2차원 전기영동 영상에서 잡영을 제거하기 위한 적응적인 문턱값 결정

Adaptive thresholding for eliminating noises in 2-DE image

최관덕* · 김미애** · 윤영우**
Kwan-Deok Choi* · Mi-Ae Kim** · Young-Woo Yoon**

요 약

2차원 전기영동 영상 분석 프로그램의 반점 검출 단계에서 해결해야할 문제점 중에 하나는 잡영 제거의 문제이다. 전처리과정에서 처리되지 않고 남은 잡영은 영역분할 결과 과분할되는 문제를 낳는다. 과분할된 배경 영역을 구분하고 제외시키기 위해서 일정한 밝기 이상의 영역을 제거하는 고정 문턱값을 사용하여 영역을 제거하면, 육안으로는 보이지 않으나 중요한 기능을 하는 미량의 단백질을 나타내는 반점들이 제외될 수도 있다. 제안 기법은 영역분할 후에 영역들의 첨도의 평균 곡선을 지수함수에 회귀분석하여 매개변수를 구한 다음, 오차의 확률분포에 따라서 매개변수들로 문턱값을 구하여 적용한다. 오차의 확률분포에 따르면 문턱값 적용의 신뢰도는 99.85%이며, 제안 기법을 실험 영상으로 실험한 결과로써 적응적 문턱값 결정 기법이 정확함을 보인다.

Abstract

One of the problems for implementing the spot detection phase in the 2-DE gel image analysis program is the eliminating noises in the image. Remained noises after the preprocessing phase cause the over-segmented regions by the segmentation phase. To identify and exclude the over-segmented background regions, if we use the fixed thresholding method that is choosing an intensity value for the threshold, the spots that is invisible by the eyes but mean a very small amount proteins which have important role in the biological samples could be eliminated. This paper propose an adaptive thresholding method that come from an idea that is got on statistical analysing for the prominences of the peaks. The adaptive thresholding method works as following. Firstly we calculate an average prominence value curve and fit it to exponential function curve, as a result we get parameters for the exponential function. And then we calculate a threshold value by using the parameters and probability distribution of errors. Lastly we apply the threshold value to the region for determining the region is a noise or not. According to the probability distribution of errors, the reliability is 99.85% and we show the correctness of the proposed method by representing experiment results.

Keywords : 2-DE, Spot detection, Eliminating noises, Adaptive thresholding, Statistical analysis

I. 서 론

2차원 전기영동(2-DE: 2-Dimensional Electrophoresis)은 분자 생물학 분야의 중요한 도구로서 생물학적 표본내의 단백질을 종류별로 분리하는 기법이며, 독물학, 임상 화학, 암 연구, 알츠하이머 연구 등 많은 분야에서 사용된다. 2-DE는 두 단계의 생화학적 분리를 차례로 수행한다.

* 대구과학대학 ** 영남대학교
논문 번호 : 2007-4-29 접수 일자 : 2007. 11. 29
심사 완료 : 2008. 1. 4
* 본 논문은 2004학년도 영남대학교 정보통신연구소 학술연구조성비지원으로 수행되었습니다..

첫 번째 단계는 등전위 초점(iso-electric focusing; IEF) 전기영동이라 하는데, 표본을 고정화 pH 구배(IPG: Immobilized pH Gradient) 겔(gel) 막대 위에 놓고 전기장을 가하면 표본내의 단백질들은 고유의 등전점(iso-electric point; pI)에 따라 수평으로 이동하여 분리된다. 두 번째 단계는 SDS-PAGE(Sodium Dodecyl Sulfate Polyacrylamide Gel Electrophoresis)라 하는데, 첫 번째 단계의 결과에 전기장을 가하면 단백질들은 각각의 분자량(molecular weight; Mr)에 따라 수직으로 이동하여 분리된다. 첫 번째 단계를 1차원 전기영동이라 하고, 두 번째 단계를 2차원 전기영동이라 하며, 두 번째 단계 처리 이후에 겔은 염색 과정을 거친다[1,2].

2-DE 처리 결과는 2차원 반점(spot) 패턴으로 나타나며, 패턴 내의 각각의 반점은 개별적인 단백질을 의미한다. 따라서 반점을 단백질로 동정(同定: identify)하는 과정을 거쳐서 필요한 분야에 이용한다. 반점을 동정하는 방법으로는 반점을 로봇으로 추출한 후 질량분광기를 이용하는 방법과 이미 분석된 2-DE 결과와 새로운 결과를 서로 시각적으로 교차 비교하여 동정하는 방법이 있다. 전자의 경우, 정밀분석이 가능하나 한번에 하나의 반점밖에 동정 못하므로 작업시간이 많이 걸린다. 후자의 경우는 단백질 반점 패턴의 복잡성 때문에 육안 분석이 매우 어려워서 컴퓨터로써 자동적으로 동정하기 위한 연구가 일찍부터 수행되어 왔다. 컴퓨터로써 자동적으로 동정하기 위해서는 반점 패턴을 고해상도의 스캐너나 레이저 광도계 등을 사용해서 디지털 영상화하고, 디지털화된 겔 영상을 겔 영상 분석 프로그램으로 분석해야한다[1,2].

겔 영상 분석 프로그램은 반점 검출 단계와 겔 정합 단계로 구성된다. 반점 검출 단계는 겔 영상 내의 반점들을 각각의 영역으로 분할하는 단계이며, 겔 정합 단계는 전 단계에서 검출된 반점들을 이미 분석된 겔 영상의 반점들과 교차 비교하여 실제의 단백질들로 동정하는 단계이다.

반점 검출 단계는 영상분할 알고리즘을 수행해서 반점의 영역들을 분할해 낸 다음, 반점의 형태적 모형을 적용해서 반점인지 아닌지를 판단하는 흐름으로 구현한다. 따라서 반점 검출 단계의 구현은 적합한 영상분할 기법의 고안과 반점의 형태적 모형을 정하는 것이 핵심이다. 한편 영상분할 알고리즘으로 영역 분할을 수행할 때에는 배경(즉, 겔)과 전경(즉, 단백질 반점)을 분리해 내야만 하는데, 겔 영상 내의 잡영이 이것을 어렵게 만든다. 따라서 영상분할의 전처리로서 잡영을 제거하는 기법도 필요하다.

