'WGCNA', SVM은 'e1071', CART는 'caret'와 'tree', KNN은 'class', RF는 'randomForest'와 'varSelRF', PAM은 'pamr'을 적용하여 분석하였다.

결 과

1. 대립형질 수에 기초한 우선순위 부여 방법

공유 대립형질 수에 기초하여 우선순위를 부여하는 방법으로

Table 1. TNSA for common relationships

Familial relationship	Parents-child (n=778)	Full siblings (n=522)	Uncle-nephew (n=859)	First cousins (n=468)	Unrelated (n=1,000)
TNSA					
Mean±SD	26.66±1.89	27.55±4.17	19.45±3.51	17.16±3.16	14.55±2.75
Range	22−33	9−44	7−29	7−30	6−24
No. of shared alleles=0					
Mean±SD	0.06±0.27	2.68±1.93	5.56±2.62	7.28±2.34	9.26±2.20
Range	0−2	0 - 13	0−16	1−16	3−17
No. of shared alleles=1					
Mean±SD	17.21±1.91	11.08±2.65	13.43±2.60	12.27±2.32	10.93±2.29
Range	11−22	0−18	4−20	5−19	4−18
No. of shared alleles=2					
Mean±SD	4.72±1.88	8.24±2.91	3.01±1.63	2.44±1.48	1.81±1.25
Range	0−11	0−22	0−8	0−9	0−8

TNSA, total number of shared alleles; SD, standard deviation.

 https://doi.org/10.7580/kjlm.2019.43.3.97

는 혈연관계에 있는 두 대상이 서로 공유하고 있는 대립형질의 수를 계산하여 혈연관계의 가능성이 있는 후보 가족을 선정하고 더 나아가 혈연관계를 식별할 수 있는 적절한 절단값을 구하는 것을 목적으로 하고 있다.

실제 혈연관계별(1 - 4촌)로 22개의 STR 마커의 44개의 대립형질 가운데 서로 공유 대립형질 수(total number of shared alleles, TNSA)를 계산한 결과, parents-child (26.66±1.89개), full siblings (27.55±4.17개), uncle-nephew (19.45±3.51개), first cousins (17.16±3.16개)인 경우로 1촌에서 4촌으로 갈수록 평균 TNSA는 점점 낮아지지만 비혈연관계(14.55±2.75개)인 경우에 비해서는 높게 나타났다 (Table 1).

또한 공유개수별로 구분하여 비교한 결과 parents-child (1촌)에서는 유전자 쌍 가운데 하나도 공유하지 않는 STR 마커 개수(numbers of shared alleles=0)는 0.06±0.27개, 한 개 공유한 마커 개수(numbers of shared alleles=1)는 17.21±1.91, 두 개 모두 공유한 마커 개수(numbers of shared alleles=2)는 4.72±1.88이다. 또한 full siblings (2촌)의 경우, 하나도 공유하지 않는 개수는 2.68±1.93개, 한 개 공유한 개수는 11.08±2.65개, 두 개 모두 공유한 개수는 8.24±2.91개이다. Uncle-nephew (3촌)의 경우, 하나도 공유하지 않는 개수는 5.56±2.62개, 한 개 공유한 개수는 13.43±2.60개, 두 개 모두 공유한 개수는 3.01±1.63개이다. First cousins (4촌)의 경우, 하나도 공유하지 않는 개수는 7.28±2.34개, 한 개 공유한 개수는 12.27±2.32개, 두 개 모두 공유한 개수는 2.44±1.48개이다. 마지막으로 비혈연관계에서 하나도 공유하지 않는 개수는 9.26±2.20개, 한 개 공유한 개수는 10.93±2.29개, 두 개 모두 공유한 개수는 1.81±1.25개이다. 이 때, 유전자 쌍 가운데 하나도 공유하지 않는 STR 마커 평균 개수는 1촌부터 4촌으로 갈수록 점점 증가(0→7.28)하며, 비혈연관계(9.26)에서 다른 혈연관계에 비해 높게 나타났다(Table 1).

