RFLP 분석을 위한 영상처리 프로그램의 개발

김기원*, 정진영**, 김석훈***

요약

RFLP 분석은 분자생물학에서 매우 많이 응용되고 있는 방법이다. 가장 일반적인 예로 유전질환을 일으킬 지도 모르는 인간 유전자의 돌연변이를 찾는 데 응용되고 있다. RFLP 방법은 글로빈(globin)유전자의 결함에 의해 유발되는 여러 유형의 지중해성 빈혈증과 어떤 유형의 혈우병 진단을 위해 유용하다는 것이 입증되었다. 이 분석 방법은 두 개체의 제한효소(restriction endonuclease) 띠 무늬(banding pattern)의 차이를 분석하는 방법으로 개체들의 띠 무늬의 개수와 무늬가 나타난 위치들의 차이를 분석하게 된다. 본 논문에서는 수작업으로 이루어지는 RFLP 분석을 자동화하는 프로그램을 개발한다. 이를 위해 입력되는 DNA 띠 무늬 패턴을 전처리 한 후 특징을 추출하고 표준 패턴과의 패턴 매칭을 통해 유사도를 계산하였다. 계산에 사용된 특징은 띠 무늬의 개수와 띠 무늬가 나타난 위치 정보이다.

Development of Image Processing Program for RFLP Analysis

Gi-Weon Kim*, Jin-Young Jung**, Seok-Hun Kim***

ABSTRACT

RFLP analysis in molecular biology is a commonly used method. For example, RFLP analysis is used to identify mutations in the genes. RFLP method caused by genetic defects in the Mediterranean can be used to diagnose anemia. It also proved useful in diagnosing hemophilia. This method analyzes the differences in banding patterns. In general, the number of band patterns and the position of a pattern is analyzed. In this paper, we develop image processing program. The program will automatically perform RFLP analysis. First, the input DNA band patterns were pre-processed. Then the features of the input patterns were extracted. By pattern matching with the standard pattern similarity was calculated. In this paper, the number of banding patterns and location of the banding pattern was used to calculate the similarity.

Key Words : RFLP, banding pattern, feature extraction, pattern matching, similarity

* 초당대학교 컴퓨터과학과(✉kwkim@chodang.ac.kr)
** 대전보건대학교 바이오정보과
*** 타임시스템(주)
· 제1저자(First Author) : 김기원 · 교신저자(Correspondent Author) : 김석훈
· 접수일(2011년 11월 7일), 수정일(1차 : 2011년 12월 7일), 게재확정일(2011년 12월 9일)

Ⅰ. 서 론

본 논문에서는 DNA 이미지 관리 프로그램의 검색 기능을 향상시키기 위한 방법으로 RFLP 분석을 기반으로 하는 검색 시스템을 개발하고자 한다. RFLP(Restriction Fragment Length Polymorphism) 분석은 분자생물학에서 많이 응용되고 있는데, 사용되는 분야에는 유전질환을 일으키는 인간 유전자의 돌연변이를 찾거나 혈우병 또는 지중해성 빈혈병을 찾는데 이용된다[1]. RFLP 분석방법은 각각의 개체의 제한효소(restriction endonuclease)를 절단하였을 때 발생된 띠 무늬(banding pattern)의 차이를 분석하는 방법이다[2,3]. 다음의 그림 1은 RFLP 분석에 사용되는 DNA 전기영동사진을 보인다.

