제18회 한국정보처리학회 추계학술발표대회 논문집 제9권 제2호 (2002.11)

DNA의 반복염기 서열 데이터베이스를 활용한 친자확인 방법

이운*, 임종태
공주대학교 컴퓨터공학과
e-mail:950916@hanmail.net

A Paternity Testing Method Using DNA Repetive Sequences

Un Lee*, Jongtae Lim
Dept of Computer Engineering, Kongju National University

요 약

DNA의 염기서열이 밝혀지면서 인간 생체에 대한 다양한 연구가 활발히 진행되고 있다. 응용분야 중 친자확인에 DNA 염기서열을 이용하려는 시도가 최근에 연구되고 있다. 본 연구는 DNA의 반복 염기서열을 이용하여 수작업으로 이루어지고 있는 친자 확인 방법을 데이터베이스 기술을 이용하여 수행하는 최초의 연구이다. 방대한 양의 자료에서 친자확률을 계산하는데 걸리는 시간은 DB를 구축하는 방법에 크게 좌우된다. 본 논문에서는 친자확률을 계산하는 시간을 최소화할 수 있는 DB를 설계하고, 또한 최소 시간내에 질의 결과를 획득하는 질의 구성하는 방법을 제안한다.

1. 서론

DNA정보를 이용한 친자확인은 현재 수작업으로 이루어지고 있다. 본 논문에서는 수작업으로 인한 친자확인에 많은 시간이 소요됨에 착안하여 컴퓨터의 데이터베이스 기술을 적용하여 사용자가 만족할 만한 시간내에 원하는 결과를 받아볼 수 있는 방법을 연구한다.

일반적으로 친자확인에 부, 모 그리고 자식의 DNA정보를 가지고 친자를 확인한다. 본 논문에서는 부와 모를 짐작할 수도 없는 경우 방대한 량의 모집단 자료에서 자신의 부 또는 모를 찾는 것을 목적으로 하고 있다.

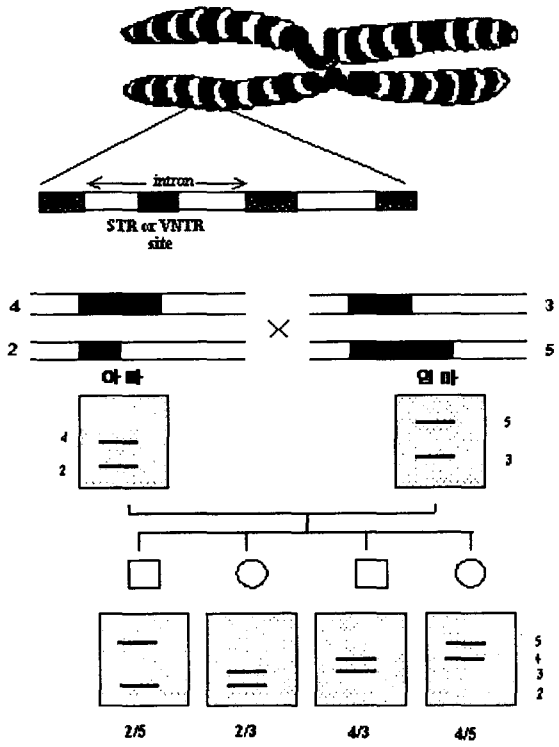
가까운 장래에 남과 북이 통일되어 현재 2천만명에 이르는 남북 이산가족이 서로의 친자들을 확인하려고 할 때에 본 시스템을 활용할 수 있도록 처리시간을 최소화하는 방법을 중점적으로 다룬다. 본 논문의 제2장은 생물학계에서 밝혀진 친자확인 방법을 설명하고, 제3장은 DB 설계방법을 설명한다. 제4장은 질의 구성 방법을 설명하고 제5장에서 결론을 맺는다.

