개인인식을 위한 DNA Genotyping의 전처리 기법

임효빈° 오옥균 공은배
충남대학교 컴퓨터공학과

adonai@cnu.ac.kr° okoh@ce.cnu.ac.kr ebkong@cnu.ac.kr

Preprocessing technology on DNA genotyping for personal verification

Hyobin Im° Okkyun Oh EunBae Kong
Dept. of Computer Engineering, Chungnam National University

요 약

최근 DNA 관련 기술의 개발이 활발하게 이루어지고 있고, 그에 따라 DNA를 개인인식에 사용하고자 하는 시도가 실시간 이용가능성이 높은 DNA 칩 기술과 합쳐져 매우 중요한 이슈로 떠오르고 있다. DNA를 분석하기 위해서는 생물학적 실험이 필수적으로 따르게 되는데 이러한 실험결과를 인식에 적용하기 위해서는 적절한 전처리가 필요하다.

본 논문에서는 여러 장점들로 인해 최근 DNA분석 기술로 주목받고 있는 모세관 전기영동법을 사용하여 DNA를 분석하고, 그 분석물을 개인인식을 위해 genotyping하는 과정에서 전처리가 요구되는 각 경우들에 대해 논하고 적절한 필터링 기법들을 제시한다.

1. 서 론

우리 삶의 전반에 걸쳐 정보화가 이루어짐에 따라 개인을 인식하고 개인의 정보를 보호하는 것의 중요성이 날로 증대되고 있다. 특히 개인 인식은 정보보호 차원에서는 물론 사회학, 법의학 등 여러 분야에 걸쳐 중요한 기술로 인식되고 있다. 개인인식의 여러 방법들 중에서 생체정보를 이용한 인식은 비밀번호와 같이 잊어버리거나 타인에게 누출될 위험이 없고, 열쇠나 보안카드를 사용하는 방법처럼 분실이나 복제의 위험이 적다는 등 많은 장점을 가지고 있다. 그 중에서도 인간의 DNA를 이용한 인식은 일란성 쌍둥이를 제외하고는 모든 사람이 다 다른 DNA를 갖고 있기 때문에 더욱 효과적이라 할 수 있다. 뿐만 아니라 신체의 어느 부위에서 추출한 표본에 대해서도 같은 결과를 갖으며 다른 방법에 비해 월등히 높은 식별력을 가진다.

DNA의 이러한 유용성을 고려할 때 개인인식을 위해 DNA를 분석하고 연구 개발하는 일은 매우 중요하다고 할 수 있다. 또한 국내에서도 최첨단 보안 관련 기술력을 자체적으로 확보하는 면에서나 유전병의 발견 및 치료, 작물의 품종 개량 등 방대한 응용 분야를 볼 때 DNA의 연구 가치는 매우 크다.

따라서 본 논문에서는 DNA를 이용한 개인인식 기법과 DNA 분석 기술에 대해 알아보고, DNA genotyping에 적합한 전처리 기법을 제시하고자 한다.

2. 관련 연구

2.1 DNA typing

인간의 DNA는 약 30억 개의 염기를 지닌 string으로, 이론적으로는 30억 길이의 DNA를 비교하면 사람을 식별해 낼 수 있지만 30억 DNA를 sequencing하여 비교하는 일은 시간적·경제적으로 비현실적이다. 인간은 99.5%~99.9%의 DNA를 공유하고 있으므로, DNA를 경제적으로 개인 식별에 사용하기 위해서는 개인적으로 큰 변동성을 보이는 염색체의 특정 부위를 찾아내어 이 부위의 서열을 비교하는 방법을 사용할 수 있다.

이러한 DNA의 특정 위치를 가리켜 '유전자좌(locus)'라 한다. 이 locus란 몇 번 염색체의 몇 번째 염기 부근을 의미하기도 한다. 사람의 염색체에는 수만 개의 locus가 존재한다. 각각의 locus들은 변이를 포함하기 때문에 서열에 어느 정도 차이가 있을 수 있는데 이렇게 locus가 가질 수 있는 염기 서열을 '대립유전자(allele)'라고 한다. 따라서 우리가 DNA를 이용하여 개인차를 알아본다 함은 특정 locus들의 allele이 같은지 알아본다는 말과 같다고 할 수 있다.