영상의 백색잡영(white noise)은 중간값 필터 등을 이용하면 제거할 수 있으며, 비백색 잡영은 top-hat 등의 수리형태적 필터를 이용하면 어느 정도 제거할 수 있다. 하지만 top-hat 필터 처리를 수행해도 구조화단위(structuring element)보다 작은 범위의 비백색잡영은 제거할 수가 없게 된다. 이것은 배경이 불균질하기 때문에 생기는 문제이며, 이렇게 처리되지 않고 남은 잡영은 분수령 알고리즘과 같은 영역분할 알고리즘 수행 시에 영역으로 분할되어, 영역분할 수행 결과 영역이 과분할되는 문제를 만든다. 따라서 과분할된 영역 중에서 배경이라고 판단되는 영역을 제외시키는 방법이 필요하다[1].

과분할된 배경 영역을 구분하고 제외시키기 위해서 일정한 밝기 이상의 영역을 제거하는 고정 문턱값을 사용한다면, 대다수의 겔 영상에 존재하는 미량의 단백질을 나타내는 육안으로는 잘 보이지 않는 반점들이 제거될 가능성이 많다. 단백질은 생명현상에 필요한 수백 가지 화학반응을 촉매하는 효소로서도 작용하므로 적은 양일지라도 중요한 역할을 한다. 따라서 고정 문턱값 사용은 적합하지 않다. 본 논문에서는 영상에 따라 문턱값을 자동적으로 계산하는 기법인 적응적인 문턱값 결정 기법을 제안한다.

본 논문의 구성은 다음과 같다. Ⅱ장에서 2차원 전기영동 영상의 잡영 제거와 관련된 연구를 살펴본다. Ⅲ장에서는 본 논문의 제안기법인 잡영 제거를 위한 적응적 문턱값 결정기법에 대해서 자세히 기술하며, Ⅳ장에서 제안기법의 실험 영상에 대한 실험과 결과 분석을 기술하며, Ⅴ장에서 결론을 맺는다.

Ⅱ. 관련 연구

겔 영상은 장시간에 걸친 복잡한 처리과정을 통해서 제작된다. 따라서 영상에는 단백질 반점과 배경(겔 자체) 외에도 잡영이 많이 있다. 겔 영상에 나타나는 잡영의 원인으로는 디지털 영상을 제작할 때에 사용된 스캐너 등의 입력장치에 의한 것(black specs), IPG 스트립 상의 염의 존재나 핵산의 오염에 의한 것(streaks), 미량의 단백질이 아닌 요소 등이 있다. 특히 미량의 단백질이 아닌 요소들이 배경을 불균질하게 만드는 원인이 되며, 잡영 제거를 더욱 어렵게 만든다[3].

대부분의 겔 영상 분석 프로그램은 잡영의 영향을 줄이기 위해서 반점 분할 단계에 앞서 전처리를 수행한다. 전처리 기법으로는 평활화 기법, 배경 보상 기법 등이 있다. 평활화 기법으로는 지역 가우시안 필터, 확산 필터, 다항 콘볼루션 필터, 지역 중간값 필터 등이 있다. 배경 보상(background subtraction) 기법은 겔 배경의 강도 레벨의 미소한 변화를 제거하기 위한 기법이다. 간단한 배경 보상 기법은 배경 내의 가장 밝은 점과 가장 어두운 점을 알아내서 전체 배경을 평균 강도로 대치하는 것이다. Tyson과 Haralick이 배경의 약화(depression)를 나타내는 국소 극소점(local minima)을 발견하여 이들 극소점 사이의 배경을 보간하는 기법을 보여준다. Melanie Ⅱ에서는 모든 픽셀 값들에서 최소 강도를 빼고, 반점을 제외한 배경 영상에 3차 다항식을 맞추어 배경을 보상한다. 다른 기법은 3차원 수리형태학에서 온 것인데, 그레이스케일 영상을 반구형 구조화 단위로써 오프닝과 클로징 연산을 수행하여 배경을 보상하는 기법인 top-hat 필터가 있다[1].

전처리 후에도 남은 잡영은 다음 단계인 반점 분할 단계에 영향을 준다. 현재 반점 분할을 위한 기법으로는 잡영에 대한 강건성 때문에 분수령 알고리즘이 주로 사용된다. 하지만 분수령 알고리즘의 가장 큰 단점중의 하나는 (국소점으로 오인되는) 잡영에 의해서 영역이 과분할되는 것이다. 과분할 문제를 해결하기 위한 기법으로는 분수령 알고리즘 수행 전에 해결하는 기법과 분수령 알고리즘 수행 후에 해결하는 기법이 있다.

마커 제어 분수령(marker controlled watersheds) 기법[4]은 반점 형태 분석 등을 수행하여 선택한 국소점(marker)만을 사용하여 분수령 알고리즘을 수행하여 과

분할을 막는다.

Pleissner[5]는 일차 미분 영상에 대해서 영역 병합(region merging) 분수령 기법을 수행한다. 이 기법은 원 영상의 일차 미분 영상에 대해서 분수령 알고리즘을 수행한 다음, 분할된 영역들이 배경 영역(즉 잡영)인지 반점 영역인지를 구분하기 위해서 두 가지 문턱값을 사용한다. 하나는 반점 영역의 평균 강도가 그것의 (배경) 이웃 영역 중에 적어도 하나 보다는 실질적으로 높을 것이라고 간주하여 정한 문턱값이고, 다른 하나는 다른 반점 영역들과 동떨어진(land-locked) 영역일 경우에 적용되는 단순한 강도 문턱값이다. 그런 다음, 반점 영역이라고 결정된 반점(실제는 부분 반점)들을 타원 형태에 맞추어 병합한다.