2. 친자확인 방법

인간의 염색체에는 염기서열의 일정한 부위가 반복해서 나타나는 일정한 초변이성 단위반복구조(tandem repeat sequence)가 존재한다. 이러한 반복단위의 반복횟수는 사람에 따라서 적게는 1회에서 많게는 수십회까지 이루어져 여러 형태로 나타나는 유전자 좌위가 있는데, 반복단위의 염기서열이 대개 14-70 개인 것을 VNTR(Variable Number of Tandem Repeats), 2-7 개인 것을 STR (Short Tandem Repeats)이라고 부른다.[2], VNTR 또는 STR은 일정한 중심염기서열(core sequence)이 직렬반복(tandem repeat)됨으로써 나타나는 반복염기서열(repetive sequence)로 사람마다 반복되는 횟수가 다르다. 이런 반복서열에 있어서의 다형성은 대립유전자의 수가 매우 많고 반복되는 중심서열의 종류도 다양하여 법의학적인 개인식별에 주로 이용된다. 또한 이런 특정 염기의 반복은 세대를 거쳐 유전되기 때문에 유전자 감식을 이용한 친자 확인에도 이용되고 있다.

DNA는 이중 나선형 구조이기 때문에 한 좌위에서의 STR도 한 쌍의 값(좌STR, 우STR)을 가진다. <그림-1>은 부(좌STR, 우STR)와 모(좌STR, 우

STR)는 하나씩의 값을 자식에게 유전하는 모식도이다.



<그림-1>STR의 유전 모식도

친자 확인을 통계학적인 방법으로 분석하여 친자 확율 값을 산출한다. 다음과 같이 4단계로 수행된다.

2.1 STR빈도

한국사람 표본의 각 좌위별 STR을 조사한다. 한 좌위의 총 자료수로 하나의 STR이 나타난 자료수를 나누어 좌위의 STR빈도를 계산한다. 좌위의 STR이 <표-1>과 같다면, 이 좌위에서 STR 7에 해당하는 STR빈도 F(7)은

$$F(7) = \frac{\text{7값의 자료수}}{\text{총자료수}} = \frac{2}{1178} = 0.0002$$

이다.[3]

2.2 친부지수(PI: Paternity Index)

모집단에서 신청자의 부를 찾기 위해서 두 사람의 친자확인 방법을 사용한다. 두 사람의 친자확인 방법은 먼저 자위의 두 쌍 STR 유형에 따라 계산 공식을 달리하여 친부지수를 계산한다<표-2>.

STR	자료수	STR빈도
7	2	0.002
8	1	0.002
9	46	0.039
10	281	0.239
11	286	0.243
12	471	0.400
13	79	0.067
14	9	0.008
15	3	0.003
Total	1178	1.000

<표-1> 좌위의 STR빈도

자	부친	친부지수 공식	예
qq	qq	1/F(q)	(5, 5) (5, 5)
pq	qq	1/{2F(q)}	(5, 7) (5, 5)
qq	qr	1/{2F(q)}	(5, 5) (5, 7)
pq	pq	{F(p)+F((q)} / {4F(p)F(q)}	(5, 7) (5, 7)
pq	qr	1/{4F(q)}	(5, 7) (7, 9)
op	qr	0	(5, 7) (6, 8)

<표-2> 유형별 친부지수 공식[1]

<표-2>에서 F(q)는 STR이 q인 경우의 STR빈도 값이다.

자의 STR이 (q, q)이고 부친도 (q, q)인 유형 즉, 자의 좌우 값이 같고 부친도 좌우 값이 같으며 자와 부친의 값이 같은 유형의 친부지수는 1/F(q)값을 갖는다.

자의 STR이 (p, q)이고 부친의 STR이 (q, q)인 유형 즉 자의 좌우 값이 다르고 부친은 좌우 값이 같으며 부친의 STR이 자의 STR 중 하나와 같은 유형의 친부지수는 1/2F(q)값을 갖는다.

자의 STR이 (q, q)이고 부친의 STR이 (q, r)인 유형 즉, 자의 좌우 값이 같고 부친은 좌우 값이 다르며 자의 STR이 부친의 STR 중 하나와 같은 유형의 친부지수는 1/2F(q)값을 갖는다.

자의 STR이 (p, q)이고 부친도 (p, q)인 유형 즉, 자의 좌우 값이 다르고 부친도 좌우 값이 다르며 자

의 좌STR과 부친의 좌STR이 같고 자의 우STR과 부친의 우STR이 같은 유형의 친부지수는 {F(p)+F((q)} / {4F(p)F(q)}값을 갖는다.