1985년 영국의 유전학자 Jeffreys가 DNA의 단편에서 VNTR(Variable Number Tandem Repeat)이라는 특정 부위를 발견하여, 이것이 마치 우리들의 손가락 지문처럼 개인 식별에 사용될 수 있다는 의미에서 'DNA 지문(DNA fingerprinting)'이라고 일컫게 되었다. 그 후 DNA 지문 분석법은 개인 식별을 위한 가장 강력한 수단이 되었으며, 현재는 'DNA typing'이라는 용어와 함께 사용되고 있다[1].

DNA typing 기법은 크게 다음과 같이 세 가지가 있다.

2.1.1 VNTR typing 기술

VNTR이란 DNA 중 어떤 패턴(core sequence)이 반복적으로 나타나는 부분을 가리키는 것으로, 이 패턴이 반복되는 횟수는 개인에 따라 다르기 때문에 이 부분을 RFLP(Restriction Fragment Length Polymorphism) 방법으로 분석하여 개인인식에 이용할 수 있다. RFLP는 사람에게서 추출한 DNA를 제한효소 처리하여 그 중에서 우리가 관찰하고자 하는 locus를 자른 후, 전기영동과 변성(denaturation), 재결합(Hybridization reaction), X선 감광 등의 과정을 거쳐 방사능 사진 이미지를 얻고 이것의 패턴을 비교하여 두 표본의 출처가 같은지를 알아낸다. 이 방법은 높은 식별력을 갖고 있지만 상대적으로 양질의 DNA가 필요하고 분석하는데 시간이 오래 걸린다는 단점이 있다[2,3].

2.1.2 PCR기반 typing 기술

PCR(Polymerase Chain Reaction)은 DNA의 어떤 부분을 가리키는 말이 아니라 적은 양의 DNA를 가지고도 분석이 용

이하도록 하기 위해 DNA의 일부를 수백만 배 증폭시키는 방법을 말한다. PCR을 이용할 경우 소량의 변질된 표본에 대해서도 적용할 수 있고 자동화하기 쉽다는 장점이 있다. STR(Short Tandem Repeat)이라 불리는 비교적 짧은 base pair(bp)를 갖는 locus들을 PCR을 통해 n번 증폭하면 대립유전자가 두드러지게 나타나 X선 감광 없이 육안으로도 분석할 수 있다. 이 방법은 여러 locus들이 동시에 분석가능하며 시간 및 비용을 절약할 수 있어 현재 가장 많이 사용되고 있다[2,3].

2.1.3 DNA 칩 기술

최근에는 DNA 칩을 이용한 인식 방법도 개발 중이다. 특히 LOC(Lab-on-a-chip)이라고 하는 DNA 칩은 이름 그대로 실험실에서 이루어지는, DNA 표본의 추출로부터 실험 및 분석까지 모든 과정을 하나의 칩 위에 올려놓은 것으로서 수 분 내지 수 초 내에 DNA를 분석할 수 있어 앞으로 DNA 인식 분야에서의 사용이 주목되고 있다.

locus / 이름	D3S1358	vWA	FGA	Amelogenin	…	D7S820
홍길동	15-17	19-19	22-23	XY	…	7-7
이영희	15-18	14-18	20-24	XX	…	7-9
김철수	13-16	16-17	20-24	XY	…	6-8
…	…	…	…	…	…	…

< 표 1 DNA typing의 예 >

2.2 모세관 전기영동(Capillary Electrophoresis;CE)법

CE는 전기장을 걸어주어 시료성분이 전하와 이동도에 따라 각각 일정한 방향과 속도로 이동하여 시료를 분리해내는 전기영동법을 모세관에 적용시켜 물질을 분리하는 것이다. 막강한 분리 능력이라는 전기영동의 장점과 고성능 액체크로마토그래피가 가지고 있는 높은 효율, 빠른 분석시간, 적은 시료 소요량, 전하에 관계없이 모든 용질을 동시 분석할 수 있다는 장점을 모두 가진 이 기법은 현재 필수적인 분석 기술로 자리 잡게 되었다. 이러한 CE는 생물학적 거대 분자, 아미노산, 키랄 약품, 단백질, 탄수화물 등의 분석에 많이 쓰이고 있다.

