유전자 및 물질의 젤 영상 분석에 관한 연구

김영원†, 전병환††
† 공주대학교 컴퓨터공학과, †† 공주대학교 정보통신공학부
전화 : 041-850-8524

A Study on the Analysis of Gel Images of Genes and Molecules

Young Won Kim†, Byung Hwan Jun††
† Dept. of Computer Engineering, Kongju National University
†† Division of Information and Communication Engineering, Kongju National University
E-mail: pattern99@hanmail.net

Abstract

With all the researches to define human genom and to look for some new bio-activated material in the bio-technology field recently, it is more highly needed to analyse DNA or so called Material than ever before.

First, the lanes are extracted based on histogram analysis and projection technique. And then three other approaches are applied for band extraction: SB, RG-1, and RG-2. In SB method, a search line is set dividing each lane equally and vertically to find peaks and valleys. And according to them, minimum enclosing rectangle of each band is determined. In RG-1 approach, on the other hand, band areas are extracted by region growing with the peaks as seeds, avoiding the overlap with the neighboring bands. In RG-2 approach, peaks and valleys are searched in two lines that trisect the lane vertically, and the pair of peaks in the same band are determined, and then used to grow the region.

To compare the accuracy of the three suggested methods, we measure the location and amount of bands. The result shows that the mean deviation of the location is 0.06, 0.03, and 0.01 for SB, RG-1, and RG-2 respectively. And the mean deviation of the amount of bands is 0.08, 0.05, and 0.02 for SB, RG-1, and RG-2 respectively. In conclusion, the RG-2 method suggested in this paper appears to be the most reliable on the degree of the accuracy in measuring the location and amount of bands.

※ 본 논문은 한국과학재단 지정 지역협력 연구센터인 공주대학교 자원재활용 신소재 연구센터의 연구비 지원에 의해 연구되었음.

Ⅰ. 서 론

최근 새로운 생리활성물질을 찾기 위한 연구들이 활발해지면서 물질의 분리, 정제 및 합성기술의 발달로 인해 대상 후보물질들의 종류와 수가 급격히 늘어나고 있다. 반면 생리활성물질의 분석을 위한 측정방법은 전기영동에 의한 젤 영상을 보고 밴드의 유무를 판단하는 이진법적인 판별이 주로 사용되고 있다. 기존의 영상처리 방법에서는 유전밴드의 유무를 판단하기 위해서 밴드 사이의 배경이 일정할 경우 미분을 사용하고, 배경이 일정하지 않은 경우 곡률을 사용하여 밴드의 위치를 분석하였다[1]. 개발된 분석 소프트웨어들도 특정형태의 정형화된 활성측정결과에는 정량 측정 가능했지만, 밴드 추출 기술에 있어서는 정밀도가 낮아 정확한 정량측정이 어려웠다.

본 논문에서는 디지털 영상처리 기법 중에서 히스토그램 분석과 프로젝션 기법[2,3], 영역성장방법[4,5,6]을 사용하여 젤 영상을 분석하는 방법을 제안한다.

Ⅱ. 레인 및 밴드 추출

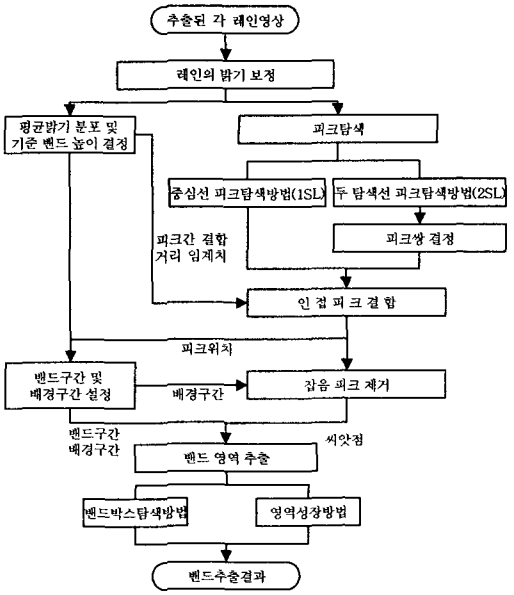
젤 영상은 일반적으로 여러 레인과 각 레인내에 여러 밴드를 포함하는 형태로 어두운 배경에 하얀색의 밴드를 갖는다. 영상의 분석은 먼저 이를 각각의 레인으로 분할하는 과정을 거친 후 각 레인에 대해서 밴드를 추출하는 과정으로 이루어진다.

