Watersheds 기반 계층적 이진화를 이용한 단백질 반점 분할 알고리즘

김 영 호[†] · 김 정 자[††] · 김 대 현[†††] · 원 용 관[††††]

요 약

생물학자가 단백질을 검색하고 분석하기 위해서는 2차원 젤 전기영동(2DGE : Two Dimensional Gel Electrophoresis) 실험을 해야 한다. 실험 결과는 2차원 영상이 생성된다. 2차원 영상에서 단백질 반점의 패턴 분석을 위해 2차원 젤 영상에 펼쳐진 단백질 반점들을 영상처리를 통해 분할하고, 대조 그룹의 단백질 패턴과 비교분석을 통해 밝히고자하는 단백질 반점을 찾아내야 한다. 단백질 반점을 분할하는 알고리즘에 있어서 기존에는 가우시안 함수를 적용하였지만, 최근 들어 형태학 분리개념에 의한 Watersheds 영역기반 분할(Watersheds region-based segmentation) 알고리즘을 활용하고 있다. 그러나 Watersheds 영역기반 분할 알고리즘은 크기가 큰 영상에서 원하는 영역을 신속하게 분할한다는 장점이 있지만, 영상 화소의 그레이 값이 연속적인 경우 실제 반점의 개수 에 비해 과다분할(over-segmentation)되거나 과소분할(under-segmentation)의 문제점을 안고 있다. 이는 마커(marker) 포인트의 설정에 의해 어느 정도 해결할 수 있지만 병합(merge)과 분할(split) 과정을 반복해야 한다. 본 논문은 Watersheds 기반 계층적 이진화 기법을 적용하여 마커 드리븐 Watersheds 영상분할의 문제점을 해결하고자 한다.

키워드 : 단백체학, 형태학 영역기반 분할, 2차원 젤 전기영동, 단백질 반점, 워터쉐이드

The Algorithm of Protein Spots Segmentation using Watersheds-based Hierarchical Threshold

Youngho Kim[†] · JungJa Kim[††] · Daehyun Kim[†††] · Yonggwan Won[††††]

ABSTRACT

Biologist must have to do 2DGE biological experiment for Protein Search and Analysis. This experiment coming into being 2 dimensional image. 2DGE (2D Gel Electrophoresis : two dimensional gel electrophoresis) image is the most widely used method for isolating of the objective protein by comparative analysis of the protein spot pattern in the gel plane. The process of protein spot analysis, firstly segment protein spots that are spread in 2D gel plane by image processing and can find important protein spots through comparative analysis with protein pattern of contrast group. In the algorithm which detect protein spots, previous 2DGE image analysis is applies *gaussian fitting*, however recently *Watersheds region based segmentation algorithm*, which is based on morphological segmentation is applied. Watersheds has the benefit that segment rapidly needed field in big sized image, however has under-segmentation and over-segmentation of spot area when gray level is continuous. The drawback was somewhat solved by marker point institution, but needs the split and merge process. This paper introduces a novel marker search of protein spots by watersheds-based hierarchical threshold, which can resolve the problem of marker-driven watersheds.

Key Words : Proteomics, Morphological Region Based Segmentation, 2DGE, Protein Spot, Watersheds

1. 2DGE의 생물학적 기초지식

포스트 지놈 시대(The post genome area)를 맞이하여, 인간 유전체(genome) 연구는 생물학적 기능을 밝히는 연구로 진행되고 있다. 단백질 분석은 생물체의 실질적 기능을 담당하고 있으며, 이에 단백체학(Proteomics) 연구가 가속화되고 있다. 단백체학은 세포, 조직 내의 수많은 단백질의 양적, 질적 상태를 입체적으로 분석하며, 단백질 분자 상호간의 물리적, 화학적 관계를 밝혀 세포의 생리현상을 통합적으로 이해하는 연구 분야이다[1]. 단백체(proteome) 연구는 일반적으로 2차원 젤 전기영동에 의해 분석하며, 세 단계로

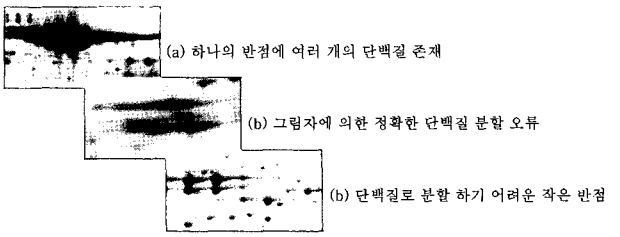
※ 본 연구는 산업자원부 지방기술혁신사업(RTI04-03-03) 지원으로 수행되었음.
† 준 회 원 : 전남대학교 전자컴퓨터정보통신공학부 컴퓨터공학, 박사과정
†† 준 회 원 : 전남대학교 바이오광 사업단, 연구교수
††† 정 회 원 : 조선이공대학교, 광전자공학과, 초빙교수
†††† 종신회원 : 전남대학교 전자컴퓨터정보통신공학부, 부교수
논문접수 : 2004년 11월 18일, 심사완료 : 2005년 3월 30일

구성된다. 첫 번째 단계는 단백질을 분리하는 작업으로서 주어진 실험 샘플을 단백질 분해자(Digestion Enzyme)에 의해 분해 시킨다. 다음으로 이차원 젤 전기연동을 이용하여 전기자극을 수행하면, 2차원 젤 평면상의 수십, 수천 개의 단백질을 반점으로 펼쳐진다. 젤 평면상의 단백질 반점 패턴을 대조군의 단백질 반점 패턴과 비교하고, 분석하고자 하는 단백질 반점을 찾아낸다. 두 번째 단계는 단백질 미량 분석 기술인 MALDI-TOF MS(Matrix Assisted Laser Desorption/Ionisation Time Of Flight Mass Spectrometry), ESI-TOF MS(Electro Spray Ionization Time Of Flight Mass Spectrometry) 등을 이용하여 단백질 분자의 질량(weight of molecular), 펩타이드 매스 핑거프린팅(peptide mass finger printing), 펩타이드 스퀀스 태그(peptide sequence tag) 등에 대한 정보를 얻는다. 마지막 단계는 획득한 정보를 구축된 여러 가지 생물학 데이터베이스를 검색하여 분석하는 반점이 어떠한 기능을 하는 단백질인가를 알아내는 과정이다[2, 3].