이러한 결과를 바탕으로 공유 대립형질 수를 0에서부터 44까지 하나씩 늘려가면서 혈연관계 분류에 있어 민감도와 특이도가 가장 높은 공유 대립형질 수의 절단값을 계산한 결과, parents-child은 23개(정확도, 99.71%), full siblings은 22개(정확도, 97.63%), uncle-nephew은 18개(정확도, 80.63%), first cousins은 16개(정확도, 66.83%)에 해당된다(Table 2).

2. LR에 기초한 우선순위 부여 방법

LR에 기초한 방법은 기존의 대립형질의 수에 기초한 방법보다 혈연관계의 식별 확률이 높아지며, 또 드물게 나타나는 대립형질(rare allele)에 대해서도 설명이 가능할 수 있다. 혈연관계 평가에서 계산된 LR이 '혈연관계 지수(Relativeness Index)'이며, 두 개체의 유전자 쌍이 관측될 결합 확률과 혈연관계가 없는 경우에서 유전자 쌍이 관측될 결합 확률의 비로 계산된다[2,11,13,22].

그 결과, parents-child의 경우, 로그 우도비(Log10LR)는 7.34±1.47으로 비혈연관계였을 때 로그 우도비인 2.15±0.93보다 높게 나타났으며(Table 3), 이 때 비혈연관계와 구분하는 데 있어 민감도와 특이도가 최대가 되는 부분인 로그 우도비의 절단값은 4.43(정확도, 98.76%)이다(Table 4). 단, 이 때 일반적으로 1촌 혈연관계 판정의 경우, 우선적으로는 모든 STR 마커에서 identity by descent (IBD)가 성립하여

Table 2. Classification of common relationships according to TNSA

Familial relationship	Cut-off value of TNSA	NF	NU	Sensitivity (%)	Specificity (%)	Accuracy (%)
Parents-child	23	776/778	997/1,000	99.74	99.70	99.71
Full siblings	22	492/522	994/1,000	94.25	99.40	97.63
Uncle-nephew	18	637/859	862/1,000	74.16	86.20	80.63
First cousins	16	337/468	644/1,000	72.01	64.40	66.83

TNSA, total number of shared alleles; NF, number of family relationships (true/total); NU, number of unrelated (true/total).

Table 3. LR for common relationships

$Log_{10}LR$	Parents-child		Full siblings		Uncle-nephew		First cousins	
	Relative	Unrelative	Relative	Unrelative	Relative	Unrelative	Relative	Unrelative
Mean	7.34	2.15	5.84	−4.76	1.08	−1.42	0.31	−0.35
SD	1.47	0.93	3.43	1.79	1.70	1.07	0.71	0.54
Min	3.77	−0.47	−8.24	−10.10	−4.20	−4.71	−1.84	−1.86
Max	12.33	5.53	20.62	2.01	6.24	2.36	3.41	1.73

LR, likelihood ratio; SD, standard deviation; Min, minimum; Max, maximum.

야 하지만 유전자 돌연변이 등의 이유로 비록 혈연관계임에도 불구하고 모든 STR 마커에서 IBD가 성립하지 않는 경우도 있으며, 아직까지 한국인의 정확한 돌연변이률을 확정할 수 없으므로 보수적인 방법으로 유전자 좌위가 맞지 않는 경우 해당 마커의 paternity index (PI)=0으로 두고 계산하였다[6]. Full siblings의 경우, 로그 우도비는 5.84±3.43으로 비혈연관계였을 때 로그 우도비인 −4.76±1.79보다 높게 나타났으며(Table 3), 이 때, 로그 우도비의 절단값은 0.15 (정확도, 98.29%)이다(Table 4). Uncle-nephew의 경우, 로그 우도비는 1.08±1.70으로 비혈연관계였을 때 로그 우도비인 −1.42±1.07보다 높게 나타났으며(Table 3), 이 때 로그 우도비의 절단값은 −0.03 (정확도, 84.19%)이다(Table 4). First cousins의 경우, 로그 우도비는 0.31±0.71로 비혈연관계였을 때 로그 우도비인 −0.35±0.54보다 높게 나타났으며(Table 3), 이 때 로그 우도비의 절단값은 −0.11 (정확도, 70.57%)이다(Table 4).

