유전자정보 기반 민족식별 시스템 개발에 관한 연구

구교찬*, 최은지**, 손혜민*, 김선욱*, 한승조***
*단국대학교 산업공학과
**단국대학교 생명과학과
***국방과학연구소
e-mail:kookyochan@dankook.ac.kr

A Study on the Development of the Ethnic Identification System Based on DNA Data

Kyo-Chan Koo*, Eun-Ji Choi**, Hye-Min Son*, Sun-Uk Kim*, Seung-Jo Han***
*Dept. of Industrial Engineering, Dankook University
**Dept. of Biological Sciences, Dankook University
***Agency for Defense Development

요 약

최근 국내에 체류하고 있는 외국인이 지속적으로 증가하고 있으며, 이와 함께 국내 외국인 범죄도 증가하는 추세를 보이고 있다. 따라서 범죄 수사를 진행할 때 DNA 프로파일링을 통해 민족을 미리 예측할 수 있다면 수사 기간 또한 단축시킬 수 있을 것이다. 이에 본 연구에서는 전 세계적으로 다양한 민족 집단의 mtDNA, Y-DNA, atDNA 변이정보를 수집하여 데이터베이스를 구축하고, 민족식별 가능성을 검증하였으며, 이를 기반으로 민족식별 시스템을 개발하였다. 그리고 민족식별 시스템 성능을 확인하기 위해 3가지 평가모델을 개발하여 시스템을 평가하였다. 첫 번째 데이터 매칭 모델은 민족식별할 DNA 프로필과 일치하는 정보를 데이터베이스에서 탐색하는 방법이고, 두 번째 데이타마이닝 모델은 데이터베이스를 클러스터링하여 구해진 민족별 인덱스를 이용해 확률식으로 계산하는 방법이다. 마지막 휴리스틱 모델은 데이터마이닝 모델에 전문가 인터뷰 결과를 반영하여 민족식별 옵션을 추가한 모델이다. 세 가지 모델에 기반한 평가시스템을 구축하여 평가한 결과 휴리스틱 모델의 성능이 가장 우수했다. 본 연구는 법유전학 및 법과학 분야에서 주로 사용하는 마커를 이용하여 추가적인 분석 없이 기존의 DNA 정보만으로도 민족 집단 식별에 적용할 수 있도록 방법을 확립하였으며, 단일 마커를 이용한 민족식별 방법에 비해 다양한 가능성을 제시하였다.

KeyWords : 유전자정보, 민족식별, 데이터마이닝, 클러스터링, 휴리스틱

1. 서 론

국내에 체류하고 있는 외국인은 대부분 장기체류자로 2013년에 약 157만 명에서 2017년에 약 218만 명으로 지속적으로 증가하고 있다. 이와 함께 국내 외국인 범죄도 2015년 이후 3만 명 이상으로 넘어서며 증가하는 추세를 보이고 있다. 한편, 수사과정에서 언어와 문화의 차이, 전문성의 부족 등의 문제로 인해 인종 차별과 외국인 인권 문제가 부각되기도 한다. 따라서 범죄 수사를 진행할 때 DNA 프로파일링을 통해 민족을 미리 예측할 수 있다면 수사를 진행하는 과정에서 용의자의 민족 계통을 미리 인지함으로써 외국인에 대한 과잉 진압 또는 인종 차별에 대한 우려를 감소시킬 수 있으며 수사 기간 또한 단축시킬 수 있을 것이다. 이러한 민족 집단의 계통을 연구하는 데에 있어 가장 유용하게 사용되는 유전적 도구로는 크게 두 가지, 즉 어머니로부터 자손에게 모계 유전되는 미토콘드리아 DNA(mtDNA)[1,2]와 아버지로부터 아들에게 부계 유전되는 Y-염색체 DNA (Y-DNA)[3]의 정보를 들 수 있다.