단백질 분석을 위한 하나의 방법론으로서 패턴 분석 및 패턴 인식은 생물학적인 샘플 내에 들어있는 단백질의 내용을 동정하고 단백질의 양을 측정하는데 많은 도움을 준다. 현재 2차원 젤 전기영동 패턴 분석기술은 2차원 젤 전기영동 이미지를 통하여 단백질들을 찾아내는 방법으로 널리 쓰이고 있다. 2차원 젤 전기영동 패턴 분석방법은 두 단계로 구성된다. 첫 번째 단계는 생물학적 실험을 거친 전기영동 젤 판을 디지털 카메라나 스캐너 등을 통하여 영상을 입력받아 배경 이미지(background image)에서 단백질 반점 영상을 분할하고, 단백질 반점의 데이터를 획득하는 과정이다. 두 번째 단계는 2개 이상의 2차원 젤 전기영동 이미지를 교차비교(cross matching)하여 단백질 반점 패턴을 분석하는 과정이다[4]. 단백질 데이터 획득 과정에 있어서 단백질 반점의 정확한 분할 결과는 단백질의 발현(expression) 양을 반영하고, 교차 비교는 생물학적 단백질의 발현 변화 차이를 감지한다. 변화의 감지를 통해 기존 데이터베이스를 검색하고 새로운 단백질 또는 유사한 기능의 단백질을 찾아내는 과정이다.

2. 2차원 젤 전기영동 이미지 분석의 문제점

2차원 젤 전기영동 이미지의 처리는 배경 이미지로부터 수백에서 수천 개에 달하는 단백질 반점을 신속하게 분할하여 단백질 데이터를 획득해야 한다. 그러나 그림 1과 같은 문제점을 안고 있다. (그림 1)의 (a) 이미지와 같이 여러 개의 단백질 반점이 하나의 반점처럼 중복되어 존재하는 하거나, 두 번째 (b) 이미지와 같이 전기연동에 의한 시료의 이동으로 단백질들의 그림자(shadow/smearing)가 발생해 배경화면과 희미한 반점 사이에 분리 기준을 마련하기 힘들다. 세 번째, 많은 잡음이 섞여 있어 희미한 반점과 잡음을 구분해내기 힘들다[5]. 이외에도 2차원 젤 전기영동 이미지의 대부분의 반점들은 배경과 확연히 구분되지 않는다.



(그림 1) 2DGE 영상에서 단백질 분할의 문제점

최근 들어 이에 대한 해법으로 Vincent와 Soille가 제안한 Watersheds 영역기반 분할 알고리즘(이하 Watersheds)[6]이 많이 사용되고 있다. 2차원 젤 전기영동 이미지에서는 연속된 그레이 값으로 인한 과다분할 및 과소분할의 발생은 형태학적 반지름을 적절히 적용해야 한다. 또한 과소, 과다 분할의 문제점을 해결하기 위해 분할된 영역을 병합(merge)과 분리(split)을 반복적으로 적용해 주어야 한다는 문제점을 안고 있다. 이에 대한 해결책으로 마커 제어기(Marker Controller)를 이용하여 위 두 가지 문제점을 해결하고 있다[7]. 2차원 젤 전기영동 영상의 마커 제어기를 사용한 마커의 선정은 형태학적 영상처리 기법인 Top-Hat이나 H-Dome를 일반적으로 적용하고 있다[7]. 또는 가우시안 적합과 같은 모델 적합에 의한 마커의 선정을 한다[7]. 그러나 이러한 기법들에 의해서 모든 마커를 선정할 수 없다. 즉 그림 1과 같은 문제점이 마커의 잘못된 선택을 유도하고, 영상의 근본적인 문제점에 의해 마커의 잘못된 선정은 경우 잡음을 마커로 선정하는 경우가 빈번하며, 또한 가우시안 적합의 경우 중심점과 분산에 의해 마커의 개수와 위치가 많은 차이를 보인다.

본 논문에서는 마커의 선정 없이 Watersheds 기반의 계층적 이진화를 적용하여 단백질 반점 분할을 위한 알고리즘을 제안함으로써 마커의 선정이나 병합과 분할을 해야 하는 문제점을 해결하였다. 영상의 분할 과정은 다음과 같다. 영상 향상 기법들인 밝기 향상(Brightness Enhancement), 대비 향상(Contrast Enhancement), 잡음 제거(Noise remove) 등의 필터를 통한 전처리 과정을 통해 영상을 향상 시킨다. 향상된 영상에서 배경 제거 과정을 통해 영상의 배경 및 그림자를 제거한다. 배경 화면이 제거된 영상에 본 논문에서 제안하는 계층적 이진화를 적용하고, 최종 Watersheds를 적용한다. 제안하는 계층적 이진화는 만약 분리된 하위 영역에 하나 이상의 반점이 존재하면, 하위 영역의 국부적인 영역에 대해 새로운 하위 이진화 및 Watersheds를 적용을 말한다. 즉, 정해진 영역에 하나의 단백질 반점이 존재 할 때까지 재귀적인 호출을 한다.

본 논문의 구성은 다음과 같다. 2장에서는 단백질 반점 분석 기법을 소개하고 3장에서는 Watersheds 분할 알고리즘을 알아본다. 4장에서는 제안하는 Watersheds 기반 계층적 이진화(Watersheds-based Hierarchical Threshold)분할 알고리즘을 서술하고, 5장에서 계층적 분할을 위한 변수와 실험의 각 과정과 결과를 보인다. 마지막으로 6장에서 결론을 맺는다.