이러한 LR을 기준으로 하는 분류 방법은 1촌에서의 민감도 98.59%, 특이도 98.90%, 2촌에서의 민감도 95.98%, 특이도 99.50%로 높은 민감도, 특이도를 나타났지만 상대적으로 3촌(민감도 76.72%, 특이도 90.60%), 4촌(민감도 73.93%, 특이도 69.00%)에서는 다소 낮은 민감도, 특이도가 나타났다(Table 4).

3. 로지스틱 회귀분석

지금까지 분석한 공유 대립형질 수, LR 또는 공유 대립형질 수와 LR을 한꺼번에 고려한 방법들은 단순히 절단값을 기준으로 정확도를 확인하는 방법이다. 이러한 방법은 추출된 표본 집단에 따라 다르게 나올 수 있으므로 이에 대한 일반화를 위해서는 더 고차원적 분석이 필요한 실정이다. 그 방법으로는 우선 로지스틱 회귀분석을 통해 혈연관계 여부를 추정할 수 있다. 로지스틱 회귀분석은 독립 변수의 선형 결합을 이용하여 오즈비(odds ratio)을 예측하는 목적으로 사용되는 통계 방법이다. 이는 TNSA, 우도비(Log LR)에 따라 혈연관계에 포함될 확률과 포함되지 않을 확률을 계산한 후 혈연관계 여부를 분류하게 된다(Table 5).

그 결과, parents-child의 경우 로지스틱 회귀분석으로 추정한 결과 정확도가 99.94%로 기존의 공유 대립형질 개수(99.71%), LR 방법(98.76%)보다 높게 나타났다. 이는 민감도 기준으로도 마찬가지로 로지스틱 회귀분석 추정 결과는 100.00%로 기존의 공유 대립형질 개수(99.71%), LR 방법(98.76%)보다 높게 나타났다(Table 5). Full siblings의 경우 정확도는 98.62%로 기존의 공유 대립형질 개수(97.63%), LR 방법(98.29%)보다 높게 나타났다. 이는 민감도 기준으로도 마찬가지로 로지스틱 회귀분석 추정 결과는 97.89%로 기존의 공유 대립형질 개수(94.25%), LR 방법(95.98%)보다 높게 나타났다(Table 5). Uncle-nephew의 경우 정확도는 84.99%로 기존의 공유 대립형질 개수(80.63%), LR 방법(84.19%)보다 높게 나타났다. 이는 민감도 기준으로도 마찬가지로 로지스틱 회귀분석 추정 결과는 86.96%로 기존의 공유 대립형질 개수(74.16%), LR 방법(76.72%)보다 높게 나타났다(Table 5). 마지막으로 first cousins의 경우 정확도는 75.48%로 기존의 공유 대립형질 개수(66.83%), LR 방법(70.57%)보다 낮게 나타났다. 이는 민감도 기준으로도 마찬가지로 로지스틱 회귀분석 추정 결과는 66.67%로 기존의 공유 대립형질 개수(72.01%), LR 방법(73.93%)보다는 낮게 나

Table 4. Classification of common relationships according to Log LR

Familial relationship	Cut-off value of $Log_{10}LR$	NF	NU	Sensitivity (%)	Specificity (%)	Accuracy (%)
Parents-child	4.43	767/778	989/1,000	98.59	98.90	98.76
Full siblings	0.15	501/522	995/1,000	95.98	99.50	98.29
Uncle-nephew	−0.03	659/859	906/1,000	76.72	90.60	84.19
First cousins	−0.11	346/468	690/1,000	73.93	69.00	70.57

LR, likelihood ratio; NF, number of family relationships (true/total); NU, number of unrelated (true/total).

Table 5. Logistic regression of TNSA and Log LR for common relationships

Familial relationship	Cut-off value of probability score	NF	NU	Sensitivity (%)	Specificity (%)	Accuracy (%)
Parents-child	0.34	778/778	999/1,000	100.00	99.90	99.94
Full siblings	0.38	511/522	990/1,000	97.89	99.00	98.62
Uncle-nephew	0.42	747/859	833/1,000	86.96	83.30	84.99
First cousins	0.38	312/468	796/1,000	66.67	79.60	75.48

TNSA, total number of shared alleles; LR, likelihood ratio; NF, number of family relationships (true/total); NU, number of unrelated (true/total).

https://doi.org/10.7580/kjlm.2019.43.3.97

타났다(Table 5).