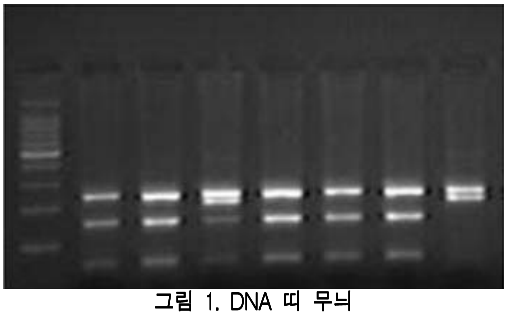


그림 1. DNA 띠 무늬
Fig. 1 DNA band Pattern

현재 띠 무늬들의 차이를 분석하는 과정은 수작업에 의해 이루어지거나 고가의 분석 장비를 사용해야만 한다. 본 논문에서는 이 같은 문제를 해결하기 위해서 DNA 이미지 분석 프로그램을 개발하여 빠른 시간 내에 정확한 DNA 분석 정보를 얻을 수 있도록 하였다. 연구 목적을 효율적으로 달성하기 위해 연구된 세부적인 내용은 다음과 같다. DNA 이미지를 입력받아 전처리를 수행한다. 전처리를 함으로써 불필요한 이미지 정보를 없앨 수 있어 연산효율을 높이고 정확한 분석이 이루어지게 된다. 그 후 전 처리된 영상에서 특징을 추출하고 추출된 특징을 이용하여 참조패턴과의 비교를 통한 유사도 검색을 실행한다. 유사도 계산에 사용된 특징으로는 띠 무늬의 개수와 띠 무늬가 나타난 위치 정보를 사용하였다.

Ⅱ. 이론적 배경

2.1 인식 시스템의 구조

인식 시스템은 모집단의 데이터로부터 패턴이 가지는 특징을 추출하여 인식하는 시스템을 말한다. 패턴 인식은 미리 정의된 군집의 집합에 속하는 표준패턴과 현재 입력된 패턴에 대한 적절한 분석을 통해 입력 패턴이 어떤 군집에 가장 유사한지를 결정하는 것이다. 인식 시스템은 다음의 그림 2와 같은 단계로 처리된다[4,5].

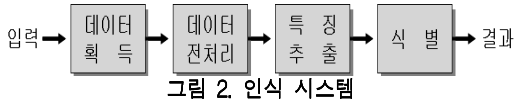


그림 2. 인식 시스템
Fig. 2 Recognition System

가. 데이터 획득

인식 문제에 있어서 A/D 변환등의 방법으로 획득해야 하는 물리적 신호는 데이터 입력 장치에 따라 종류가 다르게 된다. 예를 들면 영상처리 분야에서는 영상 입력장치를, 음성처리 분야에서는 마이크 등을 통해 입력된 신호를 그 대상으로 한다. 최근에 널리 사용되는 영상 획득장치에는 Line scan CCD 카메라, Area scan CCD 카메라와 레이저광을 주사하여 처리하는 장치등이 있다. 본 논문에서는 사전에 디지털화된 DNA 영상을 사용한다.

나. 데이터 전처리

전처리 과정에서는 신호대 잡음비의 향상이나 인식

의 단순화와 효율성 등을 위해 획득된 측정 신호를 변형하는데, 전처리 과정에서는 먼저 데이터 획득 과정에서 입력되는 잡음을 제거한다. 잡음이 제거되면 그 다음 단계로 왜곡을 제거하게 된다. 왜곡은 일반적으로 관측위치에 의한 왜곡과 장소에 따른 밝기의 변화 등에서 발생된다[5,6].

다. 특징추출

전처리된 데이터를 식별단계에서 효율적으로 사용할 수 있는 형태로 단순화시키는 단계로서 추출된 특징에 따라 복잡도가 결정된다. 패턴공간의 축소와 결정단계를 위한 특징 수의 관계에 대한 적절한 조정을 통해 시스템의 효율성을 높일 수 있다. 특징 추출의 목적은 정보의 손실을 적게 하면서 패턴의 차원을 줄여 중복성을 없애고 인식에 필요한 시간 및 기억공간을 줄이는데 있다[6].

라. 식별

식별은 미지의 입력패턴이 주어졌을 때 그 패턴이 어느 패턴의 군집에 속하는가를 결정하는 것이다. 식별을 위해서는 일반적으로 표준패턴이 만들어져 있어야 하며, 입력되는 패턴과 각각의 표준패턴들의 유사도를 비교하여 이중 가장 유사한 표준패턴을 식별 결과로 정하게 된다. 이와 같은 식별방법에는 크게 통계적 방법, 구조적 방법, 인공지능형 방법 등이 있다[5,7].