3. 2차원 젤 전기영동 이미지 분석

3.1 2차원 젤 전기영동 이미지

2차원 젤 전기영동 이미지란 미리 정의된 가로축의 산성도 값 PI(iso-electric point)와 세로축의 단백질의 분자량 무게 MWt(Molecular Weight)로 구성된 이차원 젤 판에 생물학적 단백질 시료를 떨어뜨리고 전기영동을 이용해 단백질이 해당 PI와 MWt 값에 멈추었을 때 이를 스캐너를 통해 촬영한 영상이다[3]. 따라서 생물학적 시료의 수많은 단백질은 2차원 젤 전기영동 실험의 결과로 수백에서 수천 개의 단백질 반점을 포함한 2차원 이미지로 생성된다. 생물학자는 시간 및 환경이 다른 여러 조건의 2차원 젤 전기영동 이미지를 통해 분할된 단백질 반점의 그레이 강도(gray intensity) 값을 이용해 단백질의 발현 정보를 얻을 수 있으며, 영상의 x, y 좌표를 통해 단백질의 PI와 MWt값을 알아내고, 데이터베이스 검색을 이용해 단백질의 후보를 추측함으로써 단백질의 잠재적 기능을 유추할 수 있다[5, 8].

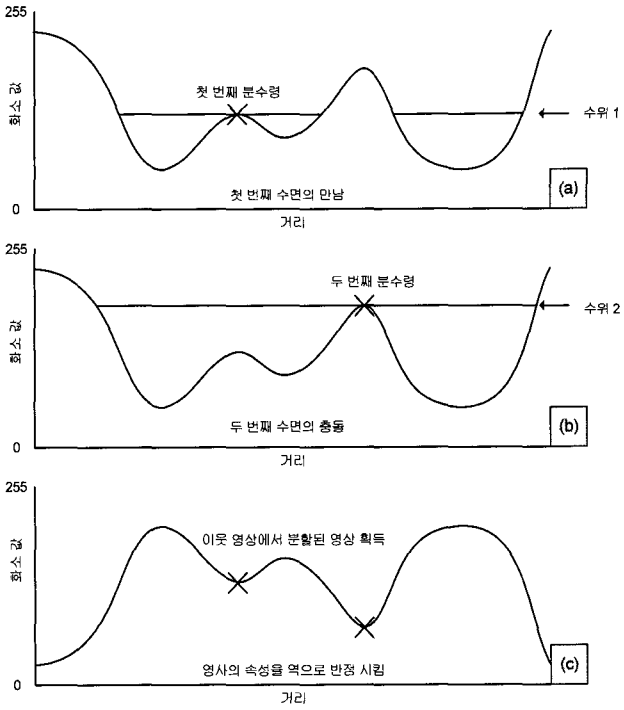
3.2 2차원 젤 전기영동 이미지 분석

서로 다른 2차원 젤 전기영동 이미지의 단백질 패턴은 너무 많은 반점이 존재하여 육안으로 분석하기는 어렵다. 현존하는 많은 상업용 소프트웨어에서는 2차원 젤 전디영동 이미지의 단백질 반점을 분리 하는데 많은 부분에서 수작업을 하기 때문에 시간과 노력이 많이 소요된다[2]. 효율적인 2차원 젤 전기영동 이미지의 분석은 디지털영상처리(Digital Image Processing) 기법을 사용하여 영상을 생성, 처리, 해석, 인식 및 패턴을 분석하는 전 분야를 의미한다. 예를 들어 배경화면과 잡음으로부터 단백질 해석과정에 의미 있는 반점 패턴을 분할하는 과정이나, 2차원 젤 전기영동 이미지에서 그레이 값으로 구성된 수백에서 수천 개의 반점의 반점외형요소(Blob shape momentum)를 찾는 작업에서 수많은 형태학(Morphological) 기법들이 쓰이고 있다[9].

4. Watersheds 영역 기반 분할

4.1 Watersheds transform

Watersheds 영역기반 분할 알고리즘은 지형학적 분야에서 연구되어온 것으로, 단순화된 영상에 대한 화소의 밝기 값의 경사를 지형학에서 고도(altitude)로 생각함으로써 영상처리에 응용된다. 이는 지역적 최소점(regional minima)을 통해 담수지역(catchment basin)을 구분하는 분수령(Watersheds)을 찾음으로써, 각 담수지역에 결합된 영역을 찾아내는 알고리즘이다. 이제까지 분수령을 찾기 위해 많은 알고리즘들이 제안되었으나, 가장 효과적인 방법은 침수 모의실험(Immersion simulation)을 이용한 방법이다[9, 10, 11, 12, 13]. 침수 모의실험이란 (그림 2)에서 보는 바와 같이 가장 낮은 수면으로부터 수면 높이를 점차로 증가시키면서, 지역적 최소점을 기점으로 호수(lakes/valleys)가 형성 되면서 담수 지역을 넓혀가는 과정이다[12, 13, 14, 15]. 수면이 계속



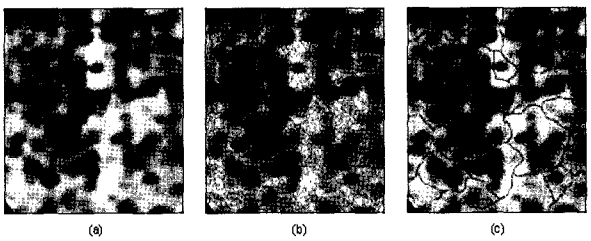
(그림 2) Watersheds 과정 (a) 침수 모의실험의 첫 번째 분수령 라인 형성 (b) 침수 모의실험의 무 번째 분수령 라인 형성 (c) 영상의 이웃 지점으로부터 분수령 라인을 통한 분할

높아지면서 서로 다른 담수지역에 접하게 되며, 접선의 경계 면에 분수령 라인(Watershed Line : Dam)을 쌓아서 담수지역을 구분한다.