4. 데이터마이닝 기법

분류분석 방법으로는 LDA, DLDA, DQDA, KNN, CART, SVM, RF, PMA를 적용한 혈연관계별 분류 방법을 고려하였다(Table 6).

그 결과, parents-child, full siblings, uncle-nephew, first cousins 모든 경우에, 민감도와 특이도가 가장 높은 경우는 다른 데이터마이닝 기법에 비해 RF 방법이다. RF 방법을 통한 parents-child 관계 분류로는 민감도 99.99%, 특이도 99.98%, 정확도 99.99%였으며, full siblings 관계 분류로는 민감도 99.11%, 특이도 99.62%, 정확도 99.44%이다. 또한 uncle-nephew 관계 분류로는 민감도 89.04%, 특이도 91.54%, 정확도 90.34%이며, 마지막으로 first cousins 관계 분류로는 민감도 61.41%, 특이도 88.25%, 정확도 79.69%이다(Table 6).

고 찰

지금까지 혈연관계 식별에 주로 사용되었던 방법은 특정 염기서열의 반복으로 이루어진 STR 마커를 통한 LR의 계산으로 비혈연관계와 혈연관계를 식별할 수 있는 절단값(cut-off value)을 확인하였다. Jeong 등(2016) [6]에 따르면 로그 우도비의 최소값인 3.77 이상, 2촌(full siblings)은 로그 우도비 1.94 이상이었을 때 오류율을 최소로 하는 절단값으로 보고하였다. 하지만 3촌(uncle-nephew) 이상의 혈연관계의 경우, 비혈연관계에 있는 사람들과 로그 우도비의 분포가 상당부분 겹쳐져 기존의 방법으로는 정확하게 구분할 수 있는 절단값을 계산하는데 한계점도 함께 제시하였다[6].

따라서 본 연구는 혈연식별을 위한 22개의 STR 마커를 기준으로 다양한 방법들을 적용하여 혈연관계를 분류하는데 있어 효과적인 모형이 무엇인지를 확인하고자 한다. 해당 방법으로는 가장 간단한 대립형질 수나 LR을 기초로 하여 절단값을 정하고 분류하는 것을 기본으로 더 확장하여 대립형질 수나 LR을 통한 로지스틱 회귀분석 방법과 함께 다양한 데이터마이닝 방법(LDA, DLDA, DQDA, KNN, CART, SVM, RF, PMA)에 대한 분석을 적용하였다.

TNSA를 이용한 분류 방법은 과거 우도비(Log LR)와 함께 혈연 분류를 위해 사용되었던 방법이며 다양한 자료에 적용하기 쉽고 계산이 간단하지만, 3촌 이상 혈연관계의 경우 비교적 적은 대립형질을 공유하게 되므로 실제의 혈연 가

Table 6. Classification of common relationships according to various classification methods

Method	Parents-child			Full siblings			Uncle-nephew			First cousins		
	Sen (%)	Spe (%)	Acc (%)	Sen (%)	Spe (%)	Acc (%)	Sen (%)	Spe (%)	Acc (%)	Sen (%)	Spe (%)	Acc (%)
LDL	99.90	99.61	99.74	95.48	99.45	98.08	84.31	84.87	84.57	68.65	73.47	71.94
DLDA	100.00	99.35	99.63	96.93	99.23	98.44	83.65	84.91	84.28	71.79	71.90	71.87
DQDA	97.38	99.66	98.66	93.72	96.63	95.63	54.93	94.15	75.99	36.24	89.87	72.77
KNN	99.66	99.16	99.38	94.20	98.98	97.34	76.73	89.01	83.29	45.32	82.06	70.35
CART	99.97	99.98	99.98	99.16	99.43	99.34	88.54	90.86	89.74	52.63	88.91	77.35
SVM	100.00	99.18	99.54	97.79	98.55	98.29	85.83	88.25	87.09	51.00	90.08	77.62
RF	99.99	99.98	99.99	99.11	99.62	99.44	89.04	91.54	90.34	61.41	88.25	79.69
PMA	99.98	99.80	99.88	92.18	99.89	97.25	82.11	85.07	83.66	54.08	83.20	73.92