CE에서 모세관은 완충용액(buffer)으로 채워져 양끝이 완충용액 vial에 담겨져 있고, 이 완충용액 vial에는 고전압 전원에 연결된 전극이 담겨져 있다. 이 전극에 높은 전압-이때 전압은 30kV까지 걸어 줄 수 있다-을 걸어 줌으로써 전기장이 생기게 된다. 전기장이나 압력에 의해 모세관의 한쪽 끝에 시료를 주입하고, 전기장에 의해 형성된 EOF(Electroosmotic Flow)라는 buffer의 흐름과, 시료성분의 전하량으로 인한 전기적 인력에 의해 모세관의 반대편 끝으로 시료성분이 이동하게 된다. 이렇게 이동된 시료성분의 검출은 시료를 주입시킨 반대편 모세관 끝부분에 위치한 검출기를 통해 측정한다[4].

CE를 이용할 경우, 기존에 사용하던 전기영동법보다 훨씬 빠른 시간 내에 DNA절편의 크기에 따른 물질의 분리가 가능하며, DNA 분석 및 검출의 대부분 과정을 자동화 할 수 있어 개인인식에 사용하는데 적합하다.

3. 제시된 필터링 기법

본 논문에서는 현재 국립과학수사연구소에서 개인 식별을 위해 주로 사용하는 10개의 STR locus(D3S1358, vWA, FGA, Amelogenin, TH01, TPOX, CSF1PO, D5S818, D13S317, D7S820)를 사용하였는데, 이 locus들로 개인인식을 할 경우 10^{-9} ~ 10^{-12} 이상의 분별력을 갖는다[5].

3.1 DNA 분석 장비 및 데이터

본 논문에서 사용한 DNA 분석데이터는 Applied Biosystems사의 ABI PRISM 310이라는 DNA 분석기기를 통해 획득되었는데, 이는 HGP(Human Genome Project)에서는 물론 현재 국내외적으로 관련기관과 연구실에서 폭넓게 사용되고 있는 상용 분석 시스템이다.

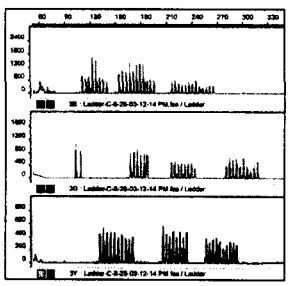
이 장비에서 출력하는 데이터는 실험 sample 분석 데이터와 그 데이터를 genotyping하기 위한 기준이 되는 ladder 데이터로 구성된다.

3.1.1 Ladder 데이터

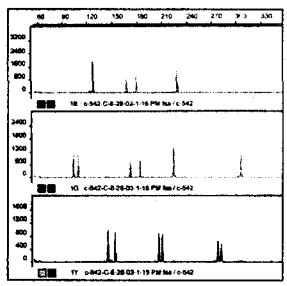
Ladder 데이터는 일종의 측정기준이 되는 데이터이다. sample 데이터가 있을 때, sample 데이터에서 나타난 peak와 ladder 데이터에서 나타난 peak들과 size값을 비교하여 sample 데이터의 allele 수치를 결정할 수 있게 된다. 따라서 ladder 데이터에는 각 locus별로 나타날 수 있는 모든 peak들이 나타나게 된다.

3.1.2 Sample 데이터

Sample 데이터는 모세관 전기영동 실험을 거친 각 사람의 DNA sample이 CCD에 의해 검출된 데이터이다. 각 사람마다 10개의 locus에서 나타나는 allele들의 조합이 일치될 확률은 매우 작기 때문에 이것으로써 개인 식별을 할 수 있게 되는 것이다. 일반적인 경우, 각 locus별로 peak는 하나 또는 2개로 나타나게 되는데 이는 부모로부터 받은 대립유전자가 같은 성질일 경우(homozygote)는 1개, 다른 성질일 경우(heterozygote)는 2개로 나타나기 때문이다. 대표적인 예로, 성염색체의 경우 여자(XX)는 Amelogenin locus에서 peak가 하나로 나타나게 되고, 남자(XY)는 peak가 2개로 나타난다. 간혹 한 locus에서 peak가 3개로 나타나는 경우가 있는데, 이것은 그 locus에 대해 3개의 allele을 갖는 돌연변이일 때로 그 확률은 매우 작다.