4.2 Watersheds 알고리즘의 문제점

Watersheds 알고리즘은 지역적 최소점과 담수지역을 통해 이웃하는 분수령의 경계면을 쉽게 찾을 수 있다는 장점이 있지만, 2차원 젤 전기영동 이미지는 연속된 그레이 값을 가지고 있어, 지역적 최소점의 위치와 개수가 명확하지 않다는 문제점을 안고 있다.

따라서 영상에 지역적 최소점이 너무 많이 존재할 경우 그림 3의 원영상 (a)에 대하여 (b)와 같이 과다분할의 문제를 야기하며, 영상에 지역적 최소점이 너무 적게 존재하면



(그림 3) (a) 2차원 젤 전기영동 영상. (b) Watersheds에 의한 지역적 최소점이 너무 많아 생기는 고다 분할. (c) Watersheds에 의한 지역적 최소점이 너무 적어 생기는 과소 분할.

(그림 3)의 (c)와 같이 과소분할이 발생한다. 이와 같은 문제점을 해결하기 위해 내부마커(internal maker)와 외부마커(external marker)를 통해 지역적 최소점의 수와 위치를 결정해 준다[12]. 이때 적절한 마커의 위치를 찾기 위한 마커 제어기가 필요하게 된다. 또한 적절한 마커를 찾더라도, 분할된 영역의 평균 그레이 값을 통해 이웃하는 분할 영역과 병합 및 분할 작업을 수행하여야 한다[12].

5. Watersheds 기반 계층적 이진화(Watersheds-based Hierarchical Threshold) 분할

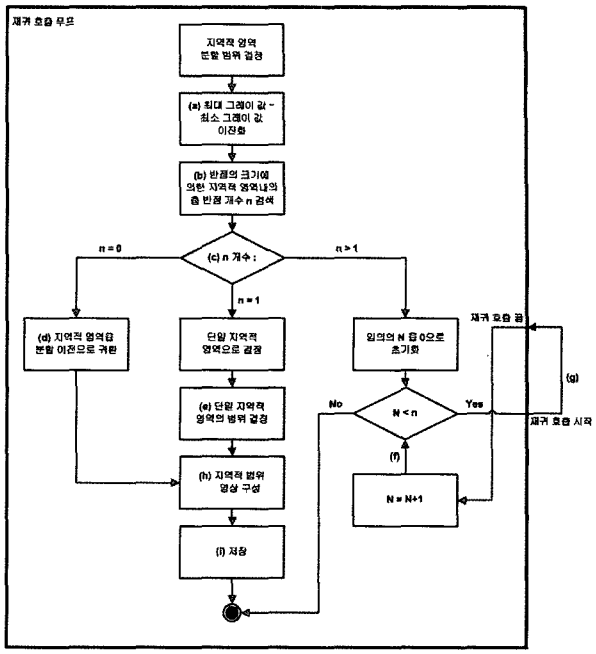
본 논문은 Watersheds 알고리즘의 과다분할과 과소분할의 문제점을 해결하고자, 2차원 젤 전기영동 이미지에 Watersheds 기반 계층적 이진화 분할 기법을 제안한다. Watershed 를 이용한 분할 대신 계층적 이진화 분할 기법을 적용한 이유는 "분할될 영역이 배경화면과 비교하여 명확히 분간할 수 있다면 Watersheds 분할 알고리즘은 전역 분할에 빠르다[6]."는 장점에 착안하였다. 즉 다차원의 이진화를 사용하여 연속된 그레이 값을 여러 단계의 불연속의 그레이 값으로 변화 시켜 2차원 젤 전기영동 이미지를 비연속된 그레이 값으로 변환 한다.

본 논문은 불연속한 영상을 만들기 위해 수식 (1)의 연산을 사용하여 배경화면을 제거시킴으로써 연속된 그레이 값을 어느 정도 상쇠 할 수 있지만, 단백질 반점 영역의 그레이 값도 함께 낮아지게 된다.

$$\text{출력 영상}(x,y) = 255 - (\text{원 영상에서 향상된 영상}(x,y) - \text{미디언 필터 영상}(x,y)) \quad (1)$$

이를 해결하려면 수식 (1)의 출력 영상에 일정 그레이 값을 갖는 화소에 대해서만 지역적 영역을 이진화 한다면 배경화면이 제거된 이미지는 불연속적인 그레이 값으로 만들 수 있다. 이는 (그림 6)의 (a)와 같은 영상으로 나타나며, 이 영상에 일정 그레이 값을 갖는 화소에 대해서 재귀적 이진화를 적용한다. 이러한 재귀적 이전화의 결과에 Watersheds 기법을 적용할 경우 (그림 6)의 (b), (c), (d)와 같은 비 연속적 영역을 생성할 수 있다.

