스팟 윤곽의 기울기를 이용한 단백질 2차원 전기영동 영상에서의 스팟 검출 방법

유혜경° 이성환
고려대학교 정보통신대학 컴퓨터학과
{hkyoo°, swlee}@image.korea.ac.kr

A Spot Detection Method in 2D Electrophoresis Images Using Gradients of Spot Boundary

Hye-Kyoung Yoo°, Seong-Whan Lee
Department of Computer Science and Engineering, Korea University

요 약

단백질 2차원 전기영동은 다양한 단백질 분리 방법 중 가장 널리 쓰이는 방법으로 실험 결과를 촬영한 영상을 분석하여 얻은 단백질 스팟의 위치나 질량, 발현 유무 등을 이용하여 각종 질병의 발생 원인, 진행 상태, 생리적인 변화 등에 대해 분석할 수 있다. 실험 영상에 다수의 단백질이 존재하므로 이를 수작업으로 처리할 경우에 많은 시간과 노력이 소요되므로 본 논문에서는 자동화된 단백질 스팟 검출 방법을 제안하였으며 단백질 스팟이 같은 위치에 겹쳐서 나타나는 경우가 많은 단백질 2차원 전기영동 실험 영상의 특성을 고려하여 여러 개의 단백질이 겹쳐진 복잡한 스팟 영역에 대해서 스팟의 형태 정보를 이용하여 스팟의 개수를 추정하고 개별 스팟으로 분리하여 보다 신뢰성 있는 분석이 가능하게 하였다. 본 논문에서 제안된 방법의 효용성을 검증하기 위해 기존에 널리 사용되고 있는 상용 소프트웨어와 비교 실험을 수행한 결과 겹친 정도가 60% 이상인 경우 기존 방법에 비해 우수한 결과를 보였다.

1. 서 론

2003년, 인간 게놈 프로젝트를 통해 인간 게놈 지도가 완성된 후 유전체 정보를 기반으로 생물의 다양한 특성을 분석하는 후기 유전체학이 등장하였다. 후기 유전체학의 다양한 분야 중 하나인 단백질체학의 발전에 기여를 한 단백질 2차원 전기영동 실험은 다양한 단백질 분리 방법 중 가장 널리 쓰이는 방법이다. 단백질 2차원 전기영동 실험은 단백질의 전기적 특성과 질량 정보를 이용하여 단백질을 분리하는데 한번의 실험으로 세포 내에 존재하는 모든 단백질의 발현 정도나 변형 상태 등을 파악할 수 있으며 이러한 정보를 바탕으로 각종 질병의 발생 원인, 진행 상태, 생리적인 변화 등에 대해 분석 할 수 있어 활발히 연구되고 있다. 단백질 2차원 전기영동 실험 영상은 다수의 단백질 스팟을 포함하고 있으며 배경과 스팟의 밝기값의 차이가 미미하고 겹친 스팟이 많이 발생하는 특성 때문에 수작업으로 처리할 경우 많은 경험과 시간이 필요하다는 단점이 있다. 이러한 문제를 해결하고 빠른 영상 분석을 위해 자동화된 방법들이 연구되고 있는데 각 스팟의 중앙에 명암의 최고값을 가지는 꼭지점이 존재한다는 가정 하에 영상의 픽셀값을 이용하여 단백질 스팟을 검출하는 픽셀값 수집 방법[1]과 스팟 후보 영역의 윤곽선 중 임의의 세 점을 이용해서 타원을 생성하고 스팟의 개수를 추정하여 그리는 방법[2]이 대표적이다. 이밖에 워터쉐드를 이용해서 스팟 후보 영역을 추출한 후에 모델링 기법을 사용하여 스팟을 검출하는 방법으로 가우시안 모델링, 확산 모델링, 통계적 모델링[3] 등이 있다. 기존 연구들은 단백질 2차원 전기영동 영상에서 많이 발생하는 겹친 스팟들을 처리하는 단계가 없어서 여러 개의 스팟을 하나로 정량화하거나 스팟 영역의 형태와 무관하게 스팟 개수를 추정하는 단점이 있다. 따라서, 본 논문에서는 기존 연구의 단점을 보완하기 위해 여러 개의 단백질이 겹쳐진 복잡한 스팟 영역에 대해서 스팟의 형태 정보를 이용하여 스팟의 개수를 추정하고 개별 스팟으로 분리하는 과정을 수행하여 보다 신뢰성있는 분석이 가능한 자동화된 단백질 2차원 전기영동 영상 분석 방법을 제안하였다.