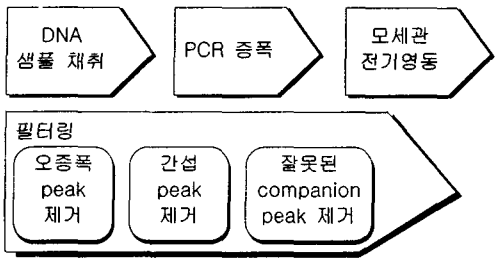
< 그림 1 Ladder 데이터 >



< 그림 2 Sample 데이터 >

ABI PRISM 310 시스템의 경우 PCR을 통한 DNA 증폭 과정에서, 혹은 모세관 전기영동 실험 및 검출과정에서 여러 가지 형태의 오류 값들이 발생할 수 있다. 예를 들어, PCR 과정에서 불필요한 DNA절편까지 포함하여 증폭하거나 모세관 전기영동 실험 결과를 검출하는 과정에서 각 CCD간에 간섭현상이 일어나기도 한다. 이런 경우 ABI 시스템 내에서는 유효한 peak인지를 가리기 위해 적절한 파라미터 조정이 필요하며, 최종 allele 결정시 'OL'이라 하여 시스템이 판단 불가인 경우가 발생하기도 하는데 이는 모두 전문 감정인만이 판단할 수 있다. 하지만 실제 그런 전문인들이 개인인식을 요구하는 모든 장소에 항상 상주하고 있을 수 없기 때문에 파라미터 조정 및 OL값들을 자동으로 처리할 수 있는 전처리 과정이 필요하게 된다.

3.2 필터링이 요구되는 경우와 그에 대한 필터링 기법



< 그림 3 DNA sample 처리과정 >

3.2.1 잘못 증폭된 경우의 필터링

DNA를 분석하기 전, PCR을 통해 원하는 부위를 증폭한다. 이때 원하지 않는 부분까지 함께 증폭되는 경우가 있는데, 이는 올바르지 못한 peak가 검출되는 결과를 초래하며 정확한 결과를 위해서는 잘못된 peak를 제거해야 한다. 반복실험에 의한 관련논문에 의거하여, 이렇게 잘못 증폭된 peak들은 각 dye별로 최대 peak값의 15% 미만 값을 갖는다[6]. 따라서 각 dye별 최대 peak area값을 찾고 그 값의 15% 미만인 값을 갖는 peak들은 모두 제거한다.

3.2.2 간섭현상이 일어나는 경우의 필터링

일반적으로 사용되는 모세관 전기영동 장비는 주어진 DNA sample에 대해 모세관 전기영동 실험을 하고 전용의 CCD 카메라를 사용하여 레이저 유도 형광법에 의해 peak 데이터를 검출한다. 여기서 검출용 CCD 카메라는 각 dye 색상에 따라 각각 검출하게 된다. 즉, blue 검출 CCD, green 검출 CCD, yellow 검출 CCD 이렇게 세 개의 CCD로 촬영하는 것이다. 이 CCD들 간에 종종 간섭현상이 일어나는데, blue에 peak가 나타날 때, green이나 yellow CCD에서도 모두 peak로 검출되는 경우를 말한다. 간섭현상에 의한 peak들은 정상적인 peak 사이즈에서 ±0.5bp의 범위 내에서 나타나게 된다[7].

이러한 경우, 각 peak들은 본래 형광염료에 의해 염색된 DNA 절편들이 형광물질의 색상에 따라 전용 CCD에 의해 검출되도록 설계되어 있기 때문에, 간섭현상에 의해 나타난 잘못된 peak들은 원래의 정상적인 peak에 비해 그 값이 작게 나타나게 된다. 따라서 같은 위치(size)에서 나타난 peak들을 비교하여 그중에서 가장 큰 값을 갖는 peak를 정상 peak로 보고, 나머지 peak들을 비정상 간섭 peak로 간주하여 제거하는 방법으로 필터링을 한다.