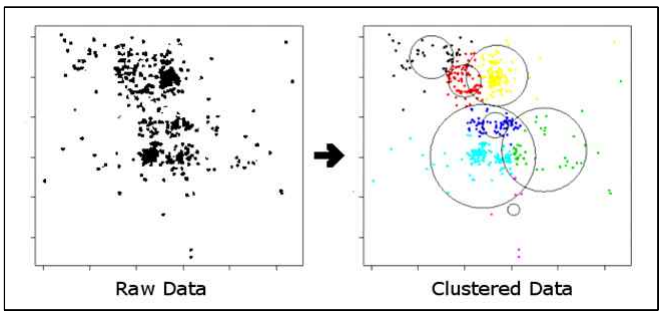
최근에는 외국인과의 혼인으로 인한 다문화 가정의 자녀가 증가하고 있어 혼혈인의 경우 모계와 부계 계통이 서로 다르게 나타날 수 있으므로 정확한 민족 계통을 판단하기 위해서는 mtDNA와 Y-DNA뿐만 아니라 부모로부터 각각 하나씩 유전되는 상염색체 DNA(atDNA)[4] 정보를 모두 조사할 필요가 있다.

이에 본 연구에서는 전 세계적으로 다양한 민족 집단의 mtDNA, Y-DNA, atDNA 변이정보를 수집하여 데이터베이스를 구축하고, 이를 기반으로 민족식별 시스템을 개발하였다.

2. 연구방법

2.1 유전자정보 데이터베이스 분석 및 설계

본 연구에서는 동아시아 집단뿐만 아니라 전 세계적으로 분포하고 있는 지역 및 민족 집단을 대상으로 mtDNA, Y-DNA, atDNA 변이정보를 분석하고 수집하여 데이터베이스를 구축하고, 이를 통해 데이터베이스 기반의 계통 및 유사성 분석[5,6,7]을 이용한 민족 집단 식별 방법을 제시하였다[그림1].



[그림1] 클러스터링 알고리즘

2.1.1 미토콘드리아 DNA

mtDNA 변이정보를 이용하여 민족식별 가능성을 검증하기 위해 mtDNA haplogroup 및 control region 서열 변이정보를 조사한 결과 아시아 대륙에 속하는 8개 집단에 대한 3,289개 표본, 유럽 대륙에 속하는 18개 집단에 대한 10,780개 표본, 아프리카 대륙에 속하는 20개 집단에 대한 2,244개 표본, 아메리카 대륙에 속하는 4개 집단에 대한 3,452개 표본, 중동에 속하는 2개 집단에 대한 227개 표본의 데이터베이스를 수집하였다.

2.1.2 Y-염색체 DNA

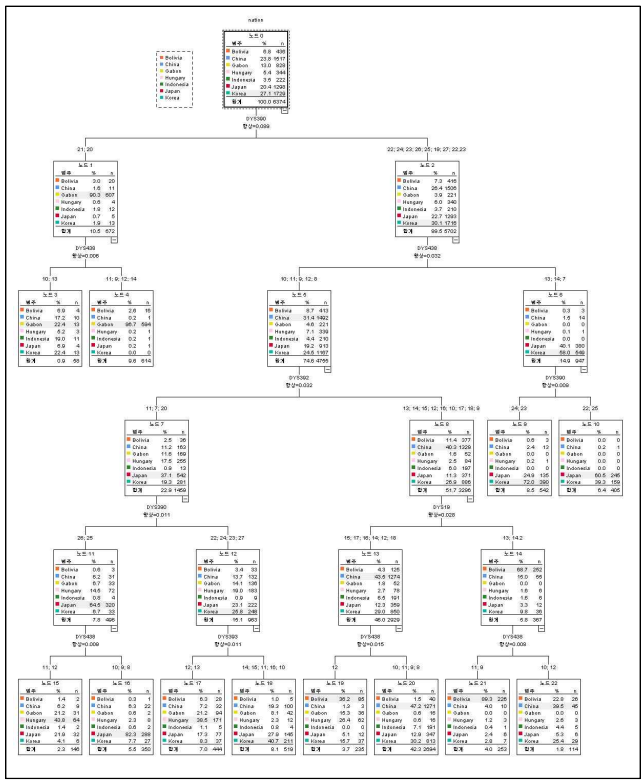
Y-DNA 변이정보를 이용하여 민족식별 가능성을 검증하기 위해 Y-DNA haplogroup 및 STR 변이정보를 조사한 결과 Y-DNA haplogroup의 경우, 아시아 대륙에 속하는 14개 집단에 대한 7,116개 표본, 유럽 대륙에 속하는 7개 집단에 대한 1,753개 표본, 아프리카 대륙에 속하는 3개 집단에 대한 1,103개 표본, 아메리카 대륙에 속하는 1개 집단에 대한 439개 표본의 데이터베이스를 수집하였고, Y-염색체 STR의 경우, 아시아 대륙에 속하는 16개 집단에 대한 24,536개 표본, 유럽 대륙에 속하는 25개 집단에 대한 14,168개 표본, 아프리카 대륙에 속하는 8개 집단에 대한 1,596개 표본, 아메리카 대륙에 속하는 9개 집단에 대한 4,455개 표본, 중동에 속하는 3개 집단에 대한 846개 표본의 데이터베이스를 수집하였다.