2. 제안된 스팟 검출 방법

본 논문에서 제안하는 방법은 단백질 2차원 전기영동 실험 영상이 입력되면 추가적인 정보나 사용자의 개입없이 자동으로 단백질 스팟을 검출하는 방법이다. 영상이 입력되면 먼저 단백질 스팟과 배경의 후보 영역들을 추출한 후 스팟 후보 영역과 배경을 구분한다. 스팟 후보 영역은 단일 스팟과 겹친 스팟으로 나눌 수 있는데, 각 단백질별로 정확한 발현량 분석을 위해 각 스팟 후보 영역에 대해 윤곽의 기울기를 계산하여 기울기 변화에 대한 프로파일을 생성한 후, 패턴을 분석하여 스팟 개수를 추정한다. 겹친 스팟으로 판단될 경우 스팟 개수에 따라 영역을 분리한 후 정량적 분석을 한다. 그림 1에 제안된 방법의 전체 흐름도를 나타내었다.

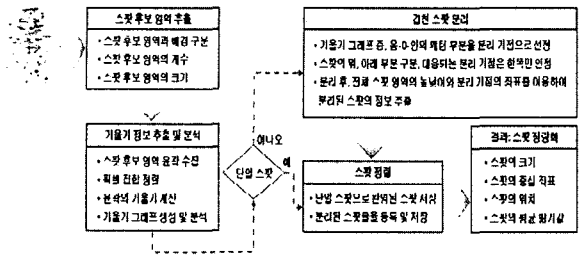

그림 1. 제안 방법의 전체 흐름도

2.1 스팟 정보 검출

2.1.1 워터쉐드를 이용한 스팟 후보 영역 추출

단백질 스팟을 검출하기 위해 워터쉐드 알고리즘을 사용하여 스팟 후보 영역을 추출하는데, 여기서 스팟 후보 영역은 단백질 2차원 전기영동 영상에서 스팟 존재 가능성이 있는 영역을 의미한다. 단백질 2차원 전기영동 영상은 지역에 따라 배경의 밝기값 변이가 심한 특성을 가지고 있는데 이러한 특성에 맞춰 효과적으로 스팟 후보 영역을 추출할 수 있는 방법이 워터쉐드 알고리즘이다[4]. 워터쉐드 알고리즘은 영상에서 픽셀의 밝기값이 높은 곳은 언덕을 이루고 밝기값이 낮은 곳은 계곡을 이루는 지형적인 높낮이를 가진다고 가정하고 각 영역별로 최저점에서 단계적으로 침수시키면서 두 영역이 만나게 되는 지점에 워터쉐드 라인을 만들고 이를 이용해 영역을 분할한다. 워터쉐드 알고리즘을 사용하면 매우 세밀한 영역 분할이 가능하지만 대부분의 경우 실제보다 많은 영역을 만들어 추가적으로 영역을 병합하는 과정을 수행해야 한다.

2.1.2 스팟 윤곽의 기울기 정보 추출

워터쉐드 알고리즘을 사용하여 검출된 영역에 대해 레이블링을 수행하여 각 영역의 크기, 높이, 넓이, 픽셀 평균 값 등의 정보를 추출한다. 추출된 정보를 이용하여 단백질 2차원 전기영동 실험 영상의 배경과 스팟 후보 영역을 구분한다. 각 영역을 배경과 스팟 후보 영역으로 나누는 기준은 영역의 가장 작은 픽셀 값이 임계값 보다 작고, 영역의 크기가 임계값 보다 크면 스팟 후보 영역이다. 그리고, 나머지 영역은 배경으로 분류한다. 이것은 스팟과 비슷한 정도의 명암값을 갖는 배경의 잡음과 스팟을 구분하기 위한 기준이다. 명암값이 높을 경우와 그 크기가 작을 경우에 영역을 잡음으로 구분하게 된다.

스팟 후보 영역과 배경이 결정되면 스팟 후보 영역을 단일 스팟과 겹친 스팟 영역으로 구분한다. 각 단백질별로 정확한 발현량을 측정하기 위해서는 겹친 스팟 영역에 대해 몇 개의 단백질이 겹쳐져 있는지 측정한 후 각 단백질별로 영역을 분할한 후 발현량을 측정하여야 하므로 두 가지 스팟의 구분이 필수적이다. 단일 스팟과 겹친 스팟 영역을 구분하기 위해 먼저 스팟 후보 영역의 윤곽 픽셀들의 집합을 구한다. 단백질 2차원 전기영동 실험 영상의 전체 영역을 R 이라 하고, 이웃 픽셀에 워터쉐드 라인을 이루는 픽셀을 하나라도 포함하고 있는 픽셀들의 집합을 $R_{watershedNeighbor}$라 하면 스팟 후보 영역의 윤곽 픽셀들의 집합 $R_{boundary}$는 다음과 같다.