본 연구에서 제안하는 적응적 문턱값 적용 기법은 분수령 수행 후에 잡영에 의한 영역을 제거하므로써 과분할 문제를 해결한다. 잡영의 판단 기준으로 기존의 기법들이 고정된 문턱값이나 직관에 가까운 문턱값을 정하는 반면에 본 연구의 제안 기법은 영상에 적응적이며 통계에 의한 계산 방법으로 문턱값을 정한다.

III. 반점 검출 처리 시스템 및 제안 기법

본 논문에서 구현한 반점 검출 처리 시스템은 그림 1과 같다. 입력된 겔 영상은 전처리 과정을 거친 후, 분수령 영상분할을 거쳐 영역으로 분할되며, 분할된 영역에 제안 기법인 적응적 문턱값 결정을 적용한다. 문턱값을 넘은 영역들은 반점 후보로 결정되며 이후의 단계인 반점 정합 단계로 넘겨진다.

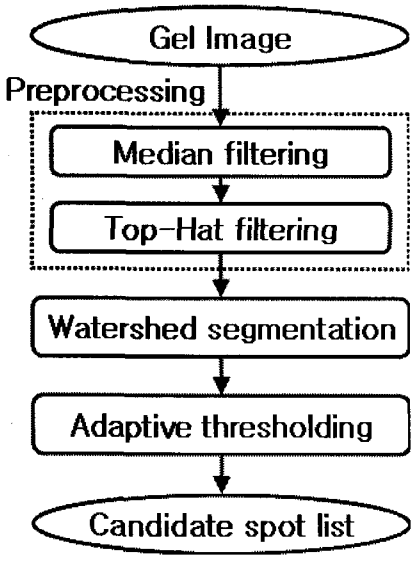


그림 1. 반점 검출 처리 시스템

Fig. 1. Spot detection processing system

3.1 전처리

겔 영상의 입력 형식은 BMP 형식이며, 입력된 겔 영상은 256 레벨의 그레이스케일 영상으로 변환하여 처리한다.

전처리 단계는 잡영을 제거하는 단계이다. 잡영은 백색잡영과 비백색잡영으로 구분할 수 있다. 백색잡영은 화소 단위로 독립적으로 발생되는 잡영이며, 잡영의 강도가 무작위적이다. 즉 모든 강도의 발생 확률이 동일하다. 비백색잡영은 강도가 인접 화소들과 연관된 잡영 즉, 화소의 군집으로 구성된 잡영이다.

본 논문에서는 백색잡영을 제거하기 위해서 중간값 필터를 사용하고, 비백색잡영을 제거하기 위해서 top-hat 필터를 사용한다. 비백색잡영은 잡영의 범위가 반점의 크기에 비해 매우 넓은 경우는 top-hat 처리로 제거되지만, 그 외의 경우에는 분수령 분할에 의해 영역으로 지정되는 문제점이 있다. 이것은 배경의 불균질성 문제로 인한 것이며, 이것을 해결하기 위한 제안 기법을 이후의 절에서 기술한다.

3.2 분수령 영상분할

겔 영상을 반점 영역별로 분할하기 위해서 분수령(watershed) 알고리즘을 사용한다. 분수령 알고리즘은 지형학 분야에서 연구된 것으로 영상을 픽셀 값을 높이로 하는 3차원 지형으로 간주하고 이 지형에 물을 채웠을 때 하나의 윤곽선으로 둘러 싸여진 웅덩이를 분할 영역으로 판단하여 영상을 분할하는 알고리즘이다.

본 논문에서는 실험 영상이 그레이스케일 영상이므로 각 픽셀의 밝기 값의 역상을 상응하는 지형적인 고도(topographic relief)로 간주하여 분수령 알고리즘을 적용하며, 분수령 알고리즘 중에서 시공간적으로 아주 효율적이라고 알려진 Vincent와 Soille[7]가 제안한 침수(immersion) 방식 분수령 알고리즘을 사용한다. 침수 방식은 범람(flooding) 방식이라고도 하는데 지형의 국소 극소점들에 구멍이 있다고 가정하고 물을 가장 낮은 곳부터 점차적으로 채워나가서 호수를 만든다. 호수에 물이 계속 차오르면 어느 시점에서 호수들은 꽉 차서 넘쳐날 지경이 되는데 이때 호수들 사이의 물이 만나는 지점에 댐을 쌓아서 물이 합쳐지지 못하도록 한다. 결과적으로 댐들을 연결한 것이 분수령이며 영역 분할의 경계선이 된다.

3.3 제안 기법 : 잡영 제거를 위한 적응적인 문턱값 결정

분수령 알고리즘의 수행에서 나온 각각의 분할 영역들을 각각의 반점 영역들로 간주할 수 있다. 하지만 top-hat 필터로 배경을 보상하여도 구조화 단위(structuring element)보다 작은 부분은 보상하지 못하므

로 분수령 알고리즘에 의해서 영역으로 분할되어서, 결과적으로 과분할 문제를 초래한다. 이러한 과분할 문제는 전체 반점 검출 알고리즘의 시공간적 비효율성을 초래하며, 또한 배경 영역에는 형태 정보가 부족하므로 이후의 단계에서 필요한 형태 분석 과정에서 오류 발생 가능성을 높인다. 따라서 영역 중에서 배경이라고 판단되는 영역을 구분하여 제외시키는 기법이 필요하다.

과분할된 배경 영역을 구분하고 제외시키기 위해서 일정한 밝기 이상의 영역을 제거하는 고정 문턱값을 사용한다면, 그림 2와 같은 대다수의 겔 영상에 존재하는 미량의 단백질을 나타내는 육안으로는 잘 보이지 않는 반점들을 제거할 수도 있는 가능성이 많으므로 적합하지 않다. 그림 2의 예는 밝기값의 역상으로 고도(범위:0~255)를 표현한 것인데, 최고치가 3인 미약한 영역이므로 고정 문턱값을 사용하면 제거될 가능성이 있으나 정밀 분석에 의해서 단백질로 밝혀진 반점 영역이다.