(그림 4)는 계층적 이진화 분할 기법의 흐름도를 보여주고 있다. Watersheds 기반 계층적 이진화 분할은 상위 전역적 영역과 하위 지역적 영역으로 표현 한다. 계층적 이진화 분할의 과정은 분할의 범위가 결정된 전역적 영역에서 먼저 (a)단계에서는 반점의 최대 그레이 값과 최소 그레이 값의 차만큼을 이진화를 적용한다. 이진화 된 이미지는 (b)단계에서 전역적 영상(global image) 내에 총 n 개의 하위 지역적 영역(local boundary)이 존재 하는지 계산한다. 이때 n은 단백질 반점의 최대 사이즈, 최소 사이즈에 의해 계산된다. 최대, 최소 사이즈는 5장에서 설명한다. (c)단계에서는 n의 개수에 따라 세가지 하부 항목으로 나눈다. n의 개수가 0일 때는 (d)-(h)-(i)단계를 거치며, 이진화 결과 영상에 반점이



(그림 4) 재귀 호출을 이용한 계층적 이진화 분할 흐름도

존재 하지 않음을 뜻한다. 따라서 이전 단계의 이진화 결과의 지역적 영역 영상을 저장한다. n이 1일 때는 (e)-(h)-(i) 단계를 거치며, 이진화 결과 영상에 1개의 반점 영역만이 존재하는 것을 의미한다. 이때 이진화 결과를 단일 지역적 영역으로 확정하고, 지역적 범위 영상을 저장한다. 마지막으로 (c)-(f)-(g)단계는 n이 1 보다 큰 경우에 해당하며, 지역적 영역 내에 2개 이상의 반점이 존재하는 것을 의미하여 임의의 N을 0으로 초기화 하고 정해진 n의 개수까지 재귀호출로 이진화 과정을 반복한다.

위와 같은 알고리즘을 반복하면 진역 이미지는 여러 개의 지역적 영역으로 분할되며, 재귀 호출에 의해 하나의 상위 지역적 영역이 다른 여러 개의 하위 지역적 영역으로 계층적으로 분할되게 된다. 본 논문에서는 이를 Watersheds 기반 계층적 이진화 분할이라 정의 한다. (그림 5)는 이에 대한 세부적인 알고리즘을 나타낸 것이다.

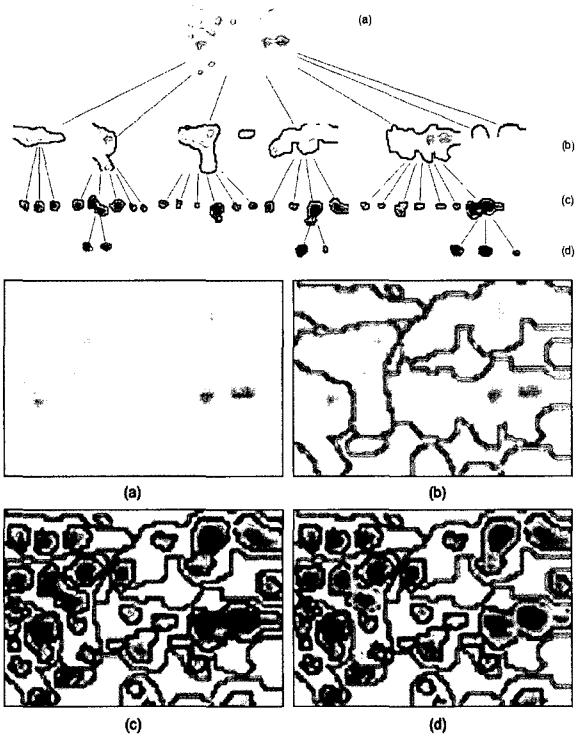
(그림 6)은 Watersheds 기반 계층적 이진화 분할을 적용한 예제이다. (a)는 2차원 젤 전기영동 이미지에 이미지 향상을 적용하고 수식 1에 의해 배경화면을 제거한 영상이다. (b)는 (a)의 영상에 제안한 Watersheds 기반 이진화를 1회 적용하여 하나의 영역에 여러 개의 지역적 영역이 들어가 있는 영상이다. (c)는 (b)의 영상에 재귀호출 함수 사용해 Watersheds 기반 이진화를 2회 적용한 영상이다. 1회 적용한 분할은 빨간색 영역을 말하며 (그림 6)의 (b)를 나타낸다. 2차 이진화 분할 과정은 (그림 6)의 (b) 의 빨간색 영역에 오버레이 된 파란색 영역이다. 3차 이진화 분할 과정은 (그림 6)의 (c)에 4차 이진화 분할을 거친 파란색 영역에 오버레이 된 녹색 영역이 최종 이진화 분할이다. 이와 같이 Watersheds의 문제점을 해결하기 위한 마커 제어기를 사용하지 않고서도 반점 영역을 명확하게 구분함을 알 수 있다.

```
threshold level initialize

apply threshold

  if does not exist boundary area                    /* n = 0 인 경우. */
    apply previous segment level
    save boundary area
  end
  else if have 1 boundary area
    if boundary area <minimum spot size /* 반점의 최소 사이즈보다 작
                                          을 때 */
      delete boundary area              /* 반점으로 인식 하지 않음 */
    end
    else if boundary area <= maximum spot size
        if boundary area <= previous boundary area  /* n = 1 인 경우*/
            save boundary area
        end
        else                                        /* n > 1 인 경우*/
            set change threshold level
            go to recursive                         /* if N < n 인 경우 */
        end
    else
     Set change threshold level
     go to recursive
    end
  else
  recursive:
    recursive to search as boundary number
  end
```

(그림 5) Watersheds 기반 계층적 이진화 분할 알고리즘



(그림 6) (a) 영상 향상 및 배경화면 제거 이미지. (b) Watersheds 기반 1차 이진화 분할을 적용한 영상 (c) Watersheds 기반 2차 이진화 분할을 적용한 파란색 영역 (d) Watersheds 기반 3차 이진화 분할을 재귀호출한 녹색 영역

6. 실험 및 결과

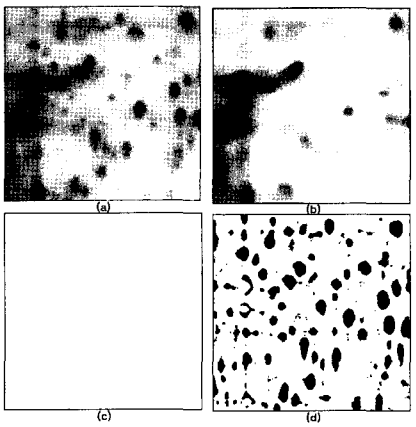
6.1 Watersheds 기반 계층적 이진화 분할 알고리즘의 적용

2차원 젤 전기영동 이미지의 경우 반점이 배경화면과 완벽하게 분리 되어 있지 않다. 즉, 단백질 시료가 젤 판에 전기연동을 통해 움직이기 때문에 지저분한 그림자를 남기게 된다. 따라서, 2차원 젤 전기영동 이미지는 특히 다른 종류의 영상과 비교해 많은 잡음(noise)이 존재하게 된다. Watersheds 분할 알고리즘의 경우에도 3.2 (그림 3)에서 언급했던 바와 같이 단백질 시료의 움직임에 의한 잡음으로 아주 많은 과다분할과 과소분할이 발생한다.

