다차원 클러스터링 기반의 단백질 2DE 이미지에서의 자동화된 기준점 추출 방법

심 정 은[†] · 이 원 석[††]

요 약

2DE는 조직 내의 단백질을 규명하는 단백질 분리 기술이다. 그러나 2DE 이미지는 실험 조건, 스캐닝 상태와 같은 환경에 민감하게 영향을 받는다. 이러한 이미지간의 변화를 극복하기 위해서 사용자는 각각의 서로 다른 이미지에 수동으로 기준점을 입력해주어야 한다. 그러나 이 과정은 에러를 발생시키며 긴 시간을 요구하는 작업으로, 빠른 분석에 장애 요인이 된다. 따라서 본 논문에서는 기준점 프로파일에 기반 하여 기준점을 자동으로 추출하는 방법을 개발하였다. 기준점 프로파일은 이미 확인된 이미지들의 기준점들에 대한 클러스터링 방법을 통하여 생성하며, 각 클러스터의 다양한 속성을 정의한다. 새로운 이미지가 입력되면 기준점의 후보 스팟들을 대상으로 프로파일과 비교하여 기준점을 추출한다. 그리고 A^*알고리즘을 이용하여 기준점 선정 과정을 최적화한다. 본 논문에서는 실제 사람의 간 조직 이미지를 이용하여 기준점 추출 방법의 성능을 분석하였다.

키워드 : 프로테옴 인포매틱스, 데이터마이닝, 클러스터링, 2DE

Automated Method of Landmark Extraction for Protein 2DE Images based on Multi-dimensional Clustering

Jung Eun Shim[†] · Won Suk Lee[††]

ABSTRACT

2-dimensional electrophoresis(2DE) is a separation technique to identify proteins contained in a sample. However, the image is very sensitive to its experimental conditions as well as the quality of scanning. In order to adjust the possible variation of spots in a particular image, a user should manually annotate landmark spots on each gel image to analyze the spots of different images together. However, this operation is an error-prone and tedious job. This thesis develops an automated method of extracting the landmark spots of an image based on landmark profile. The landmark profile is created by clustering the previously identified landmarks of sample images of the same type. The profile contains the various properties of clusters identified for each landmark. When the landmarks of a new image need to be found, all the candidate spots of each landmark are first identified by examining the properties of its clusters. Subsequently, all the landmark spots of the new image are collectively found by the well-known optimization algorithm A^*. The performance of this method is illustrated by various experiments on real 2DE images of mouse's brain-tissues.

Key Words : Proteome Informatics, Data Mining, Clustering, 2DE(Two-Dimensional Electrophoresis)

1. 서 론

단백질체학 연구는 주어진 셀이나 조직, 생물체에 표현된 단백질의 프로파일에 대한 조직적인 분석을 다룬다. 특히, 단백질체학 연구의 주요 목적은 임의의 조직에서 어떤 단백질이 발현되며, 특정 조건 하에서 단백질이 어떻게 상호 작용 하는 지를 분석하는 것이다. 이런 목적에서, 임의의 조직에서 특정 조건에 따른 단백질의 발현량의 변화 분석은 조직의 기능 장애를 일으키는 단백질의 도출에 있어서 핵심 이슈 중 하나이다. 이처럼 발현량의 변화를 통한 단백질 분석 방법을 발현 단백질체학(expression proteomics)이라고 한다. 발현 단백질체학에는 크게 두가지 분석 방법이 존재한다. 이차원 전기영동 방법(Two Dimensional gel Electrophoresis, 2DE)과 Non-2DE의 두 가지 기술이 사용되며, 전자는 전기영동 방식을 사용하여 임의의 조직 내에 존재하는 단백질들을 분리한다[1, 2]. 후자는 주로 ICAT(isotope coded affinity tag)[3]이나 MCAT(mass-coded abundance tag)[4]와 같은 특정의 친화 태그를 붙이는 방법과 LC-MS (liquid chromatography-mass spectrometry)를 사용하며, 자

† 준 회 원 : 연세대학교 컴퓨터과학과 박사과정
†† 종신회원 : 연세대학교 컴퓨터과학과 교수
논문접수 : 2005년 2월 23일, 심사완료 : 2005년 9월 21일

동화에 유용하고 처리율이 높다. Non-2DE 기술이 보다 정확한 결과를 제공하지만, 여전히 2DE 기술이 가격 효율성 등으로 인해 단백질의 발현 패턴 분석에 주로 이용되는 기술이다.

단백질 이미지에 나타나는 스팟(spot)의 패턴은 해당 샘플을 채취할 때 약의 투여 정도, 면역성 정도, 주변 환경 조건에 따라 차이가 있으며 이미지를 생성하는 전기 영동 실험에서도 처리 조건, 젤의 특성, 이미지 스캔 작업의 오차로 인해 동일한 샘플의 젤 이미지간에 차이가 발생할 수 있다. 그리고 한 이미지에는 평균적으로 1000개, 최대 3000개 이상의 스팟이 나타나기 때문에, 두 개 이상의 이미지에 대해서 나타난 스팟들의 양상을 비교한다는 것은 상당히 어려운 일이다. 따라서 두개 이상의 이미지를 비교하기 위해서는 좌표축 보정이 필요하다. 이 때 사용되는 상대적 위치가 유사한 스팟이 해당 기준점 스팟이다. 유사 샘플들에 대한 기준점 집합은 각 이미지 전체에 고르게 분산되어야 하며, 육안으로 모든 기준점들을 지정해주어야 하므로 100%의 정확도를 보장할 수 없다. 따라서 정확하고 객관적인 판단에 의해 이미지 상의 기준점을 추출할 수 있는 기준점의 자동 추출 방법이 필요하다.