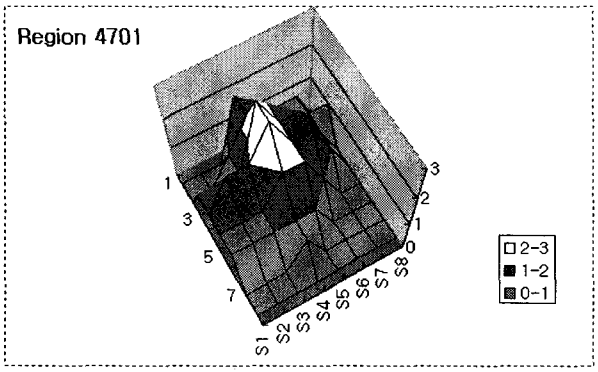


그림 2. 미약한 영역의 예
Fig. 2. An example of a very small region

본 논문에서는 영상에 따라 문턱값을 자동적으로 계산하는 기법인 적응적인 문턱값 결정 기법을 제안한다. 제안 기법은 봉우리의 삼차원적 형태를 구별하기 위한 형태 척도로서 첨도(prominence)를 제안하고, 봉우리의 첨도를 통계적으로 분석한 결과를 이용한 것으로, 영역분할 후에 영역들의 첨도의 평균 곡선을 지수함수로 회귀분석하여 지수함수의 매개변수를 구한 다음, 오차의 확률분포를 이용해서 매개변수들로 문턱값을 구한다. 오차의 확률분포에 따르면 문턱값 적용의 신뢰도는 99.85%이다.

3.3.1 첨도

높지만 넓은 불균질한 배경 영역으로부터 낮고 작아서 눈에 보이지 않지만 삼차원적 형태가 뚜렷한 산봉우리(peak)를 구별할 수 있는 형태 척도로서 영상 내 영역의 집합을 R이라 할 때 영역 r∈R의 첨도 p_r을 식 (1)과 같이 정의한다. 식 (1)에서 h_r은 영역의 높이, A_r은 영역의 정점에서 αh_r, $0<\alpha<1$ 아래의 밑면적을 나타낸다.

$$\text{정의} : p_r = \log(h_r^2/A_r) \qquad (1)$$

분수령 분할은 봉우리 사이의 경계를 짓는 것이므로 봉우리에 속하지 않는 평지부분도 영역에 포함된다. 따라서 p_r은 그림 3과 같이 선명한 봉우리 부분 즉, 봉우리를 구성하지 않는 산자락 부분을 제외한 부분으로 계산된다. 부분식 h_r^2/A_r은 무차원이므로 영역의 크기와는 무관한 척도이다. 여러 가지 영상에 대한 실험을 통계적으로 분석한 결과, h_r이 높아짐에 따라 대체적으로 부분식 h_r^2/A_r도 커지는 관계를 보인다. 그러나 h_r이 높아질수록 그러한 관계가 교란되는 현상을 보인다. h_r에 따라 오차가 커지는 현상을 억압하기 위하여 부분식 h_r^2/A_r의 대수값을 취하여 첨도 p_r로 정의한다.

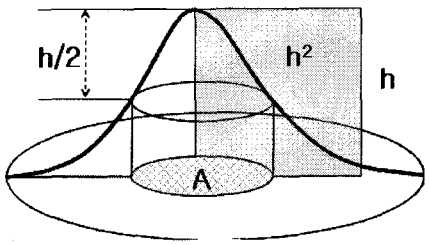


그림 3. 첨도 p_r의 설명도
Fig. 3. Prominence p_r

3.3.2 회귀분석

다양한 겔 영상의 실험을 통계적으로 분석한 결과, 영역의 높이가 매우 낮은 경우를 제외하고, 영역의 높이에 따른 평균 p 곡선이 지수분포(포화곡선)의 특성을 가짐을 알았다. 실험 결과의 한 예로 그림 4의 겔 영상 DICTYSLUG에서 영역의 고도 h에 따른 p의 평균은 그림 5와 같은 곡선을 그린다.

영역 중에서 높이가 높은 영역들이 배경이 될 가능성은 매우 낮으므로 높이가 thBG보다 높은 영역들의 p_r을 지수함수 회귀분석한다. 식 (2)와 같이 문턱값 thBG는 영역 r의 높이 h_r이 thBG 보다 높은 영역의 개수가 전체 영역 개수의 1/2이 되는 값으로 결정한다.

$$|\{r|h_r > thBG\}| = \frac{|R|}{2} \qquad (2)$$

영역의 높이에 따른 p_r의 평균곡선(단, h_r>thBG)에 식 (3)과 같은 지수함수를 회귀분석한다.

$$f(h)=a[1-\exp(-bh)]+c \qquad (3)$$

회귀분석의 결과로 지수함수의 매개변수 a, b, c의 값이 정해지며, 이들을 이용해서 문턱값을 계산한다. 이 예에서 thBG=7이며, 지수함수로 회귀분석한 결과 a=5.836135, b=0.015, c=0.1이다. 그림 6에 곡선으로써 나타내었다.