Sen, sensitivity; Spe, specificity; Acc, accuracy; LDL, linear discriminant analysis; DLDA, diagonal linear discriminant analysis; DQDA, diagonal quadratic discriminant analysis; KNN, K-nearest neighbor; CART, classification and regression trees; SVM, support vector machines; RF, random forest; PMA, penalized multivariate analysis.

Table 7. Summary classification of common relationships according to various classification methods

Method	Parents-child (n=778)			Full siblings (n=522)			Uncle-nephew (n=859)			First cousins (n=468)		
	Sen	Spe	Acc	Sen	Spe	Acc	Sen	Spe	Acc	Sen	Spe	Acc
TNSA	99.74	99.70	99.71	94.25	99.40	97.63	74.16	86.20	80.63	72.01	64.40	66.83
LR	98.59	98.90	98.76	95.98	99.50	98.29	76.72	90.60	84.19	73.93	69.00	70.57
logistic regression	100.00	99.90	99.94	97.89	99.00	98.62	86.96	83.30	84.99	66.67	79.60	75.48
RF	99.99	99.98	99.99	99.11	99.62	99.44	89.04	91.54	90.34	61.41	88.25	79.69

Sen, sensitivity; Spe, specificity; Acc, accuracy; TNSA, total number of shared alleles; LR, likelihood ratio; RF, random forest.

족임에도 불구하고 배제시킬 가능성이 높다. 그렇기 때문에 TNSA나 우도비(Log LR)만으로는 3촌 이상 혈연관계를 구분하기에는 다소 한계점이 있기 때문에 보다 확장된 분석 방법인 데이터마이닝 기법의 도입이 필수적이다.

따라서 최종적으로 지금까지의 모든 분석 방법들을 총괄적으로 비교한 결과, 데이터마이닝 기법 중의 하나인 RF가 기존의 대립형질 수(TNSA), 우도비(Log LR), 로지스틱 회귀분석 방법보다 정확도(1촌, 99.99%; 2촌, 99.44%; 3촌, 90.34%; 4촌, 79.69%)가 높게 나타났다(Table 7). 하지만 그럼에도 불구하고 민감도 부분에서 여전히 1촌(99.99%), 2촌(99.11%)에 비해 3촌(89.04%), 4촌(61.41%)에서 다소 낮은 결과를 얻었지만, 기존의 방법들보다는 민감도 또는 특이도가 증가하였다(Table 7). 하지만 다른 혈연관계와는 달리 4촌의 경우 RF를 적용하였을 때 전체적인 정확도(79.69%)나 특이도(88.25%)는 다른 기법들보다 높게 나타났지만 민감도(61.41%)는 다소 낮게 나타났다(Table 7). 일반적으로 민감도와 특이도를 모두 높일 수 있을 경우가 가장 좋은 판단 기준이 될 수 있으나 민감도와 특이도는 서로 음의 상관을 나타내는 경향이 있기 때문에 어느 부분을 좀 더 우선 고려해야 되는 기준이 필요할 것이다. 혈연관계 판정에 있어서 민감도는 사촌관계를 혈연관계로 정확하게 판단하는 것이며, 특이도는 사촌이 아닌 경우 이를 비혈연관계라고 판단하는 수치를 의미하며, 이들을 모두 통합적으로 고려한 것이 정확도에 해당된다. 본 연구는 연구 특성상 혈연관계를 혈연관계로, 비혈연관계를 비혈연관계로 판정하는 모든 경우를 고려해야 되기 때문에 정확도를 기준으로 분석 기법들을 비교하였고, 정확도 기준으로는 RF 기법이 가장 높게 나타났다. 하지만 보다 정확한 판정을 위해서는 어떤 한가지 기법보다는 민감도가 높게 나타났던 대립형질 수 비교나 LR 분석을 서로 병합하여 사용하는 것이 효과적이라 판단한다.