1. 레인(lane)추출

젤 영상에서 배경과 레인을 구분하는 임계치를 찾기 위해 영상 히스토그램과 엔트로피를 사용하였다. 찾아진 임계치로 영상을 이진화하고 수직으로 프로젝션시킨다. 이때 '골(valley)-마루(peak)-골(valley)'의 형태를 띈 영역을 레인으로 추출한다. 수직프로젝션에 의해 추출된 각 레인의 평균 레인폭을 구하고, 평균 레인폭의 70%를 레인추출을 위한 임계치로 정하여, 노이즈에 의해 추출된 레인을 제외한 실제 레인을 추출한다.

2. 밴드(band) 추출

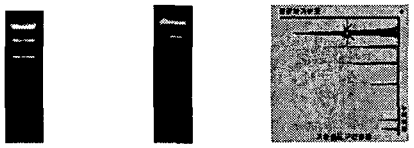
밴드 추출과정은 [그림 1]과 같다.



[그림 1] 밴드 추출 과정

1) 레인의 밝기 보정

[그림 2]의 (a)와 같이 추출된 각 레인영상에서 좌우 배경 평균밝기를 구한 후 평균밝기만큼 뺀 밝기로 영상을 변환하면 [그림 2]의 (b)와 같다.



(a) 레인영상 (b) 밝기 보정영상 (c) 평균밝기 분포

[그림 2] 레인 영상의 평균밝기 분포

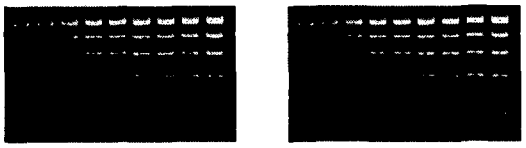
2) 평균 밝기 분포 결정

밝기가 보정된 레인의 수직위치에서 평균밝기는 [그림 2]의 (c)와 같다. 이때, 최대 평균밝기에 대해 50%에서 형성되는 최대 높이를 기준 밴드 높이로 사용한다.

3) 피크 탐색

각 레인에 대해서 레인을 이등분하는 하나의 탐색선(1SL) 또는 레인을 삼등분하는 두 개의 탐색선(2SL)을 결정하고, 화소 밝기값의 변화에 따라 피크 및 밸리를 찾는다. 두 탐색선에서 찾아진 피크들에 대해서는 수직위치의 유사성에 근거하여 동일한 밴드에 속하는 피크쌍을 결정한다.

[그림 3]은 피크 추출 결과영상이다.



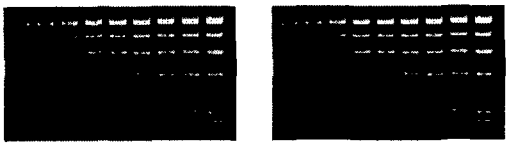
(a) 두 탐색선 피크 추출 (b) 중심 탐색선 피크 추출

[그림 3] 피크 추출 결과

4) 인접 피크 결합

먼저 수직으로 이웃하는 피크들간의 거리를 구하고 그 중 최소값이 주어진 임계치보다 작으면 밝기가 작은 피크를 제거하는 과정을 반복하여 지나치게 인접하는 피크들을 제거한다. 이때, 거리 임계치는 기준 밴드 높이의 80%로 정한다. 또한, 각 밸리마다 상부와 하부 피크와의 밝기차를 구하고 그 중 큰 값의 30%를 밝기 임계치로 정한다. 만일 작은 밝기차가 임계치 미만이면 낮은 피크와 밸리를 제거한다.