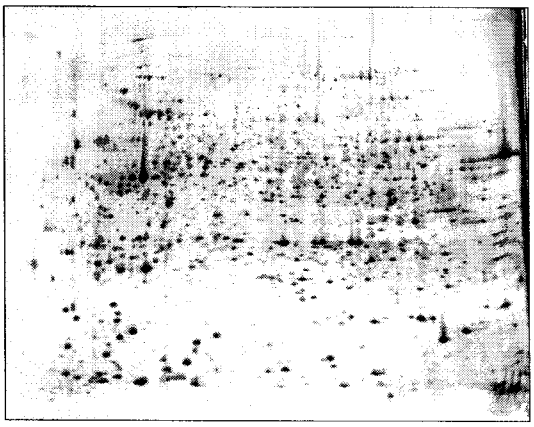


그림 4. 겔 영상 DICTYSLUG
Fig. 4. Gel image DICTYSLUG

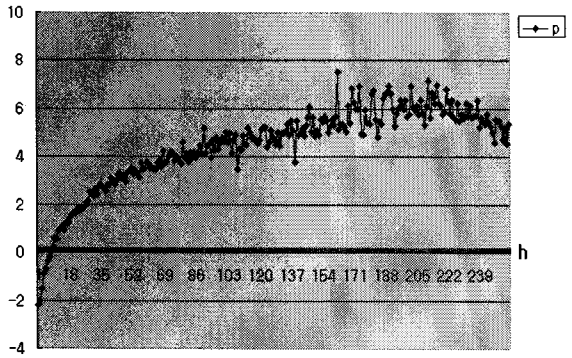


그림 5. 영역의 높이에 따른 평균 P 곡선 : DICTYSLUG
Fig. 5. Average P curve for region height : DICTYSLUG

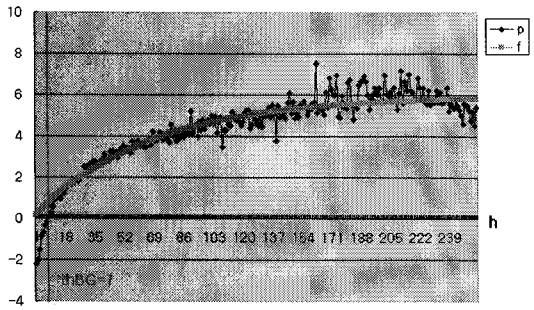


그림 6. 그림 5의 회귀분석 결과
Fig. 6. Curve fitting result for Fig. 5.

3.3.3 문턱값 결정

문턱값 결정은 오차의 확률분포에 따른다. 영역에 대한 평균 p 곡선과 지수함수 f(h)와의 오차는 식 (4), 오차의 평균은 식 (5), 오차의 편차는 식 (6)과 같다.

$$\text{오차: } e_r=|p_r-f(h_r)|,\ r\in R\ (R\text{:영역집합}) \qquad (4)$$

$$\text{오차의 평균: } \bar{e}=\frac{1}{|R|}\sum_{r\in R}e_r \qquad (5)$$

$$\text{오차의 편차: } \sigma=\sqrt{\frac{1}{|R|}\sum_{r\in R}(e_r-\bar{e})^2} \qquad (6)$$

배경을 봉우리로 오인하면 다음 처리 단계에서 배경으로 판별할 수 있는 기회가 있으나, 봉우리를 배경으로 오인하면 제거되어버리므로 그 기회가 사라진다. 따라서 문턱값은 가짜 부정(false negative)보다 가짜 긍정(false positive)의 기준으로 결정한다.

오차의 확률분포에 따르면 배경에 포함되지 않은 영역들의 99.7%는 $\bar{e}-3\sigma < e_i \le \bar{e}+3\sigma$ 범위 안에 존재한다. 따라서 영역 중에서 $p_i < f(h_i)-\bar{e}-3\sigma$인 영역은 배경에 속할 가능성이 매우 높으므로 제외할 수 있다. 문턱값 적용의 오류 발생 확률은 0.15%가 되므로 적응적 문턱값의 신뢰도는 99.85%로 추정된다. 3.3.2절의 예를 예로 들어 문턱값을 계산해 보이면, 문턱값은 각 영역 마다 적응적으로 결정된다. 해당 영역의 고도 h에 의한 $f(h)$ 즉, $5.836135[1-\exp(-0.015\times h)]+0.1$에서 $\bar{e}$와 3σ를 뺀 것이 해당 영역에 대한 문턱값이 되며, 해당 영역의 첨도 p_i가 문턱값보다 크면 해당 영역은 배경이 아닌 반점 영역으로 결정된다.

Ⅳ. 실험 및 분석

실험용 목적 시스템은 Microsoft Windows XP 기반의 PC이며, 전체 알고리즘은 Microsoft Visual C++ 6.0을 사용해서 구현하였다.

4.1 실험 방법

실험을 위한 겔 영상은 표 1과 같으며 SWISS-2DPAGE[8]의 겔 영상들이다. 표 1에서 동정 반점 수는 이미 정밀한 실험을 통해서 확인된 단백질 반점의 개수를 나타내며, 본 논문에서 분석한 적응적 문턱값의 정확도를 검증하는데 사용된다. 실험을 위해서 각 동정 반점들의 좌표를 담은 텍스트 파일을 겔 영상마다 작성하여 사용한다.

표 1. 실험용 겔 영상
Table 1. Gel image for experiment

	Gel image	No. of identified spot
1	Arabidopsis	79
2	Dictyslug	25
3	Ecoli4-5	70
4	Ecoli4.5-5.5	97
5	Ecoli5-6	106
6	Ecoli5.5-6.7	16
7	Ecoli6-9	2
8	Ecoli6-11	34
9	Ecoli	206
10	Elc_human	35
11	Hl60_human	26
12	Kidney_human	44
13	Lymphoma_human	60
14	Nuclei_liver_human	27
15	Nucleoli_hela_2D_human	51
16	Platelet_human	41
17	Staphylococcus	250
18	U937_human	42
19	Yeast	101

실험 방법은 겔 영상을 입력받아 적응적 문턱값 적용까지 처리한 후, 그 결과인 후보 반점 리스트 내에 동정 반점이 얼마나 포함되어 있는 지를 분석하는 방법을 사용하며, 전체 동정 반점 수에 대한 리스트에 포함된 동정 반점 수의 비율을 마크율(mark ratio)로 정의하여 결과 분석에 사용한다. 그림 4의 DICTYSLUG 겔 영상을 실험의 한 예로 들어 차례대로 살펴본다. DICTYSLUG 영상의 분수령 알고리즘 수행 후의 영역분할 모습은 그림 7과 같다. 분수령 알고리즘 수행 후에 분할된 영역의 수는 6821개이다.