2.2 영상 전처리 및 영상 특징

영상처리에서 인식에 불필요한 정보나 잡음을 제거하여 빠르고 정확한 인식결과를 얻기 위해 사용되는 전처리 및 특징들에는 다음과 같은 방법이 사용된다.

가. 히스토그램 평활화

히스토그램 평활화(histogram equalization)는 영상 명도값의 분포를 나타내는 히스토그램이 균일하게 되도록 변환하는 처리로서 출력 영상이 각 명도에서 동일한 숫자의 픽셀을 갖도록 입력영상의 명도를 조정하는 것을 의미한다. 이 변환을 쓰는 이유는 콘트라스트(contrast)가 나쁜 영상의 개선에 유효한 방법이기 때문이다[5].

$$h(i) = \frac{G}{N^* H(i)} \quad (1)$$

, G: 최대밝기
, N: 영상크기
, $H(i)$: 정규화 누적값

나. 이진화

이진화(thresholding) 방법은 영상의 조명 상태가 심각하게 나쁜 경우 콘트라스트가 심각하게 낮은 경우가 있다. 이럴 때 보다 확실하게 배경과 이미지를 구분하기 위해 이진화를 시행한다[4]. 다음의 식 2는 이진화 방법을 설명하고 있다[5].

$$B_{(m,n)} = \begin{cases} 1, & f_{(m,n)} \le T \\ 0, & f_{(m,n)} \rangle T \end{cases} \quad (2)$$

, $f_{(m,n)}$: 픽셀의 명도
T : 명도 임계값

다. 명도히스토그램

명도 히스토그램(bright histogram) 특징은 전 처리된 영상을 2, 4, 8, 16, 32, 256 레벨의 그레이 스케일 히스토그램으로 표현한 것으로 원 영상이 칼라 또는 흑백인 경우 방대한 이미지 정보를 단순화 시켜 빠르고 간편한 패턴 매칭을 처리할 수 있게 한다[5].

$$h(t,b) = \sum_{n=0}^{N-1} 1, \quad if \ In(t) \in Lb$$

$$Lb = \{In(t) \mid b \le In(t) < b+1\}, \ b : brightness \quad (3)$$

라. 색상 히스토그램

이는 영상에서 제한된 N개의 색상에 대한 히스토그램을 구하는 방법이다. 명도 히스토그램과 유사하게 N 색상 히스토그램을 구하게 된다. 두개의 영상이 서로 다른 경우 두 이미지 사이의 색상 상관관계가 상대적으로 약하게 나타나기 때문에, 이를 영상의 패턴 매칭에 이용할 수 있다. 두 영상의 색상 상관계수를 구하는 식은 다음과 같다[5].

$$h(t,\ c) = \sum_{n=0}^{N-1} 1, \quad if\ \ Co\ (t) \in Lc$$
$$Lc\ = \{Co(t)\ \ | c \le Co(t)\ \ < c + 1\}, \quad c : color \tag{4}$$

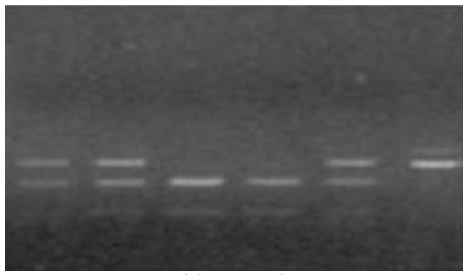
III. 시스템 설계

3.1 띠 무늬 이미지 특징 분석

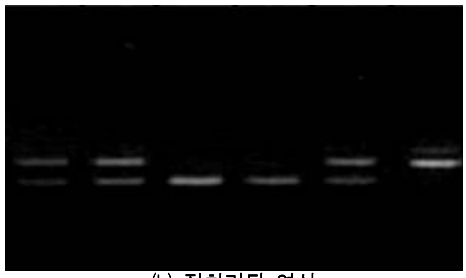
유전자 진단에 사용되는 DNA 이미지는 영상학적인 특징이 존재한다. 각각의 개체는 대나무 줄기와 같은 하나의 막대 이미지로 대응되며 각 막대는 마디처럼 보이는 몇 개의 DNA 정보를 가지게 되는데, 각 마디를 띠 무늬(band pattern)라고 부른다. 이들 띠 무늬 이미지는 명도차가 있는 색상과 위치 정보와 같은 몇 가지 특징을 갖게 된다. 본 연구에서 사용한 이미지는 컴퓨터에 저장된 그레이스케일 단계의 이미지로서 흰색계열의 마디 개수와 마디 위치가 비교적 뚜렷하게 나타난다. 본 논문에서는 영상에서 띠 무늬가 나타난 개수와 위치 정보를 이미지의 특징으로 사용하였다.