$$R_{boundary} = \{p \in R \mid p \in R_{watershedNeighbor}\}$$

다음으로 그림 2에서와 같이 스팟 후보 영역의 윤곽 픽셀들의 집합 $R_{boundary}$를 스팟 후보 영역의 왼쪽 최상단 픽셀부터 시계 방향으로 정렬시킨다. 이때 스팟 후보 영역의 위쪽 즉, 왼쪽 최상단 픽셀에서 시작해서 오른쪽 최상단 픽셀까지는 같은 x축의 여러 픽셀 중 y축의 값이 가장 작은 픽셀을 선택한다. 아래쪽의 경우에는 같은 x축의 여러 픽셀 중 y축의 값이 가장 큰 픽셀들을 선택한다. 각 x축에서 y축의 값이 제일 작은 또는 제일 큰 픽셀들을 선택함으로써 스팟 윤곽의 작은 굴곡에 강건하게 기울기 정보를 추출할 수 있고 큰 굴곡의 기울기를 극대화하는 효과가 있다.

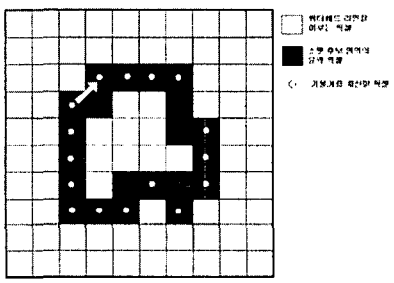

그림 2. 스팟 후보 영역의 윤곽 픽셀 집합 생성

2.2 스팟의 개수 추정 및 분리

앞 단계에서 단백질 2차원 전기영동 실험 영상에서 스팟 후보 영역을 추출한 후, 스팟 후보 영역 윤곽 픽셀들의 기울기를 구했다. 이 기울기의 정보를 분석하여 각 후보 영역에 포함된 정확한 스팟 개수를 추정한다.

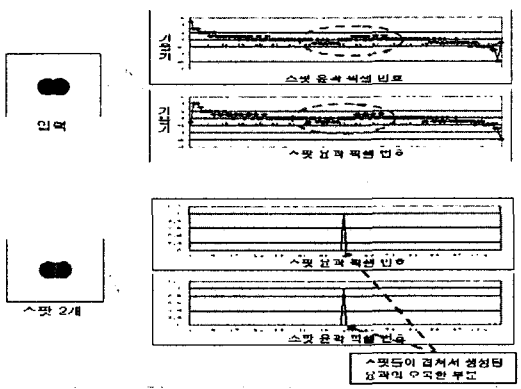

그림 3. 스팟 분리를 위한 기울기 분석

그림 3을 보면, 스팟 후보 영역 윤곽의 기울기 그래프를 스팟의 위와 아래 부분으로 구분하여 그린 후 각 그래프에서 음과 양이 바뀌는 부분을 찾아 스팟 분리 기점으로 지정하여 새로운 그래프를 생성하였다. 스팟 분리 기점을 정하는 방법은, 가로로 겹친 복잡한 스팟 후보 영역에서 주기에 나타나는 패턴을 찾은 후, 음수를 나타내는 마지막 픽셀과 양수를 나타내는 첫 번째 픽셀 좌표 값들의 중간 점을 구해서 스팟 분리 기점을 정한다. 스팟의 위 부분과 아래 부분의 대칭 되는 이러한 분리 기점이 발견되면, 그 좌표 값들을 이용하여 스팟을 분리한다. 이 방법을 사용하면 2개 이상의 스팟이 겹친 경우에도 정확하게 식별할 수 있는 장점이 있다.

3. 실험 결과 및 분석

본 논문에서 제안한 방법의 효용성을 증명하기 위해 인공적으로 생성한 겹친 스팟과 실제 영상을 이용하여 실험을 하였다. 단백질 스팟 검출 실험의 성능을 평가하기 위해서는 TPF(True Positive Fraction)를 사용하였다. TPF는 겹친 스팟의 개수에 대한 정보를 알고 있는 영상에서 겹친 스팟의 개수를 정확히 추측하고 분리시킨 정확도를 의미한다.