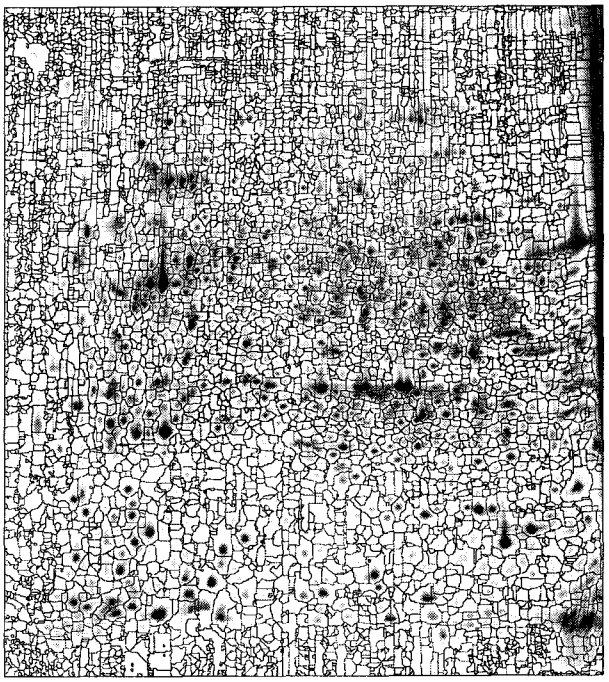


그림 7. DICTYSLUG의 분수령 수행결과
Fig. 7. Watershed result for DICTYSLUG

그림 8은 적응적 문턱값 결정 적용 결과, 단백질일 가능성이 높은 후보 반점들의 영역들을 나타낸다. 후보 반점들의 영역의 봉우리들이 + 기호로 표기되어 있다. 분할 결과인 6821개의 영역 중에 5596개의 영역이 후보 반점들의 영역임을 보인다.

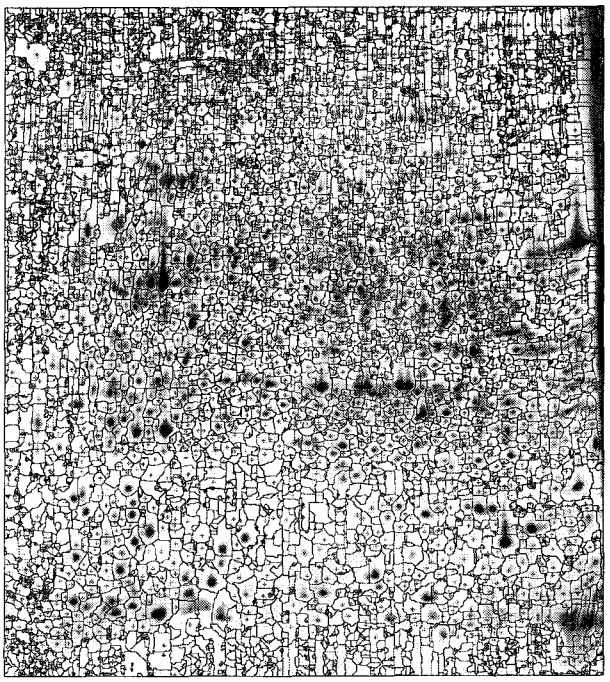


그림 8. DICTYSLUG의 후보 반점
Fig. 8. Candidate spots for DICTYSLUG

그림 9에 문턱값 결정에 의해서 후보 반점으로 결정된 반점의 예를 보인다. 실험에 의한 매개변수 a의 값은 5.836135, b의 값은 0.015, c의 값은 0.1이며, 통계치인 $\bar{e}$는 0.654089, σ는 0.557774이다. 영역의 h는 7이므로, f(h)는 0.68172, 문턱값은 -1.645691이다. 따라서 영역의 첨도 p의 값인 -1.13206은 문턱값보다 크다. 그러므로 후보 반점으로 결정된다.

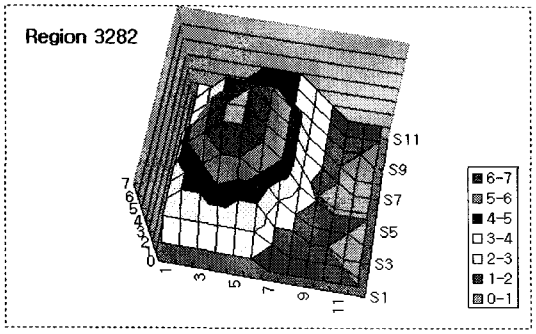


그림 9. 후보 반점의 예
Fig. 9. An example of candidate spot

그림 10은 적응적 문턱값 결정에 의해서 탈락된 반점들, 즉 단백질일 가능성이 적은 반점들을 나타낸다. 탈락된 반점들의 영역의 봉우리들이 '+'로 표기되어 있다. 분할 결과인 6821개의 영역 중에 1225개의 영역이 후보 반점들의 리스트에서 탈락되었음을 보인다.

그림 10. DICTYSLUG의 탈락 반점
Fig. 10. Non-candidate spots for DICTYSLUG

그림 11에 문턱값보다 낮아서 탈락된 예를 보인다. 실험에 의한 매개변수 a의 값은 5.836135, b의 값은 0.015, c의 값은 0.1이며, 통계치인 $\bar{e}$는 0.654089, σ는 0.557774이다. 영역의 h는 7이므로, f(h)는 0.68172, 문턱값은 -1.645691이다. 따라서 영역의 첨도 p의 값인 -1.854383은 문턱값보다 작다. 그러므로 후보 반점 리스트에서 제거된다.