3.2 전처리 및 영상 분할

본 연구에서는 처리 속도의 향상을 위해 입력된 이미지에 대해 전처리 과정을 실행하였다. 전처리과정에서는 이미지의 밝기 값과 대비 값을 재계산하여 배경이나 잡음으로부터 띠 무늬를 분리하였다. 그래서 보다 빠르고 정확하게 띠 무늬의 존재 여부를 계산할 수 있게 하였다. 다음의 그림3은 DNA 이미지 원영상과 전 처리 후의 이미지를 보인다.



(a) 원 영상



(b) 전처리된 영상

그림 3. DNA 이미지

Fig. 3 DNA Image

영상 분할 과정에서는 각 개체에 대응되는 DNA 이미지로 세로로 분할을 하였다. 세로로 분할된 이미지에 대해 띠 무늬의 개수와 위치값을 효율적으로 계산하기 위해 동일 간격으로 가로 분할을 수행하였다. 그림4는 영상 분할된 DNA 이미지를 보인다.

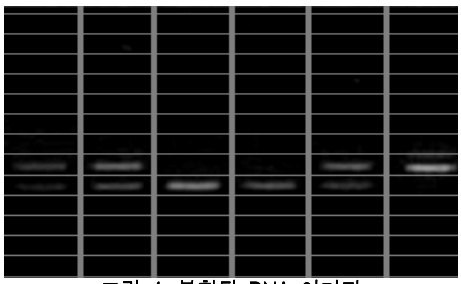


그림 4. 분할된 DNA 이미지

Fig. 4 Split DNA Image

3.3 식별과정

식별과정에서는 우선 각각의 개체에 대한 특징값을 계산하게 된다. 본 논문에서 제시한 특징값인 띠 무늬의 개수와 각 위치값을 계산하여 저장한다. 그 후 한 개체마다 다른 나머지 개체들과의 유사도 검사를 실행한다. 패턴간의 비교 결과로 유사도가 계산되면, 유사도가 가장 높은 패턴을 각 개체별 검색 결과로 보여주게 된다.

IV. 시스템 구현

4.1 영상처리 및 특징 추출

본 연구에서는 사전에 디지털화 되어 컴퓨터에 입력되어 있는 DNA 이미지를 획득하여 프로그램의 입력 영상으로 사용하였다. 원 영상에서 띠 무늬들이 선명하게 나타나게 하기 위해서 이미지의 밝기 값과 대비 값을 재계산하여 노이즈를 제거하였다.

그림 5. 영상처리 프로그램
Fig. 5 Image Processing program

그림 5는 이와 같은 프로그램 실행결과를 보이고 있다. 그림 5의 실행 화면에서 좌측에 크게 보이는 모양이 읽어 들인 DNA 영상이며, 우측에 흑백으로 보이는 영상이 전처리를 실행한 결과이다. 입력된 영상보다 배경 잡음이 상당히 소멸되어 띠 무늬들이 선명히 나타난 것을 알 수 있다.

4.2 개체간의 유사도 측정

개체간의 유사도 계산은 다음의 과정을 거친다. 전처리된 영상은 띠 무늬 판별을 쉽게 하기 위하여 영역 분할이 이루어진다. 영역 분할은 2단계로 이루어지는데, 우선 각 개체에 해당되는 영역인 세로 막대 형태로 분할을 한다. 그 후 각각의 세로 막대 영역을 다시 작은 직사각형 형태로 분할한다. 그 후 분할된 특정 직사각형 영역에 띠 무늬가 존재하는지 여부를 판별한다.