3.2.3 잘못된 companion peak의 필터링

특정한 locus에서 대립유전자가 2개일 때, 한 peak area값이 다른 peak값의 60% 미만인 경우, 이 peak는 잘못된 peak로 간주하여 제거하도록 한다[8]. 즉, allele이 하나인 것으로 간주한다. 이것은 정상적인 peak의 경우 PCR 과정에서 충분히 증폭되므로 그 값이 다른 peak의 값과 큰 차이를 보이지 않기 때문이다. 따라서 일반적으로 한 peak가 다른 peak의 60% 미만의 값을 갖는다면, 그것은 잘못된 peak로 보는 것이 올바르다.

3.3 실험 결과

본 시스템을 국립과학수사연구소에서 제공받은 200여명의 실제 DNA sample 데이터에 적용해 본 결과, 앞에서 언급한 여러 오류 값들에 대한 필터링이 국립과학수사연구소의 전문 감정인들이 내린 결과와 비교하여 90%이상의 정확도를 보이는 것을 알 수 있었다. 또한 여러 데이터들에 대해 반복 실험한 결과, 잘못 증폭된 데이터들의 경우 최고 peak의 20% 이하의 값을 갖는 peak들은 noise로 간주하여 제거한 경우 기존의 15% 미만의 peak를 제거하는 방법과 유사한 결과를 나타냄을 알 수 있었다. 뿐만 아니라 companion peak에 대한 필터링에서 종전에 사용하던 60% 필터링 대신 minor peak가 major peak area값의 70% 미만의 값을 갖는 경우를 잘못된 companion peak로 간주해 제거함으로 비교해야 할 peak수를 줄이면서도 기존의 방법에서와 동일한 결과를 얻을 수 있었다.

4. 결론

본 논문에서는 개인 인식을 위한 DNA genotyping의 필터링 기법을 제시하였는데, 이 기법에 의해 전문 감정인의 추가적인 조작 없이도 필요한 필터링을 자동적으로 처리할 수 있었다.

PCR을 포함한 생물학적 실험과정이 완전 자동화 및 실시간화 된다면 본 필터링 방법을 개인 인식 시스템에 적용하여 엄격한 출입통제가 요구되는 보안기관의 출입관리를 위해 폭넓게 사용할 수 있다.

감사의 글

본 연구는 한국전자통신연구원의 "DNA를 이용한 개인인식 시스템에 적합한 전처리 기술에 관한 연구" 위탁과제에 의해 지원되었음. 이 연구를 위해 관련 정보와 실험 데이터를 제공해 주신 국립과학수사연구소 한면수실장님과 홍승범선생님에게 감사드립니다.

참고문헌

1. 정연보, *"DNA Typing"*, 인제대학교 출판부, 1996

2. National Research Council, *"DNA Technology in Forensic Science"*, National Academy Press, 1992

3. National Research Council, *"The Evaluation of Forensic DNA"*, National Academy Press, 1996

4. D. R. Baker, *Capillary Electrophoresis*, John Wiley & Sons, Inc., New York, U. S. A., 1995(Chapter 2).

5. Xiuling Wang*, Toshiko S., Akiko S., *Analysis of STP polymorphisms in the northeast Chinese population*, Forensic Science International 132:161-163. 2003

6. Clayton TM., Whitaker JP., Sparkes R., Gill P., *Analysis and interpretation of mixed forensic stains using DNA STR profiling.* Forensic Science International. 91(1):55-70. Jan 1998

7. Gill P., Sparkes B., Buckleton J.S., *Interpretation of Simple Mixtures When Artefacts Such as Stutters are Present - with Special Reference to Multiplex STRs Used by the Forensic Science Service.* Forensic Science International. in press 1998,

8. Gill P., Sparkes R., Kimpton C., *Development of Guidelines to Designate Alleles Using an STR Multiplex System.* Forensic Science International. 89:185-197. 1997