2.1.3 상염색체 DNA

atDNA 변이정보를 이용하여 민족식별 가능성을 검증하기 위해 상염색체 STR 변이정보를 조사한 결과 아시아 대륙에 속하는 8개 집단에 대한 2,856개 표본, 유럽 대륙에 속하는 6개 집단에 대한 4,893개 표본, 아프리카 대륙에 속하는 2개 집단에 대한 643개 표본, 아메리카 대륙에 속하는 4개 집단에 대한 5,474개 표본의 데이터베이스를 수집하였다.

2.2 유전자정보를 이용한 민족식별 가능성 검증

민족식별 가능성 검증을 위하여 mtDNA, Y-DNA, atDNA 변이정보 데이터베이스를 이용하였다. mtDNA의 경우, haplogroup 및 control region 변이정보, Y-DNA는 haplogroup 및 STR 변이정보, atDNA는 STR 변이정보를 이용하여 민족 집단 식별 가능성 검증을 수행하였다. 검증 방법으로는 SPSS 프로그램을 이용하여 의사결정나무(Decision Tree) 분석을 하였으며, 분류모형에 민족 정보를 알 수 없는 표본의 프로필을 대입한 결과 각 노드에서 가장 가능성 높은 민족 집단을 예측하였다[그림2].

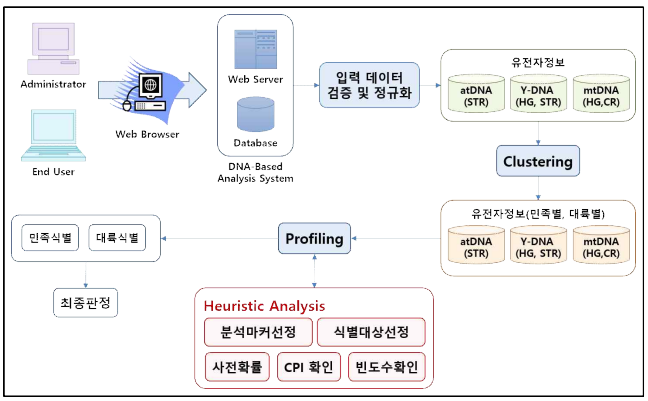


[그림 2] Y-DNA 기반 의사결정나무 모델 결과

3. 연구결과

3.1 시스템 구성

민족식별 시스템은 데이터가 입력되면 검증 및 정규과정을 거쳐 데이터베이스에 저장된다. 검증 및 정규화 과정에서는 데이터 누락, 입력 형식에 맞지 않는 데이터, 입력 불가능한 값 등은 입력되지 않도록 하고, 사용자에게 재확인 후 입력하도록 정보를 제공해 준다. 데이터 검증과 정규화 과정이 끝나면 데이터 클러스터링 과정을 통하여 민족별 유전자정보의 빈도를 이용하여 인덱스를 구한다. 여기서 인덱스는 민족을 식별할 때 사용하는 확률식에 사용된다. 민족식별(Profiling) 결과는 확률식을 이용하여 계산된 결과를 민족별 식별지수가 높은 순위대로 출력한다. 단, 식별지수가 낮거나, 여러 민족이 동일하게 높게 나오는 등 민족식별이 모호한 경우에 5가지 분석 옵션을 추가로 사용하여 민족식별 정확도를 높일 수 있다[그림3].