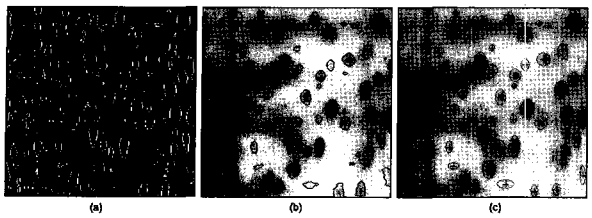
제안하는 기법은 잡음에 대해 정확한 분할을 저해하는 요소들을 최소화하기 위해 영상 향상과 배경화면 제거를 적용하였다. 또한 단백질 반점의 계층적 이진화 분할 알고리즘을 위해 다음과 같은 변수를 설정하였다.

최대 반점 크기(maximum spot size)는 단백질 반점의 최대 영역 크기를 결정한다. 최소 반점 크기는(minimum spot size)는 단백질 반점의 최소 영역 크기를 결정한다. 희미한 반점(faint spot)은 반점의 최소 그레이 값을 결정한다. 그레이 값의 차(distance value from maximum gray level to minimum gray level of spot)는 분할 하고자 하는 구간에서 최대 그레이 값과 최소 그레이 값의 차이를 나타낸다. 그레이 경사(slop of gray level)는 반점에서 급격한 그레이 경사를 가지는 영역을 경계선으로 결정한다.

(그림 7)의 (c)영상은 수식 (1)에 정의된 출력 영상을 구하기 위하여, 원영상 (a)와 미디어 필터 영상 (b)에 대하여 배경화면을 제거한 차를 나타내며, (d)는 제안하는 Watersheds 계층적 이진화 분할 알고리즘을 통한 영상이다. 그러나 (그림 6)과 다르게 Watersheds 알고리즘을 한번도 적용하지 않았다. Watersheds 알고리즘을 한번도 적용하지 않은 이유는 단백질 반점의 분할 시간을 줄이기 위해 동일한 결과를 내는 Watersheds 과정을 1회로 한정함으로써 시간의 효율성을 얻기 위함이다.



(그림 7) (a) 2차원 젤 전기영동 원영상. (b) (a) 원영상에 미디언 필터를 처리한 영상. (c) 수식(1)을 이용해 배경화면을 제거한 영상. (d) 배경화면이 제거된 영상 (c)에 대하여 계층적 이진화 분할 기법을 적용한 영상.



(그림 8) (a) 그림 7의 (d)에 기울기 연산에 의한 영상. (b) Watersheds 알고리즘을 적용하 분수령 라인을 구한 영상. (c) 식별성을 높이기 위한 단백질 반점의 타원 경계선을 표현.

(그림 8)의 (a)는 (그림 7)의 (d)영상에 기울기 연산자(gradient operator)를 취한 영상이며, Watersheds 알고리즘을 적용하여 (b)와 같은 결과를 얻을 수 있다. 그림 8의 (c)는 결과 영상에 단백질 반점의 식별성을 높이기 위해 타원 경계선(Ellipse Boundary)과 중심 좌표를 나타낸 영상이다.

6.2 연구 결과

우리는 본 논문의 실험결과의 타당성을 검토하기 위해 맨체스터 대학(The University of Manchester)대학의 Imaging Science and Biomedical Engineering의 Mike Rogers의 실험 결과를 근거로 하였다. Mike Rogers는 범용 2차원 젤 전기영동 영상을 분석하기 위해 범용으로 사용되는 툴인 ImageMaster 2D Elite, Melanie 3D, PSQuest, Progenesis, Z3의 True Positive와 TPF(True Positive Fraction)의 검출 결과를 측정하였다. 이에 대한 결과는 <표 1>과 같다.[16]

〈표 1〉 2차원 젤 전기영동 분석을 위한 범용 툴의 단백질 반점 추출 결과

패키지	Accurancy	TPF	FP
ImageMaster	3.15%	85.1%	40
MElanie	3.40%	81.5%	26
PDQuest	5.02%	78.2%	20
Progenesis	3.46%	84.6%	26
Z3	4.46%	68.2%	30

그러나 이러한 방법에 의한 측정은 <표 1>과 같은 파라미터를 가지고는 정확한 측정이 어렵다. 즉 TPF나 FP만을 가지고 Accuracy를 측정하는 것은 정확한 측정 결과를 도출하기 힘들다. 본 논문에서는 Watersheds 기반 계층적 이진화 분할 알고리즘의 성능을 평가하기 위해 다음과 같은 테스트를 하였다. 총 115개의 단백질 반점이 있는 것으로 알려진 2차원 젤 정기영동 이미지에서 Watersheds 계층적 이진화 분할 알고리즘을 적용하여 <표 2>와 같은 TP(True Positive), TN(True Negative), FP(False Positive), FN (False Negative)의 개수를 측정하였다. 측정된 각 개수에 의해 신뢰도(Confidence)와 정확도(Sensitivity)을 구할 수 있다. 이를 이용하여 더욱 정확한 단백질 반점의 검출 확률을 알 수 있다.