많은 실험실에서는 이미 상용화되어 있는 Melanie[5] 및 Progenesis[6]등과 같은 다양한 이미지 분석 소프트웨어를 사용하고 있다. 기존의 대부분의 이미지 분석 소프트웨어는 두 개 이상의 이미지의 차이를 보정하기 위해 기준점 또는 관심영역(AOI : Area Of Interest)을 임의로 지정해야 한다. 두 개의 이미지를 비교할 때 기준점은 두 이미지의 스팟들의 위치를 보정하여 상대적 위치 값을 계산하는 기준이 되며 관심영역은 각 이미지의 분석 관심 영역으로 상대적 위치가 같은 영역을 설정하여야 의미 있는 분석결과를 얻을 수 있다. 그러나 기준점과 관심영역의 입력은 사용자 임의로 선택하므로 입력 내용에 따라 분석 결과의 정확도에 큰 차이가 있을 수 있으며 잘못된 입력이 들어갔을 경우 분석 결과가 틀릴 수 있고 상당수의 기준점을 모든 대상 이미지에 지정해야 하므로 입력 과정이 번거롭다. 따라서 본 논문에서는 기존에 사용하는 다양한 이미지 분석 소프트웨어를 이용해 스팟 검출(Spot Detection) 과정을 거친 데이터를 기반으로 대상 이미지들을 비교할 수 있는 기준이 되는 기준점을 자동으로 추출하는 알고리즘을 제안하고 실험을 통하여 알고리즘의 성능과 효과를 평가한다.

본 논문의 2장에서는 기존의 상용 소프트웨어의 특징 등의 기존 연구를 설명하며, 3장은 기준점 추출을 위한 프로파일 생성 방법을 제안하며, 4장에서는 A^*탐색 기법을 이용한 기준점 추출 알고리즘을 제안한다. 5장에서는 본 논문에서 제안하는 알고리즘의 성능을 실제 단백질 이미지 데이터에 대해 적용한 실험 결과를 제시하고 알고리즘의 성능을 분석한다. 6장에서는 결론 및 향후 연구 방향에 대해 기술한다.

2. 관련 연구

2DE 이미지에는 유전자 이미지와 단백질 이미지가 있으며 유전자의 특성과 단백질의 특성이 서로 상이하므로 동일한 이미지 분석 알고리즘을 적용할 수 없으며 이미지의 데이터의 특성과 분석 목적에 적합한 알고리즘을 선택하여 적용하여야 한다.

2.1 디루니 삼각 분할(Delaunay triangulation)[7]을 이용한 DNA 이미지 분석

디루니 삼각 분할은 유전자 이미지를 분석하기 위한 방법으로 최소 신장 트리(Minimal Span Tree)를 생성하고 Voronoi Diagram[8]을 통해 디루니 삼각 분할을 한다. 디루니 삼각 분할을 이용하여 DNA 이미지에 있는 스팟의 지리적 위치를 RLGS(Restriction Landmark Genomic Scanning)라는 프로파일로 생성한다. 두 이미지를 비교하기 위해 두 이미지에 대한 RLGS 프로파일을 비교 분석한다[9]. DNA 2DE 이미지는 단백질 이미지와는 달리 패턴이 비교적 정형화되어 있다. 인간의 유전체는 30억개의 염기로 구성되어 있으나 중 0.005~0.001%에 해당하는 10만여개의 염기 차이가 사람마다 각각 다른 차이를 만들어낼 정도로 개인의 유전체의 차이는 크지 않다. 따라서 두 이미지는 유사한 형태로 삼각 분할 되기 때문에 RLGS 프로파일 역시 유사한 패턴으로 나타나므로 두 이미지의 비교는 상대적으로 용이하다. 단, RLGS 프로파일을 만들때 초기 스팟을 지정해 주어야 프로파일 비교를 수행할 수 있다는 단점이 있다.