[그림 3] 민족식별 시스템 프로세스

3.2 시스템 구현

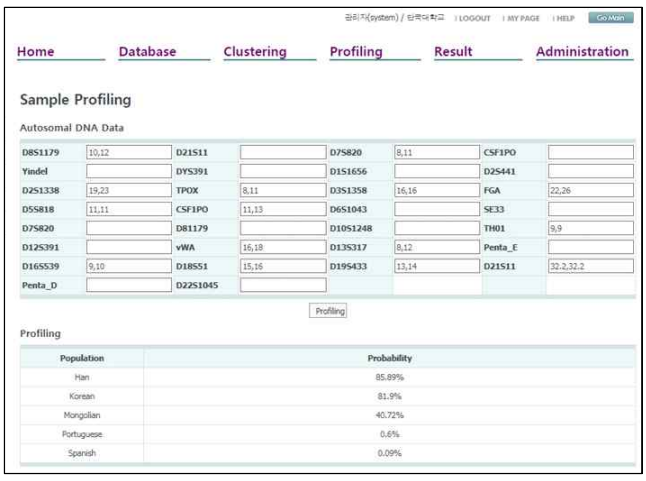
3.2.1 유전자정보 클러스터링

mtDNA 데이터베이스를 클러스터링하면 haplogroup 및 control region(HVR-I, HVR-II, HVR-III)상의 mutation sites의 인덱스를, Y-DNA 데이터베이스를 클러스터링하면 haplogroup 및 Y-STR 마커의 인덱스를 민족별로 계산하여 저장한다. atDNA 데이터베이스 클러스터링하면 A-STR 마커의 빈도와 인덱스를 민족별로 계산하여 저장한다.

3.2.2 유전자정보를 이용한 민족식별

mtDNA 프로필을 민족식별 하려면 mtDNA 데이터베이스의 haplogroup 및 control region(HVR-I, HVR-II, HVR-III)상의 mutation sites의 인덱스 이용하고, Y-DNA 프로필을 민족식별 하려면 Y-DNA 데이터베이스의 haplogroup 및 각각의 Y-STR 마커의 인덱스를 이용한 확률식을 계산하여 민족식별 결과가 높은 순으로 출력한다.

atDNA 프로필을 민족식별 하려면 atDNA 데이터베이스의 각각의 A-STR 마커의 인덱스를 이용하여 확률식을 계산하여 민족식별 결과가 높은 순으로 출력한다[그림4].