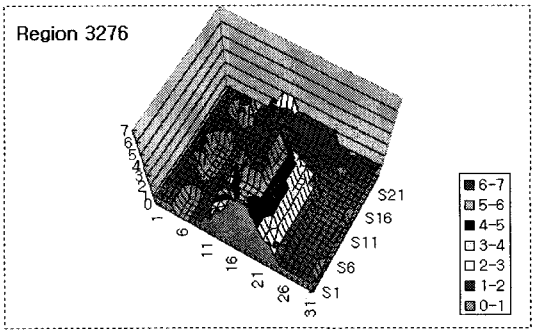


그림 11. 탈락 반점의 예
Fig. 11. An example of non-candidate spot

다음 단계의 실험은 후보 반점 리스트에 동정 반점들이 포함되어 있는 지를 분석하는 것이다. DICTYSLUG의 동정 반점 수는 25개이며, 각 동정 반점의 좌표가 텍스트 파일에 담겨 있다. 그림 12는 후보 반점 리스트에 포함되어 있는 동정 반점을 보이는데 전체 25개의 동정 반점 중에 24개가 포함되어있다.

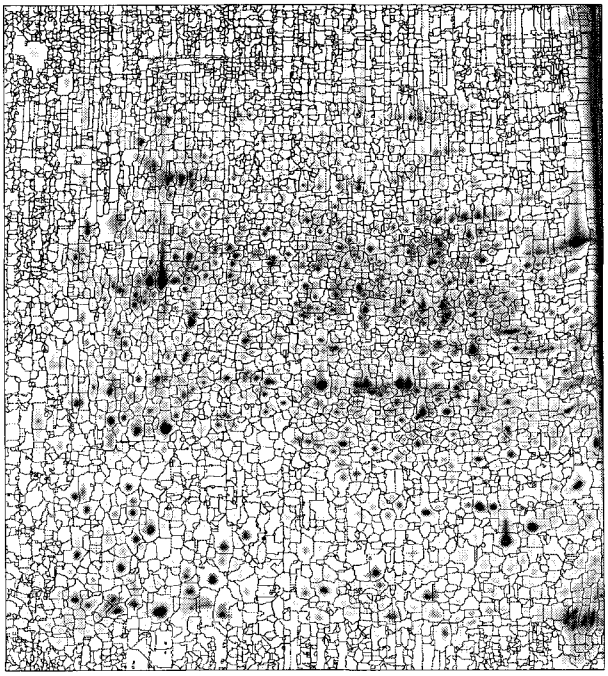


그림 12. 후보 반점 리스트에 포함되어 있는 동정반점
Fig. 12. Identified spots included in candidate spot list

그림 13은 후보 반점 리스트에 포함되어 있지 않은 동정 반점을 보인다. 그림 13에서 빨강색 '+' 표기는 한 영역에 중복되어 있는 동정 반점을 나타낸다. 실험 결과를 보면 한 영역에 중복되어 있는 동정 반점과 분수령 라인에 놓여 있는 동정 반점이 존재하는데 이것은 분수령 알고리즘의 수행결과 영역으로 분할되지 못한 결과로 인한 것이다.

그림 13. 후보 반점 리스트에 포함되어 있지 않은 동정반점

Fig. 13. Identified spots not included in candidate spot list

DICTYSLUG 영상에 대한 실험 결과를 요약하면 후보 반점 리스트에 포함되어 있는 동정 반점의 개수는 24개이고, 한 영역에 중복되어 있는 동정 반점의 개수는 1개, 분수령 라인에 놓여 있는 동정 반점의 개수는 0개이며, 적응적 문턱값 적용에 의해서 탈락된 동정 반점의 수는 0개이다. 따라서 후보 반점 리스트에 동정 반점이 얼마나 포함되어 있는 지를 나타내는 마크율은 24/25*100, 즉 96%가 된다.

4.2 결과 및 분석

4.1절의 실험방법으로 표 1의 실험 영상들을 실험한 결과는 표 2와 같다. 표 2에서 MK(마크수)는 후보 반점 리스트에 존재하는 동정 반점의 개수, UM(탈락수)은 적응적 문턱값 적용에 의해서 후보 반점 리스트에서 탈락된 동정 반점의 개수, DP(중복수)는 한 영역에 중복되어 존재하는 동정 반점의 개수, WL는 분수령 라인에 존재하는 동점 반점의 개수를 의미한다.

표 2. 실험 결과

Table 2. Experiment results

	Gel image	ID	mark ratio	MK	UM	DP	WL
1	Arabidopsis	79	96.20	76	0	3	0
2	Dictyslug	25	96.00	24	0	1	0
3	Ecoli4-5	70	97.14	68	0	2	0
4	Ecoli4.5-5.5	97	98.97	96	0	0	1
5	Ecoli5-6	106	99.06	105	0	0	1
6	Ecoli5.5-6.7	16	100.00	16	0	0	0
7	Ecoli6-9	2	100.00	2	0	0	0
8	Ecoli6-11	34	100.00	34	0	0	0
9	Ecoli	206	99.03	204	0	0	2
10	Elc_human	35	100.00	35	0	0	0
11	Hl60_human	26	100.00	26	0	0	0
12	Kidney_human	44	95.45	42	0	2	0
13	Lymphoma_human	60	95.00	57	0	3	0
14	Nuclei_liver_human	27	100.00	27	0	0	0
15	Nucleoli_hela_2D_human	51	96.08	49	0	0	2
16	Platelet_human	41	97.56	40	0	0	1
17	Staphylococcus	250	96.00	240	0	10	0
18	U937_human	42	97.62	41	0	1	0
19	Yeast	101	99.01	100	0	1	0
	Σ	1312	97.71	1282	0	25	5

ID : no. of identified spots
MK : no. of marked spots
UM : no. of unmarked spots
DP : no. of identified spots in the same region
WL : no. of identified spots on the watershed line

전체 19개의 겔 영상에 대한 마크율의 평균은 97.71%이므로 이론적인 신뢰도인 99.85%보다 낮다. 하지만 마크율을 낮춘 주원인은 한 영역에 중복되어 포함된 동정 반점과 분수령 라인에 놓여 있는 동정 반점이 존재하기 때문이다. 또한 적응적 문턱값 적용에 의해 탈락된 동정 반점의 개수가 전체 19개 영상에서 하나도 없다는 점을 보면 제안 방식의 정확성이 높음을 알 수 있다.