[그림 4]는 피크 결합된 결과영상이다.



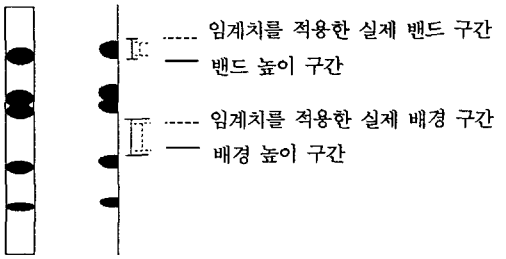
(a) 두 탐색선 피크 결합 (b) 중심 탐색선 피크 결합

[그림 4] 피크 결합

5) 밴드 구간 및 배경구간 설정

추출된 각 피크의 위치에서 평균 밝기의 지역최대(local maximum)값을 찾고, 이 밝기값의 10%를 임계치로 설정하여 밴드구간 및 배경구간을 구분한다.

[그림 5]는 밴드 구간 및 배경구간 설정 예이다.



[그림 5] 밴드 구간 및 배경구간 설정 예

6) 잡음 피크 제거

만일 피크의 위치가 배경구간에 속할 경우 잡음 피크로 간주하여 제거한다. 이와같이 추출된 피크들은 밴드 영역 성장의 씨앗점으로 사용된다.

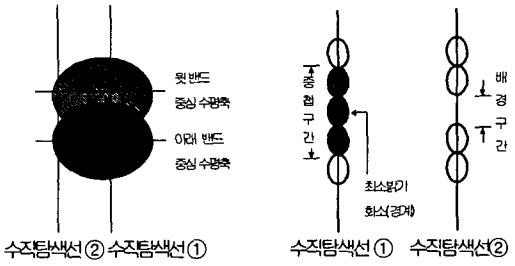
7) 밴드 영역 추출

레인내 탐색선의 수와 밴드 영역의 추출 방법에 따라, 하나의 탐색선과 밴드 박스 탐색을 채택한 SB방법, 하나의 탐색선과 영역성장기법을 채택한 RG-1방법, 두 탐색선과 영역성장기법을 채택한 RG-2방법을 제안한다. 이때, 각 밴드에 대해 윗 배경의 평균 밝기와 아래 배경의 평균 밝기를 구하고 그 중에 큰 값을 밴드간 기준 배경 밝기로 정한다. 밴드의 최대 평균 밝기와 기준 배경의 밝기와의 차이를 구하여, 그 값의 15%를 기준 배경 밝기에 더해서 밴드 영역 추출을 위한 임계치로 사용하였다.

8) 밴드간 중첩영역 분할

영역 성장의 경우 분자량이 유사하여 밴드간 중첩영역이 발생할 수 있다. 밴드간 중첩을 분할하기 위해서는 각 밴드의 피크를 수평축으로 정하고 수평축 사이의 중첩된 밴드영역을 수직으로 탐색하면서, 최소 밝기점을 상하밴드 영역의 경계로 한다.

[그림 6]은 상하 밴드가 중첩인 경우 분할하는 방법이다.



[그림 6] 밴드 중첩 분할

Ⅲ. 실험 및 결과 분석

1. 실험환경

실험을 위해서 단백질과 DNA시료 4가지를 전기영동한 총 26장의 영상을 사용한다. 이 때 정확한 밴드의 위치 및 밴드양을 측정하기 위해, 레인내에 투여하는 물질의 양을 일정하게 증가시킨 후, 전기 영동 시간에 따라 다양한 영상을 취득하였다.

2. 평가기준

앞서 제안한 SB방법, RG-1방법, RG-2방법이 밴드의 위치 및 양을 얼마나 정확하게 측정하는지 평가하기 위해 단순평가방법 및 정교평가방법을 사용한다.