[그림 4] atDNA 프로필을 민족식별한 결과 화면

3.3 시스템 평가

3.3.1 평가모델

평가모델은 3가지로 데이터 매칭, 데이터마이닝, 휴리스틱 기반 모델이다. 첫 번째 데이터 매칭 모델은 기존의 유전자정보 데이터를 이용한 시스템들에서 주로 사용하고 있는 방법이다. 이 모델은 민족을 식별할 DNA 프로필 정보와 일치하는 정보를 데이터베이스에서 탐색하여 일치하는 민족이 있는 경우, 그 민족을 해당 민족으로 식별하는 방법이다. 데이터 매칭 모델은 DNA 프로필과 완전히 일치하는 데이터베이스를 탐색하는 것을 원칙으로 하나, 완전히 일치하는 정보가 없을 때에는 가장 유사한 데이터베이스를 탐색하여 민족으로 결정한다. 이 모델은 정확히 매칭되는 샘플이 있는 경우에는 그 식별력이 강력하지만 매칭이 되는 정보가 없는 경우에는 식별이 불가능하다. 또한 모든 데이터베이스의 정보와 매칭해야 하므로 식별하는데 많은 시간이 소요된다. 두 번째 데이터마이닝 모델은 데이터베이스 정보를 클러스터링하여 구해진 민족별 상대적 빈도를 이용해 인덱스를 구하고, 그 값을 확률식에 대입하여 계산하는 방법이다. 이 모델은 데이터베이스에 있는 모든 민족에 대해 상대적인 확률을 구해주는 방법으로 데이터베이스에 해당 민족이 없어도 유사한 민족을 식별해 줄 수 있다. 또한 클러스터링은 실행해 놓으면 민족별 빈도와 인덱스 값을 미리 계산하여 저장해 놓기 때문에 매칭 모델에 비해 처리 속도가 빠르다. 세 번째 휴리스틱 모델은 데이터마이닝에 다양한 분석 옵션을 추가하여 사용자 중심의 분석이 가능하게 한 모델이다. 이 모델은 데이터마이닝 모델을 이용한 민족식별 결과가 모호한 경우에 사용자가 추가로 옵션을 선택하여 식별하는 방법이다. 분석옵션은 전문가 인터뷰를 통하여 결정하였으며, 옵션에는 분석마커선정, 분석대상민족분류, 사전확률, CPI 확인, 빈도수 확인이 있다.

3.3.2 평가결과

데이터 마이닝 모델을 이용해 민족식별한 결과는 다음과 같다. 민족 수준에서의 정확도는 mtDNA 변이정보에서 61%, Y-DNA 변이정보에서 15%, atDNA 변이정보에서 35%로 도출되었고, 대륙 수준에서의 정확도는 mtDNA 변이정보에서 89%, Y-DNA 변이정보에서 72%, atDNA 변이정보에서 75%로 도출되었다. mtDNA 변이정보를 이용하여 민족 집단을 식별했을 때 가장 높은 정확도를 보였으며, Y-DNA 변이정보를 이용하여 민족 집단을 식별했을 때 가장 낮은 정확도를 나타내었다. 이는 Y-DNA 변이정보의 데이터베이스는 atDNA에 비해 많지만 샘플수가 적은 민족 데이터들도 많이 있어서 오히려 민족 식별력을 저하시키며, 구축한 데이터베이스에 따라 사용한 Y-STR 분석 키트가 달라 확률적으로 계산할 수 있는 Y-STR 마커 수에 대해

한계가 있기 때문이다. 즉, 데이터베이스로 민족식별의 가능성을 검증하였지만, 민족식별을 위해서는 복합적인 분석이 필요하다. 이에 복학적인 분석을 위하여 전문가들의 조언을 받아 5가지 분석 옵션을 데이터마이닝 모델에 추가하여 개발한 휴리스틱 모델을 이용해 민족식별 하였다. 그 결과 민족 수준에서의 정확도는 Y-DNA 변이정보에서는 15%에서 67%로 상승하였고 atDNA 변이정보에서는 35%에서 54%로 상승하였다. 또한 대륙식별 정확도는 Y-DNA 변이정보에서는 72%에서 91%로 상승하였고, atDNA 변이정보에서는 75%에서 91%로 상승하였습니다.

4. 결론 및 고찰

본 연구에서는 법유전학 및 법과학 분야에서 주로 사용하는 마커를 이용하여 추가적인 분석 없이 기존의 DNA 정보만으로도 민족 집단 식별에 적용할 수 있도록 방법을 확립하였다. 이와 같이 mtDNA, Y-DNA, atDNA의 변이정보를 이용하여 도출된 결과를 통합적으로 판단한다면 단일 마커를 이용한 민족식별 방법에 비해 다양한 가능성을 제시하거나 정확도를 높일 수 있고 더 나아가 알고리즘을 개발하여 자동화된 시스템으로 구축한다면 국내뿐만 아니라 국제적으로도 널리 사용 가능할 것이다. 또한 과학 수사를 포함한 법과학 분야를 위한 근거자료로 활용할 수 있으며, 향후 표현형 마커 등과 연계하여 유전적 특징과 함께 신체적 특징을 식별할 수 있는 유전자 감식 기술의 개발에 활용될 수 있을 것이다. 한편 정확한 민족식별 방법을 확립하기 위해서는 전 세계의 다양한 민족 집단을 대상으로 한 DNA 정보 분석 및 확장된 데이터베이스 구축이 필수적이다. 데이터베이스 기반의 방법은 데이터베이스의 수가 많을수록 정확도가 높아지기 때문에 데이터가 부족한 민족 집단의 연구 보완이 필요할 것이며 이에 따라 주기적인 데이터베이스 확장 및 관리가 지속적으로 이루어진다면 더 체계적이고 정확한 결과를 얻을 수 있을 것이다.