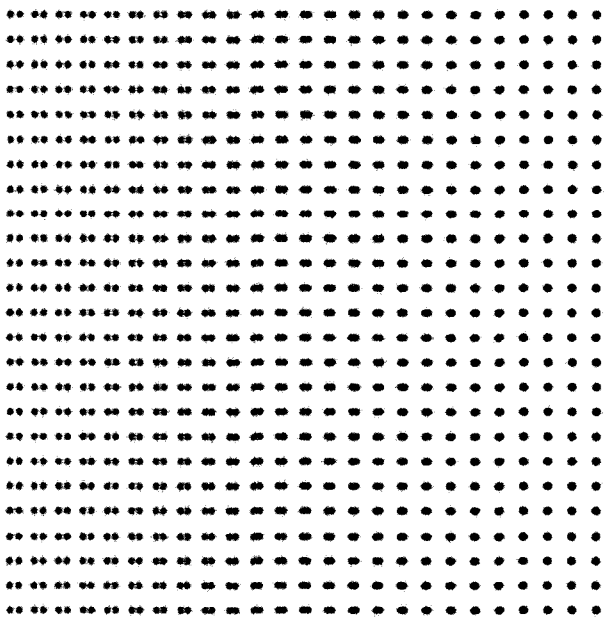
〈표 2〉 계층적 이진화 분할 알고리즘을 이용한 신뢰도 측정

Index	Total Spot	TP	TN	FP	FN	Specificity	Confidence	PS-PDP
Gel1	115	106	9	1	0	0.900	0.990	0.913
Gel2	120	110	10	2	0	0.833	0.982	0.901
Gel3	118	108	10	1	0	0.909	0.990	0.907
Total	353	324	29	4	0	0.878	0.987	0.907

〈표 3〉 맨체스터 대학의 Imaging Science and Biomedical Engineering의 Mike Rogers가 실험한 Silver Stained Gel Image의 계층적 이진화 분할 알고리즘을 이용한 신뢰도 측정

Index	Total Spot	TP	TN	FP	FN	Specificity	Confidence	PS-PDP
Silver Statin Gel1	964	947	14	3	0	0.8235	0.9968	0.9824
Silver Statin Gel2	1206	1194	10	2	0	0.8333	0.9983	0.9900
Total	2170	2141	24	5	0	0.8275	0.9976	0.9866

Confidence = TP / (TP + FP), Specificity = TN / (TN + FP)
PS-PDP : Protein Spot Pre-Detection Probability = (TP + FN) / (TP + FP + FN + TN)



(그림 9) Gassian Noise 0.05, Overlapping 이미지.

본 논문의 실험 결과의 특성을 살피기 위해 Overlapping된 Gaussian 노이즈 반점 925개 이미지를 생성하였다.

가로×세로의 반점의 개수가 25×25이며, 좌측 12번째 까지는 Overlapping된 영상에 가까우며, 13번째는 부터는 겹치기 시작하여 우측으로는 하나의 이미지로 합쳐진다. 13번째까지 이미지가 합쳐진다면 총 925개의 반점이 존재하게 된다.

〈표 4〉 Overlapping Image 인식률 측정 결과

Overlapping Image	Total	TP	TN	FP	FN	Specificity	Confidence	PS-PDP
1'th-12'th column	600	580	20	0	0	1	1	0.9667
13'th column	25	12	13	0	0	1	1	0.48
14'th-25'th column	300	298	2	0	0	1	1	0.9933
Total 25 column	925	890	35	0	0	1	1	0.9622

좌측 12번째 까지 625개 중에 20개를 반점으로 인식 하지 못하였으며, 13번째 열은 25개중 13개를 인식하지 못하였다. 우측 14번째부터 25번째 반점에 대해서는 단 2개만 반점으로 인식하지 못하였다. 또한 FP에 대한 오류가 존재하지 않아 Specificity와 Confidence의 측정이 무의미 하다. 특성으로 보아 본 시스템은 서로 가까이 붙어 있는 반점일수록 인식률이 낮은 것으로 평가할 수 있다. 그러나 실제 젤 영상에서는 이러한 Overlapping에 대한 반점이 하나로 존재하는 반점보다 그 비율이 극소수이기 때문에 실제 영상에서는 인식률이 높게 나타나는 것으로 판단된다.

7. 결 론

2차원 젤 전기영동 이미지는 단백질 반점 패턴을 영상처리를 통해 목표 단백질을 찾기 위한 방법이다. 영상처리 관점에서 2차원 젤 전기영동 이미지의 경우 전역적 영역에서 일정한 패턴을 이루지 않기 때문에 기존의 영상 분할 방법으로 처리할 때 많은 한계점을 가지고 있다. 최근에 연구된 Watersheds 분할 알고리즘의 경우 전역 분할에 빠르다는 장점이 있지만, 영상의 연속적인 그레이 값에 의해 마커의 적절하지 못한 선택은 과다분할과 과소분할의 문제점을 야기한다.

본 논문에서는 Watersheds 알고리즘의 마커 제어기의 역할을 대체하는 계층적 이진화 분할 기법을 적용하여 Watersheds 알고리즘의 문제점을 해결하였다. 계층적 이진화 분할의 역할은 연속적인 그레이 값의 영상을 단계적으로 불연속한 그레이 값 영상으로 변환함으로써 그레이 값에 의한 병합 및 분할 과정이 제거되어 과다분할과 과소분할에 대한 문제점을 해결하였다.

또한 기존의 PDQuest, ImageMaster, Melanie, Progenesis, Z3와 비교해 볼 때, 본 논문의 Watersheds 기반 계층적 이진화 시스템은 간략한 방법에 비해 거의 같은 반점 검출 성능 또는 향상된 성능을 보였다.