본 연구는 기존의 혈연식별 방법과는 달리 데이터마이닝 기법을 통해 식별력을 높일 수 있는 기법을 연구하였다. 이러한 결과는 추후 먼 친족 간의 혈연관계 거리 계산에도 응용가능할 것으로 판단되며, 이를 바탕으로 과학수사에 활용 가능할 수 있을 것이라 판단된다.

ORCID
Su Jin Jeong: https://orcid.org/0000-0001-6754-8925;
Jae Won Lee: https://orcid.org/0000-0002-3718-2704

Conflicts of Interest

No potential conflict of interest relevant to this article was reported.

Acknowledgments

This research was supported by the KU Future Research Grant(KU-FRG, K1720021, 2017) and was supported by the research project for practical use and advancement of forensic DNA analysis, of the Supreme Prosecutors' Office, Republic of Korea (1333-304-260, 2014).

References

1. Butler JM, Hill CR. Biology and genetics of new autosomal STR loci useful for forensic DNA analysis. Forensic Sci Rev 2012;24:15-26.
2. Bieber FR, Brenner CH, Lazer D. Human genetics: finding criminals through DNA of their relatives. Science 2006;312:1315-6.
3. Myers SP, Timken MD, Piucci ML, et al. Searching for first-degree familial relationships in California's offender DNA database: validation of a likelihood ratio-based approach. Forensic Sci Int Genet 2011;5:493-500.
4. Schneider PM. Scientific standards for studies in forensic genetics. Forensic Sci Int 2007;165:238-43.
5. Lee JW, Lee HS, Lee HJ, et al. Statistical evaluation of sibling relationship. Commun Stat Appl Methods 2007;14:541-9.
6. Jeong SJ, Lee JW, Lee SD, et al. Statistical evaluation of common relationships using STR markers in Korean population. Korean Acad Sci Crim Invest 2016;10:107-15.
7. Evett IW, Weir BS. Interpreting DNA evidence: statistical genetics for forensic scientists. Sunderland: Sinauer Associates; 1998.
8. Yang IS, Lee HY, Park SJ, et al. Analysis of Kinship Index distributions in Koreans using simulated autosomal STR profiles. Korean J Leg Med 2013;37:57-65.
9. Gaytmenn R, Hildebrand DP, Sweet D, et al. Determination of the sensitivity and specificity of sibship calculations using AmpF lSTR Profiler Plus. Int J Legal Med 2002;116:161-4.
10. Budowle B, Shea B, Niezgoda S, et al. CODIS STR loci data from 41 sample populations. J Forensic Sci 2001;46:453-89.
11. Cowen S, Thomson J. A likelihood ratio approach to familial searching of large DNA databases. Forensic Sci Int Genet Suppl Ser 2008;1:643-5.
12. Curran JM, Buckleton JS. Effectiveness of familial searches. Sci Justice 2008;48:164-7.
13. Cox DR. The regression analysis of binary sequences. J R Stat Soc Ser B Methodol 1958;20:215-42.
14. Fisher RA. The use of multiple measurements in taxonomic problems. Ann Eugen 1936;7:179-88.
15. Bickel PJ, Levina E. Some theory for Fisher's linear discriminant function, 'naive Bayes', and some alternatives when there are many more variables than observations. Bernoulli 2004;10:989-1010.
16. Dudoit S, Fridlyand J, Speed TP. Comparison of discrimination methods for the classification of tumors using gene expression data. J Am Stat Assoc 2002;97:77-87.
17. Vapnik VN. The nature of statistical learning theory. Berlin:

Springer; 2000.
18. Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and regression trees. Monterey: Wadsworth & Brooks/Cole Advanced Books & Software; 1984.
19. Altman NS. An introduction to kernel and nearest-neighbor nonparametric regression. Am Stat 1992;46:175-85.
20. Breiman L. Random forests. Mach Learn 2001;45:5-32.
21. Witten DM, Tibshirani R, Hastie T. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. Biostatistics 2009;10:515-34.
22. Buckleton JS, Triggs CM, Walsh SJ. DNA evidence. Boca Raton: CRC Press; 2004.