자의 STR이 (p, q)이고 부친도 (q, r)인 유형 즉, 자의 좌우 값이 다르고 부친도 좌우 값이 다르며 자의 좌우 STR 중 하나와 부친의 좌우 STR 중 하나가 같은 유형의 친부지수는 1/4F(q)값을 갖는다.

자의 STR이 (o, p)이고 부친도 (q, r)인 유형 즉, 자의 좌우 값이 다르고 부친도 좌우 값이 다르며 자의 좌우 STR 중 하나와 부친의 좌우 STR 중 같은 값이 없는 유형의 친부지수는 0이다. 하지만 실제 계산에서는 0.00001로 계산하여 돌연변이성의 유전을 감안한다.

2.3 결합 친부지수

검사한 좌위별로 각 유형에 맞는 친부지수 계산 공식에 의하여 친부지수를 구한다. 각 좌위의 친부지수를 다 곱하여 결합 친부지수를 계산한다.

2.4 친자확률(Probability of Paternity)

결합 친부지수로 친자확률(%)을 구하는 공식은 다음과 같다.

$$\text{친자확률}(\%) = \frac{\text{결합 친부지수}}{(\text{결합 친부지수}+1)} \times 100$$

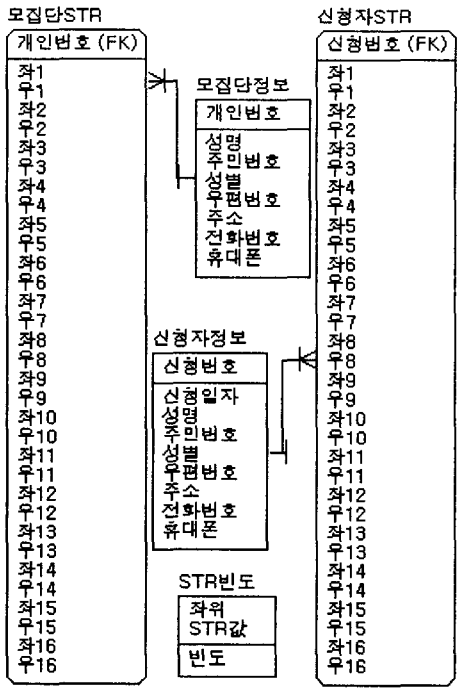
친자확률이 99%이상이면 두 사람을 친부자지간으로 간주한다.

3. DB 설계

DB설계는 모집단정보, 모집단STR, 신청자정보, 신청자STR, STR빈도 Table로 구성된다. 본 논문은 일반적인 설계 방법과 개선된 설계방법의 두 가지 유형으로 DB를 설계 및 질의 구성하였다. 개선된 방식의 설계를 제안코자 한다.

3.1 일반적 설계

일반적 설계는 <그림-1>과 <표-2>의 직관적 활용에 기반을 둔다. 즉, DNA검사 결과 최대 16좌위 STR을 이용하여 친자확률을 계산하며 각 좌위의 좌우 STR을 Table의 한 Column으로 DB를 설계한다. 한사람의 STR의 개수는 16*2이며 하나의 Row에 저장한다. STR 저장 Table의 기본키(Primary Key)는 '개인번호'이다. <그림-2>는 일반적 설계의 개념을 보인 것이다.

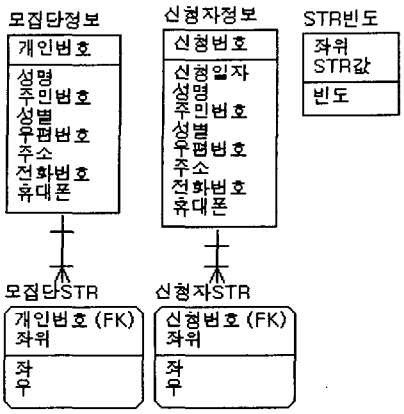


<그림-2> 일반적 설계

3.2 개선된 설계

개선된 설계는 모집단STR과 신청자STR의 Table의 구조를 변경하여 처리시간을 단축을 주목적으로 한다. 그러므로 위의 정보를 이용한 일반적인 설계 방식과 자의 STR과 부의 STR이 모두 다른 경우가 3번 이상인 경우는 친자일 확률이 극히 낮다는 점에 착안한 설계이다. 즉, 한 사람의 STR (좌,우) 한 쌍의 값을 하나의 Row에 저장하는 DB를 설계한다. 즉, 한 사람의 16쌍의 STR을 저장하기 위해서 16행(Row)이 필요하다.