제안한 방법을 이용한 단백질 반점 분할 결과는 단백질 반점의 양적 발현 양을 반영하고, 추후 연구로써 두 장 이상의 2차원 젤 전기영동 단백질 반점의 발현 양 교차비교를 통해 단백질 발현의 변화와 새로운 단백질의 발견에 관한 연구가 현재 진행 중이다.

참 고 문 헌

[1] Blackstock, W. P. and M. P. Weir. "Proteomics : quantitative physical mapping of cellular proteins", *Trends Biotechnol.* Vol.17, No.3, pp.121-127. Mar., 1999.

[2] Wilkins, M.R., Williams, K.L., Appel, R.D. and D.F. Hochstrasser, "Proteome research : New frontiers in functional genomics", *Springer-Verlag*, Berlin, ISBN 3540627537, Nov., 1997.

[3] Kuster, B. and M. Mann, "Identifying proteins and post-translational modifications by mass spectrometry", *Current Opinions in Structual Biology*, Vol.8, No.3, pp.393-400, Jan., 1998.

[4] 김영호, 원용관, "발현 단백체학 : 2DGE의 디지털 영상 분석 및 포인트 매칭 기술", *한국정보과학회 바이오정보기술 연구회지*, Vol.1, No.1, pp.13-24, Oct., 2003.

[5] Daniel C. Liebler, "Introduction to Proteomics : Tools for the New Biology", *Humana Press*, Chapter 4, pp.31-54, Dec., 2001.

[6] L Vincent and P Soille, "Watersheds in digital spaces : an efficient algorithm based on immersion simulations", *Pattern Analysis and Machine Intelligence, IEEE Transactions*, Vol.13, No.6, pp.583-598, Jan., 1991.

[7] Rafael C. Gonzalez and Richard E. Woods "Digital Image Processing", Second Edition, *Prentice Hall*, Chapter 10, pp.617, Jan., 15, 2001.

[8] Lars Pedersen, "Analysis of Two-dimensional Electrophoresis Gel Images" *Information and Mathematical Modeling*, Revision 1.36 Exp, pp.28-72, Feb., 2002.

[9] S Beucher, "Extrema of grey-tone functions and mathematical morphology", *In Proc. of the Colloquium on Math. Morp., Stereol. and Image Analysis, Prague, Tchecoslovaquia*, pp.59-70, Sep., 1982.

[10] S Beucher and C Lantuejoul, "Use of watersheds in contour detection", *International Workshop on Image Processing : Real-Time Edge and Motion Detection/ Estimation*, Rennes, France, pp.17-21, Sep., 1979.

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

[12] F. Meyer and S. Beucher, "Morphological segmentation", *J.Visual Commun. Image Representation*, Vol.1, No.1, pp. 21-46, Sep., 1990.

[13] Stanislav L. Stoev and Wolfgang Straßer, "Extracting Regions of Interest Applying a Local Watershed Transformation", *Visualization 2000. Proceedings*, pp.21-28, 8-13 Oct., 2000.

[14] J. B. Roerdink and A. Meijster, "*The watershed trans-*

form : definition, algorithms and parallelization strategies," Fundamental Information, Vol.41, No.1-2, pp.187-228, 2001.

[15] Wang, D., Labit, C., and Ronsin, J. "Segmentation-based motion-compensated video coding using morphological filters" *IEEE Transactions on Circuits and Systems for Video Technology*, Vol.7, No.3, pp.549-555, Jun., 1997.

[16] M. Rogers, J. Graham and R.P. Tonge "Using Statistical Image Models to Conduct Objective Evaluation of 2D Gel Image Analysis Packages." *In Proceedings of 5th Siena Meeting From Genome To Proteome : Functional Proteomics, Siena, Italy. Proteomics 2003*, No.6, pp.879- 886, Sep., 2002.

김 영 호

e-mail : yhkim@grace.chonnam.ac.kr

2000년 조선대학교 제어계측공학과(공학사)

2002년 전남대학교 컴퓨터공학과(공학석사)

2004년 전남대학교 컴퓨터공학과(박사수료)

2004년 (주)도울정보기술 연구원

관심분야 : 영상처리, 패턴인식, 바이오인포메틱스, 네트워크

김 정 자

e-mail : j2kim@chonnam.ac.kr

1985년 전남대학교 자연과학대학 계산통계학과(이학사)

1988년 전남대학교 자연과학대학 전산통계학과(이학석사)

2002년 전남대학교 자연과학대학 전산통계학과 (이학박사)

현 재 전남대학교 BioIT 파운드리센터 연구교수로 재직중

관심분야 : 데이터베이스, 데이터 마이닝, 바이오 인포매틱스

김 대 현

e-mail : daehkim@chosun.ac.kr

1996년 조선대학교 제어계측공학과(학사)

1998년 조선대학교 제어계측공학과(공학석사)

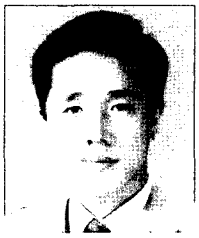
2001년 조선대학교 제어계측공학과(공학박사)

2001년~2004년 (주)남인 선임연구원

2001년~2003년 조선대학교 기계공학과 겸임조교수

2004년~현재 (주)수성전자 연구소장

관심분야 : 임베디드 시스템, 햅틱 렌더링, SoC 설계, 지능형 비젼



원 용 관

e-mail : ykwon@grace.chonnam.ac.kr

1986년 한양대학교 전자공학과(학사)

1991년 Univ. of Missouri, 컴퓨터공학과 (공학석사)

1995년 Univ. of Missouri, 컴퓨터공학과 (공학박사)

1991년~1995년 Univ. of Missouri, Research/Teaching Assistant

1995년~1996년 한국전자통신연구원

1996년~1999년 한국통신 연구개발본부

1999년~현재 전남대학교 컴퓨터공학과 부교수

관심분야 : 생물정보학, 네트워크 관리, 데이터마이닝, 정보보호, 컴퓨터 비젼