그림 4는 인공적으로 생성한 겹친 스팟 영상[5]을 이용한 스팟 검출 실험 결과를 보여주고 있다. 인공 스팟 영상을 이용한 실험에서, 조금 겹친 경우에는 비슷한 성능을 보이지만 겹친 정도가 심해지면 제안된 방법이 Melanie보다 약간 우수하거나 비슷한 성능을 보이며 Easy Pro와 Z3에 비해서는 매우 좋은 성능을 보인다. TPF를 측정한 결과 제안된 방법은 72.5%의 성능을 보이는 반면 EasyPro와 Z3는 60% 미만, Melanie는 70%의 성능을 보였다. 본 논문에서 제안한 방법은 겹친 스팟 영역의 윤곽의 기울기 정보를 이용하여 스팟의 개수를 추정하기 때문에 스팟을 이루는 픽셀 값의 차이가 적은 영역에서도 스팟 윤곽의 기울기를 이용하여 스팟을 분리할 수 있기 때문으로 보인다.

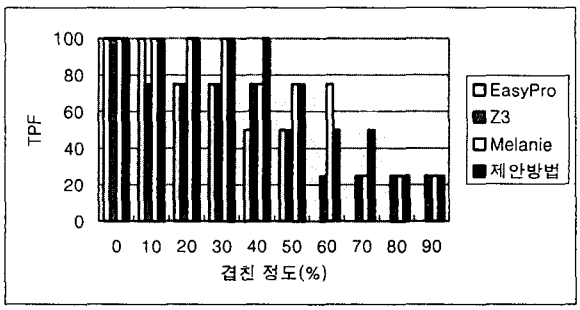

그림 4. 성능 비교 평가 그래프

실제 단백질 2차원 전기영동 영상을 이용한 실험 결과는 그림 5에 나타내었다. 그림 5를 보면 제안된 방법과 Easy Pro는 스팟의 중심을 찾은 후 스팟의 윤곽을 타원형으로 표시했으며, Z3와 Melanie의 경우에는 스팟의 윤곽을 따라 기하학적으로 스팟을 표시하였다. 최종적으로 스팟을 검출한 결과를 보면 스팟의 개수를 잘못 추정하거나 필요없는 주변 배경까지 포함한 다른 방법들에 비해 본 논문에서 제안한 방법이 가장 정확하게 스팟을 식별한 것을 알 수 있다.

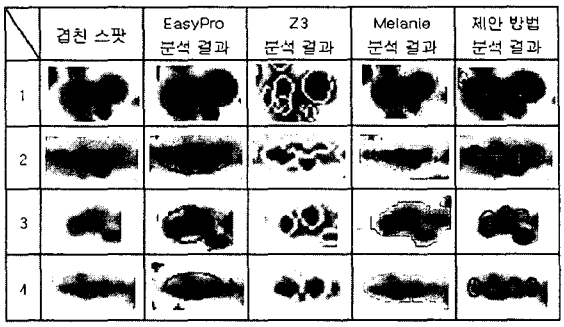

그림 5. 단백질 2차원 전기영동 영상을 이용한 실험 결과

4. 결론 및 향후 연구 방향

본 논문에서는 정확한 단백질 발현량 분석을 위해 단백질 스팟들이 다양하게 겹친 영역에 대해 스팟의 수를 추정하고 각 단백질별로 스팟 영역을 분할하는 과정을 추가한 단백질 2차원 전기영동 실험 영상 자동 분석 기법을 제안하였다. 제안된 방법의 효용성을 검증하기 위해 인공적으로 생성한 스팟과 실제 단백질 2차원 전기영동 실험 영상을 사용하여 스팟 검출 실험을 수행한 결과 기존 상용 소프트웨어들과 비교하여 가장 좋은 성능을 보였다. 향후 과제로는 겹친 정도가 심한 경우에도 스팟을 식별할 수 있는 알고리즘의 개발과 배경값을 정확히 제거하는 것에 대해 추가 연구가 필요하다.

참고문헌

[1] P. Cutler, G. Heald, I. R. White and J. Ruan, "A novel approach to spot detection for two-dimensional gel electrophoresis images using pixel value collection," *Proteomics*, Vol. 3, No. 4, 2003, pp. 392-401

[2] A. Efrat, F. Hoffmann, C. Knauer, K. Kriegel, G. Rote and C. Wenk, "Covering shapes by ellipses," *Proc. of ACM-SIAM Symposium on Discrete Algorithms*, San Francisco, USA, 2002, pp. 453-454

[3] M. Rogers, J. Graham and R. P. Tonge, "Statistical models of shape for the analysis of protein spots in two-dimensional electrophoresis gel images," *Proteomics*, Vol. 3, No. 6, 2003, pp. 887-896

[4] L. Vincent, "Morphological grayscale reconstruction in image analysis: applications and efficient algorithms," *IEEE Transactions on Image Processing*, Vol. 2, No. 2, 1993, pp. 176-201

[5] M. Rogers, J. Graham and R. P. Tonge, "Using statistical image models for objective evaluation of spot detection in two-dimensional gels," *Proteomics*, Vol. 3, No. 6, 2003, pp. 879-886