상기와 같이 기본적으로 데이터베이스가 더 확충되어야 하지만 민족식별 시스템의 정확도를 더 높이기 위해서는 지금의 모델에는 한계가 있다. 이는 유전자 변이정보 외에도 민족식별을 하는데 추가적으로 고려해야하는 요소들이 존재하기 때문이다. 이러한 문제를 해결하는데 딥러닝(Deep Learning) 기술이 하나의 대안이 될 수 있을 것으로 사료된다. 딥러닝 기술은 최근 급속히 발전하고 있어 인간이 발견하기 어려운 내재되어 있는 영향변수 또는 중요한 특성들을 자동으로 식별하는데 큰 기여를 하고 있다. 따라서 민족식별 시스템의 정확도를 높이기 위한 한 수단으로 딥 뉴럴 네트워크(Deep Neural Network)와 같은 딥러닝 모델링 연구가 큰 도움이 될 것으로 기대된다.

참고문헌

[1] Maruyama S., Komuro T., Izawa H., and Tsutsumi H, “Analysis of human mitochondrial DNA polymorphisms in the japanese population”, Biochem Genet, 51, pp. 33-70, 2013.

[2] Sanches N.M., Paneto G.G., Figueiredo R.F., de Mello A.O., and Cicarelli R.M, “Mitochondrial DNA control region diversity in a population from Espirito Santo state, Brazil”, Mol Biol Rep, 41, pp. 6645-6648, 2014.

[3] Watahiki H., Fujii K., Fukagawa T., Mita Y., Kitayama T., and Mizuno N. “Polymorphisms and microvariant sequences in the Japanese population for 25 Y-STR markers and their relationships to Y-chromosome haplogroups”, Forensic Sci Int Genet, 41, pp. e1-e7, 2019.

[4] Nothnagel M., Fan G., Guo F., He Y., Hou Y., Hu S., Huang J., Jiang X., Kim W., Kim K., Li C., Li H., Li L., Li S., Li Z., Liang W., Liu C., Lu D., Luo H., Nie S., Shi M., Sun H., Tang J., Wang L., Wang C.C., Wang D., Wen S.Q., Wu H., Wu W., Xing J., Yan J., Yan S., Yao H., Ye Y., Yun L., Zeng Z., Zha L., Zhang S., Zheng X., Willuweit S., and Roewer L. “Revisiting the male genetic landscape of China: a multi-center study of almost 38,000 Y-STR haplotypes”, Hum Genet, 136, pp. 485-497, 2017.

[5] 148 Hong S.B., Kim K.C., and Kim W. “Population and forensic genetic analyses of mitochondrial DNA control region variation from six major provinces in the Korean population”, Forensic Sci Int Genet, 17, pp. 99-103, 2015.

[6] Toscanini U., Vullo C., Berardi G., Llull C., Borosky A., Gomez A., Pardo-Seco J., and Salas A. “A comprehensive Y-STR portrait of Argentinean populations”, Forensic Sci Int Genet, 20, pp. 1-5, 2016.

[7] Kampuansai J., Völgyi A., Kutanan W., Kangwanpong D., and Pamjav H. “Autosomal STR variations reveal genetic heterogeneity in the Mon-Khmer speaking group of Northern Thailand”, Forensic Sci Int Genet, 27, pp. 92-99, 2017