밴드 위치의 단순평가는 총 밴드수에 대한 레인내 밴드수 오차의 합의 비율(R_p)을 사용하고, 밴드 양의 단순평가는 총 밴드양 증가 횟수에 대한 관찰된 밴드양 증가의 총 오류 횟수 비율(R_q)을 사용한다.

$$- R_p = \frac{N_p}{N_B \times N_L} \quad (1)$$

$$- R_q = \frac{N_q}{N_B \times (N_L - 1)} \quad (2)$$

$$N_p = \sum_{i=1}^{N_L} | N_B - N_{Bi} |, \quad N_q = \sum_{j=1}^{N_B} \sum_{i=1}^{N_L - 1} f(i,j)$$

$$f(i,j) = \begin{cases} 0 & q_{ij} < q_{i+1,j} \\ 1 & q_{ij} \geq q_{i+1,j} \end{cases}, \quad N_L = \text{레인수}$$

N_B = 레인당 밴드수

N_{Bi} = i번째 레인에서 추출된 밴드수

i = 레인번호, j = 밴드번호

밴드 위치의 정교평가는 첫 번째 밴드의 중심에서 마지막 밴드의 중심까지의 수직 거리를 1로 정규화 한 후, 레인간에 대응하는 각 밴드 위치에 대한 평균편차(Δy_j)를 사용한다. 또한 밴드 양의 정교평가는 각 레인내 모든 밴드 양의 합을 1로 정규화를 한 후, 레인간에 대응하는 밴드 양의 평균편차(Δq_j)를 사용한다.

$$- \Delta y_j = \sum_{i=1}^{N_L} \frac{| y_j - y_{ij} |}{N_L} \quad (3)$$

$$- \Delta q_j = \sum_{i=1}^{N_L - 1} \frac{| q_j - q_{ij} |}{N_L} \quad (4)$$

$$y_j = \sum_{i=1}^{N_L} \frac{y_{ij}}{N_L}, \quad q_j = \sum_{i=1}^{N_L - 1} \frac{q_{ij}}{N_L}$$

y_j = j번째 밴드의 정규화된 대표 위치

y_{ij} = I번째 레인의 j번째 정규화된 위치

$$0.0 \leq y_{ij} \leq 1.0$$

q_j = j번째 밴드의 정규화된 밴드 양

q_{ij} = I번째 레인의 j번째 밴드의 정규화된 양

$$0.0 \leq q_{ij} \leq 1.0$$

3. 실험결과

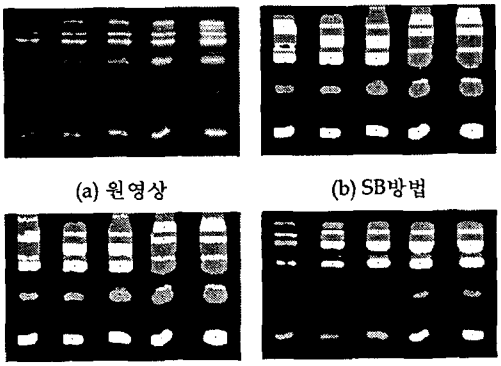
실험결과 추출된 밴드에 대해서 단순평가와 정교평가 방법으로 결과를 분석하였다. 결과는 [표1]과 같다.

[표 1] 밴드 추출 평가 결과

평가기준 \ 분석방법		SB	RG-1	RG-2
단순평가	밴드 수의 오류율 (R_p)	13%	10%	7%
	밴드 양증가의 오류율(R_q)	15%	11%	8%
정교평가	밴드위치의 평균 편차 (Δy_j)	0.06	0.03	0.01
	밴드 양의 평균 편차 (Δq_j)	0.08	0.05	0.02

이때 정교평가는 단순평가에 의해서 밴드 수가 정확하게 추출된 레인에 대해서만 평가하였다.