또한 4.1절의 그림 9와 그림 11의 실험 예를 살펴보면, 두 영역의 높이는 동일하지만 그림 9는 후보 반점으로 판정되고, 그림 11은 탈락 반점으로 판정된다. 고정 문턱값을 사용하였다면 두 영역 모두 후보 반점이거나, 두 영역 모두 탈락 반점으로 판정되었을 것이다. 따라서 고정 문턱값을 사용한 기법보다 제안 기법이 보다 더 효과적이다.

V. 결 론

본 논문에서는 2차원 전기영동 영상 분석 프로그램의 반점 검출 단계에서 해결해야할 문제점 중에 하나인 잡

영을 제거하기 위한 새로운 기법으로 영상에 따라 자동적으로 문턱값을 결정하는 기법을 제안하였다. 제안 기법은 봉우리의 첨도를 통계적으로 분석한 결과에 착안한 기법으로, 영역분할 후에 영역들의 첨도의 평균 곡선을 지수함수에 회귀분석하여 지수함수의 매개변수를 구한 다음, 오차의 확률분포에 따라서 매개변수들로 문턱값을 결정하여 적용한다. 오차의 확률분포에 따르면 문턱값 적용의 신뢰도는 99.85%이며, 제안 기법을 실험 영상으로 실험한 결과로써 적응적 문턱값 결정 기법이 정확함을 보았다.

향후 연구 방향은 반점 형태 분석에 관한 연구와 구현이며, 더 나아가서 겔 정합에 대한 연구와 구현을 완료하여 2차원 전기영동 영상 분석 프로그램을 완성하는 것이다.

참고문헌

[1] A. W. Dowsey, M. J. Dunn, and G.-Z. Yang, "The role of bioinformatics in two-dimensional gel electrophoresis", Proteomics 2003, pp. 1567-1597, 2003

[2] L. Addario-Berry, "2D Gel Electrophoresis - An Overview", http://www.mcb.mcgill.ca/~hallett/GEP/PLectures/Plecture3.pdf, May 9, 2002

[3] A. Roy, K. R. Lee, Y. Hang, M. Marten, and, Technical Reports, Dept. of Mathematics and Statistics, University of Maryland, 2003

[4] L. Pedersen, Analysis of two-dimensional electrophoresis gel images, PhD Thesis, Informatics and Mathematical Modelling, Technical University of Denmark, 2002

[5] K.-P. Pleissner, F. Hoffman, K. Kreigel, C. Wenk, S. Wegner, A. Sahlstrom, H. Oswald, H. Alt, and E. Fleck, "New algorithmic approaches to protein spot detection and pattern matching in two-dimensional electrophoresis gel databases", Electrophoresis 1999, 20, pp. 755-765, 1999

[6] E. Bettens, "Peak characterization using parameter estimation methods", PhD thesis, University of Antwerpen, 1999

[7] L. Vincent, and P. Soille, "Watershed in digital spaces: An efficient algorithm based on immersion simulations", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 12, No. 7, July, pp. 583-598, 1990

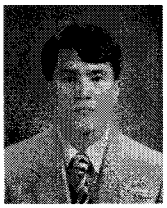
[8] C. Hoogland, K. Mostaguir, J.-C. Sanchez, D. F. Hochstrasser, R. D. Appel, SWISS-2DPAGE, ten years later. Proteomics 2004, 4(8), pp. 2352-2356, 2004

[9] A. Efrat, F. Hoffman, K. Kriegel, C. Shultz, and C. Wenk, "Geometric Algorithms for the Analysis of 2D-Electrophoresis Gels", Journal of Computational Biology, pp. 114-123, 2002

[10] K. Takahashi, M. Nakazawa, Y. Watanabe, and A. Konagaya, "Fully-Automated Spot Recognition and Matching Algorithms for 2-D Gel Electrophoretogram of Genomic DNA", Genome Informatics 1998, December 10-11. 1998, Yebisu garden place, Tokyo, Japan, pp. 161-172, 1998

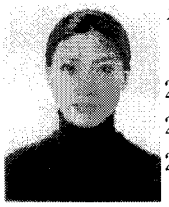
[11] E. Bettens, P. Scheunders, J. Sijbers, D. Van Dyck, and L. Moens, "Automatic segmentation and modeling of two-dimensional electrophoresis gels", Proceedings of IEEE International Conference on Image Processing, Vol. 2, pp. 665-668, Lausanne, Switserland, September, 1996

[12] S. Beucher, Recent advances in mathematical morphology" Invited lecture, Société Italienne de Microscopie Electronique, Taormina, Sicile, 1992



최 관 덕(Kwan-Deok Choi)

1989년 2월 영남대 전산공학과(공학사)
1991년 2월 영남대 전산공학과(공학석사)
1995년 2월 영남대 전산공학과(박사수료)
1996년 3월 ~ 현재 대구과학대학 컴퓨터계열 조교수
관심분야 : 바이오인포매틱스, 영상처리, 프로그래밍언어



김 미 애(Mi-Ae Kim)

2004년 2월 영남대 정보통신공학과(공학사)
2006년 2월 영남대 컴퓨터공학과(공학석사)
2006년 3월 ~ 현재 영남대 컴퓨터공학과 박사과정
2007년 8월 ~ 현재 경북전략산업기획단 연구원
관심분야 : 바이오인포매틱스, 영상처리

윤 영 우(Young-Woo Yoon)

1972년 2월 영남대 전자공학과(공학사)
1983년 2월 영남대 전자공학과(공학석사)
1988년 2월 영남대 전자공학과(공학박사)
1988년 9월 ~ 현재 영남대학교 전자정보공학부 컴퓨터공학전공 교수
관심분야 : 바이오인포매틱스, 영상처리, 컴퓨터비전, 컴퓨터설계, 프로그래밍언어