이는 하나의 직사각형 영역에서 흰색 부분의 면적이 주어진 임계값 보다 크게 되면 띠 무늬가 존재하는 것으로 판단하였다. 이와 같은 과정을 모든 영역에 대해 처리하여 생성된 띠 무늬의 개수와 위치 정보를 2차원 배열에 저장한다. 최종적으로 생성된 2차원 배열에서 각 개체별 유사도를 계산하여 각 개체별로 유사도 정보를 생성하게 된다.

본 논문에서 사용된 개체들은 한우 암소와 젖소의 DNA이며 표 1에서는 이에 대한 유사도 결과를 보이고 있다. 두 개체가 완전히 일치하는 경우 최고값 10을 갖게 된다.

표 1. 개체별 유사도 결과
Table 1. Similarity results of the objects

	1	2	3	4	5	6
1		7	2	3	8	2
2	7		2	3	9	2
3	2	2		8	3	3
4	3	3	8		3	3
5	8	9	3	3		2
6	2	2	3	3	2	

V. 결 론

본 논문에서는 DNA 입력 패턴에 대해 방대한 관련 표준 패턴들에서 유사도가 가장 높은 표준 패턴을 찾아낼 수 있는 영상처리 기반 검색 기술에 대해 연구하였다. 이를 위해 잡음과 배경으로부터 영상을 분리하는 전처리를 수행하였다. 그 후 전처리된 이미지에서 특징이 추출되고 추출된 특징을 이용하여 표준패턴과의 패턴매칭을 실행하였다.

판별에 사용된 특징은 흰색 마디의 위치 정보와 개수이다. 이것을 사용한 이유는 DNA 이미지들을 구분할 수 있는 중요한 특징이 띠 무늬이기 때문이다. 각각의 띠 무늬 정보는 2차원 배열 형태로 저장이 되며, 최종적으로 생성된 2차원 배열 정보를 기초로 각 개체별 유사도를 계산하여 각 개체별로 유사도 정보를 생성하였다.

참고문헌

[1] 강희규, 이진종, 정병균, 주세익, *진단 분자생물학*, 현문사, pp. 370-412, 8. 2011.

[2] 전문진, *현대의 생물공학과 생물산업*, 아카데미서적, pp. 70-78, 8. 2003.

[3] 한국분자생물학회, *분자생물학*, 아카데미서적, pp. 3-15, 9. 1997.

[4] 김기원, 김봉기, "*해태 생산라인에서의 실시간 시각검사 시스템*", 한국해양정보통신학회논문지, 제11권 6호, pp1136-1140, 2007.

[5] 신종홍, 장선본, 지인호, *디지털 영상처리 입문*, 한빛미디어, pp.46-61. 7. 2008.

[6] 양원근, 조아영, 정동석, "*영상 식별을 위한 전역 특징 추출 기술과 그 성능 비교*", 멀티미디어학회논문지, 제 14권 제 1호, pp.1-14. 1. 2011.

[7] Zin-suk Kim, *Conservation and Manipulation of Genetic Resources*, pp. 230-243, 7. 1995.

저자소개

김기원(Gi-Weon Kim)

1989년 숭실대학교 전자계산학과 (공학석사)
2001년 한남대학교 컴퓨터공학과 (공학박사)

1996년 ~ 현재 초당대학교 컴퓨터과과 교수
※ 관심분야 : 멀티미디어, 실시간영상처리, 음성인식

정진영(Jin-Young Jung)

1994년 한남대학교 컴퓨터공학과 (공학석사)
2002년 한남대학교 컴퓨터공학과 (공학박사)

1997년 ~ 현재 대전보건대학교 바이오정보과 교수
※ 관심분야 : 바이오인포메틱스, 웹기반정보시스템, 스마트폰 프로그래밍

김석훈(Seok-Hun Kim)

2003년 한남대학교 컴퓨터공학과 (공학석사)
2006년 한남대학교 컴퓨터공학과 (공학박사)

2007년 ~ 2010년 (주)파라곤베이스
2010년 ~ 현재 타임시스템(주)
※ 관심분야 : VoIP, 모바일컴퓨팅, 웹데이터베이스