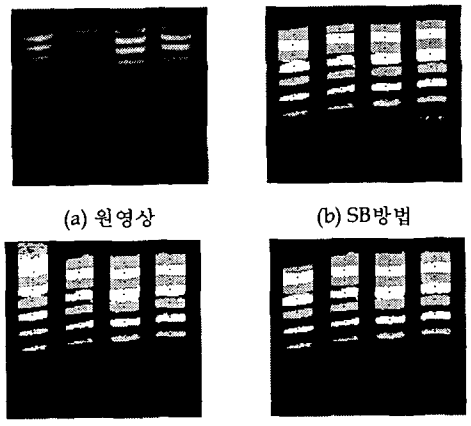
[그림 7] ~ [그림 9]는 본 논문에서 제안하는 세 가지 방법으로 밴드 영역을 추출한 결과의 예이다.



(a) 원영상 (b) SB방법

(c) RG-1방법 (d) RG-2 방법

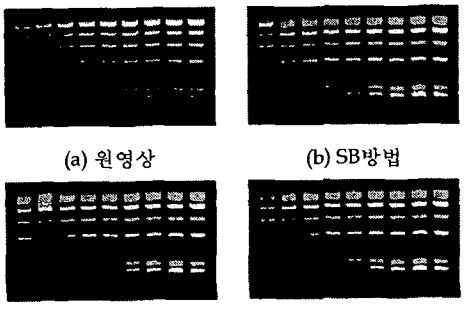
[그림7] SDS-7 시료



(a) 원영상 (b) SB방법

(c) RG-1방법 (d) RG-2 방법

[그림8] 100bp ladder 시료



(a) 원영상 (b) SB방법

(c) RG-1방법 (d) RG-2 방법

[그림 9] Lambda HindIII Digest 시료

Ⅳ. 결 론

본 논문에서는 유전자 및 물질의 정보추출을 위한 젤 영상 분석 시스템을 구현하였다.

각 레인에서 밴드를 추출하기 위해서는 하나의 탐색선과 밴드박스탐색을 결합한 SB방법과 하나의 탐색선과 밴드영역성장방법을 결합한 RG-1방법, 두 개의 탐색선과 밴드영역성장방법을 결합한 RG-2방법을 사용하였다. 이 세 가지 방법으로 추출된 밴드의 위치와 밴드의 양을 단순평가와 정교평가로 비교 실험한 결과, 본 논문에서 제안하는 RG-2방법이 두 가지 평가에서 모두 가장 정확한 밴드의 위치와 양을 추출하는 것으로 나타났다. 특히 밴드의 모양이 아령모양이나 가운데가 끊어진 형태의 밴드도 추출하는 것으로 확인되었다.

향후 연구과제로는 레인을 추출하기 위해서 히스토그램에서 배경영역과 레인영역을 구분하는 임계치를 자동으로 설정하는 부분이 필요하고, 밴드영역 추출에서도 배경과 밴드의 화소 밝기차이가 매우 작을 경우도 추출할 수 있도록 적응적인 임계치 설정방법에 대한 연구가 필요하다. 또한 시스템의 성능을 평가하고 보완하기 위해서 다양한 물질에 대한 방대한 데이터가 필요하다.

[참 고 문 헌]

[1] 황덕인, 공성곤, 조성원, 조동섭, 이승환, "영상의 지형적 특징에 의한 유전밴드 인식", *정보과학회논문지(B)*, 제26권, 제11호, pp.1350-1358, 1999.

[2] 최형일, 이근수, 이양원, *영상처리 이론과 실제*, 홍릉과학출판사, 1998.

[3] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*, Addison Wesley, 1993.

[4] R. Adams and L. Bischof, "Seeded Region Growing," IEEE Trans. on Pattern Analysis and Machine Intell, vol. 16, pp.641-647, 1994.

[5] 신민수, 김휘용, 김성대, "반자동 영역분할을 위한 씨앗점 검출과 연결 알고리즘", *한국통신학회 추계종합 학술발표회 논문집*, pp.1567-1570, 1999.

[6] S. W. Zucker, "Region Growing: Childhood and Adolescence," Comp. *Graphics Image Process.*, vol. 5,