STR 저장 Table의 기본키(Primary Key)는 '개인번호' + '좌위'이다<그림-3>. 모집단과 신청자의 같은 좌위의 STR을 비교하기 위하여 좌위가 기본키(Primary Key)가 되도록 설계하였다.



<그림-3> 개선된 설계

4. 질의 구성 방법

질의를 사용하여 친자확인을 처리하는 일련의 과정은 다음과 같다.

1단계 : 신청자의 STR을 DB에서 가져온다.

2단계 : 친부 모집단 중 1명의 STR을 DB에서 가져온다.

3단계 : 신청자와 모집단 중 선택된 1명의 좌위 STR의 유형별로 STR빈도 값을 이용하여 친부지수를 계산한다.

4단계 : 친부지수를 이용 결합 친부지수를 계산한다.

5단계 : 결합 친부지수를 이용 친자확률을 계산한다.

6단계 : 계산된 결과를 DB에 저장한다.

4.1 일반적 설계에서의 질의

모든 친부 모집단의 수만큼 질의를 반복 처리한다. 위의 2단계 처리를 위한 질의문장으로써 친부 모집단 DB에서 1명씩 STR을 가져오기 위한 SQL 문장은 <그림-4>와 같다.

```
DECLARE 커서명 CURSOR FOR
 SELECT 개인번호, 좌1, 우1, 좌2, 우2, 좌3, 우3,
        좌4, 우4, 좌5, 우5, 좌6, 우6, 좌7, 우7,
        좌8, 우8, 좌9, 우9, 좌10, 우10,
        좌11, 우11, 좌12, 우12, 좌13, 우13,
        좌14, 우14, 좌15, 우15, 좌16, 우16
   FROM 모집단STR ;
```

<그림-4> 일반적 설계 SQL문장

4.2 개선된 설계에서의 질의

검색조건에 부합하는 대상, 즉 친부일 확률이 매우 높은 자료만을 대상으로 질의를 처리한다.

<표-2>에서 자가 oq이고 부친이 qr인 경우에 자의 STR 두 개와 부친의 STR 두 개가 같은 것이 없는 좌위가 3개 이하인 모집단을 대상으로 하는 SQL은 <그림-5>와 같다.

```
DECLARE 커서명 CURSOR FOR
 SELECT DISTINCT popul_str_new.p_no
          FROM popul_str_new
         WHERE popul_str_new.p_no not in(
  SELECT 모집단STR.개인번호
    FROM 모집단STR,
         신청자STR
 WHERE ( 신청자STR.좌위 = 모집단STR.좌위 ) AND
         ( ( 신청자STR.신청번호 = :검사자 ) AND
         ( 신청자STR.좌 <> 모집단STR.좌 ) AND
         ( 신청자STR.좌 <> 모집단STR.우 ) AND
         ( 신청자STR.우 <> 모집단STR.좌 ) AND
         ( 신청자STR.우 <> 모집단STR.우 ) )
 GROUP BY 모집단STR.개인번호
  HAVING ( count(*) >= 3 ) ) ;
```

<그림-5> 개선된 설계 SQL문장

5. 결론

본 논문에서는 친자확인을 위하여 DNA 반복염기 서열 DB를 설계하였다. 개선된 설계는 일반적 설계의 Table보다 Row수가 많아지고 친부의 확률이 극히 낮은 모집단이 확인 대상에서 제외시킬 수 있다. 친부 대상 모집단 중 친부의 확률이 매우 높은 소량의 자료만 친자 확률 계산의 대상이 되므로 시간을 크게 줄일 수 있다.

현재 제안한 설계 및 질의처리 방법이 시스템 사용자가 만족할 만한 성능을 나타내는지에 대한 성능 평가가 수행 중에 있다.

참고문헌

[1] www.dna-view.com/patform.htm

[2] www.dnatyping.co.kr

[3] DNA 프로필연구회 '유전자 감식' 탐구당출판사, 2001.12.5