2.2 CAROL : 디루니 삼각 분할을 이용하는 단백질 이미지 분석

단백질 이미지는 정형성과 재현성이 상당히 낮다. 따라서 단백질 이미지에서 RLGS 프로파일을 생성하고 비교하기에는 유전자 이미지 분석보다 더 많은 어려움이 따른다. 디루니 삼각 분할을 이용하는 단백질 이미지에서의 분석 방법은 크게 두 가지를 살펴볼 수 있다. 첫 번째 방법은 지역 패턴(local pattern)만을 찾는 방법이다. 이것이 CAROL[10]이라는 시스템이다. CAROL은 전체 스팟을 매칭하는 대신에 스팟들 중에 부분 패턴의 특성만을 모델링하여 특정 영역에 국한된 패턴만을 찾고, 지역 패턴으로부터 전체 패턴 매칭을 수행하는 방법이다. 그러나 기본적으로 기존의 방법들은 하나의 기준 이미지를 선택하고 나머지 대상 이미지를 기준 이미지와 비교하기 때문에, 기준 이미지에 포함된 오류들이 분석에 그대로 반영될 수밖에 없다. 두 번째 방법은 이미지 영역을 분할하여 분석하는 방법이다. 전체 이미지를 분할할 때 비교하려는 이미지의 상대적 위치가 비슷하도록 나눠준다. 상대적 위치가 비슷하도록 이미지를 분할하면 이질적인 이미지들이 유사한 규격으로 재조정 되어 결국 기준점을 입력하는 것과 마찬가지의 효과를 준다. 비록 영역을 지정하는 것이 기준점을 직접 입력하는 것보다는 사용자의 부담을 적게 하지만 영역을 나누는 방법 역시 결국은 사용자가 꾸준히 입력해야 하므로 다량의 이미지 분석하기는 어렵다.

2.3 상용 소프트웨어에서의 단백질 이미지 분석

단백질 이미지 분석에 사용되는 대표적인 소프트웨어로

Melanie와 Progenesis 등을 들 수 있다. Melanie series[5]는 이미지 분석에 사용되는 상용화된 대표적인 소프트웨어로 서로 다른 두개 이상의 이미지 내에 존재하는 스팟을 검출하고 검출된 스팟들 중, 서로에 부합하는 스팟들을 매칭하는 과정을 수행한다. 특히 스팟 매칭을 위해서는 비교의 기준이 되는 기준 이미지를 선정하며, 각 이미지들을 서로 비교하기 위해 이미지의 비교 축이 될 수 있는 기준점들을 사용자가 수동적으로 입력해야 한다. 또한 각 이미지에 대해 입력하는 기준점은 사용자의 주관적 판단에 의해 결정되며 동일 단백질에 해당하는 스팟으로 이미지 전반에 걸쳐 고르게 퍼져 있어야 좋은 이미지 분석 결과를 얻을 수 있다. 따라서 Melanie series는 다량의 분석이 어렵고 정확도가 떨어질 수 있는 단점을 갖는다. 이 단점을 보완하기 위해 개발된 소프트웨어가 Progenesis[6]이다. Progenesis에서는 사용자는 입력 매개변수인 관심영역을 설정해야 한다. 사용자는 이미지 분석을 시작하기 전에 모든 이미지에서 분석을 하려는 특정 사각 영역을 지정해 주어야 한다. 그러나 관심영역 역시 모든 이미지에 대해 입력되어야만 하고 관심 영역이 이미지마다 제대로 대응되지 않을 경우, 스팟 매칭의 정확도를 떨어뜨린다. 따라서 Progenesis 역시, 사용자의 판단에 의해 정확하게 입력해야 하는 정보이기 때문에 다량의 이미지 동시 분석에는 적합지 않으며, 본 논문에서는 Melanie III의 기준점 입력 과정의 불편함과 부정확성을 극복하기 위한 기준점 추출 방법을 제안한다.

3. 기준점 프로파일의 모델링

유사한 샘플 그룹에 나타난 단백질 스팟의 각 기준점의 공통된 특성을 파악하기 위해 분석자에 의해 이미 검증된 학습용 이미지 내의 기준점들로 각 기준점의 특성을 요약한 프로파일을 생성한다. 학습용 이미지는 생물학자들에 의해 눈으로 확인되고 Matrix-Assisted Laser Desorption/Ionization(MALDI) 등의 생물학적 실험에 의해 검증되었으며, 이 때 기준점의 특징을 보다 상세하게 요약하기 위해서 본 논문은 클러스터링 기법을 기준점 학습 과정에 도입하였다. 클러스터링 후 생성된 각 클러스터는 평균과 표준편차, 스팟의 범위를 모델링하여 기준점의 특징을 담은 기준점 프로파일을 생성한다.

3.1 기준점과 기준점 집합의 속성

이미지에는 수많은 스팟들이 존재하며 각 스팟은 위치 정보와 농도 정보의 속성을 갖는다. 단일 스팟은 이미지 상에서의 위치 정보 (x,y)를 갖고 발현량을 나타내는 농도 정보로 OD, Vol을 갖는다. OD는 2차원의 관점에서 산출된 스팟의 농도 값이며 Vol은 2차원 이미지를 3차원으로 보정하여 3차원 도형의 부피 값을 통하여 계산된 값이다. 그러나 이미지 생성 과정에 있어서 전반적으로 이미지가 짙은 농도로 나오거나 또는 옅은 농도로 생성될 수 있으므로 절대적인 농도 값에 기반을 두어 이미지들의 스팟을 비교할 수 없다. 따라서 이미지의 농도 값 역시 정규화 과정이 필요하다. 정규화를 위해 이미지 내의 모든 스팟의 전체 농도 값을 100으로 했을 경우 하나의 스팟의 농도 값의 비율을 백분율로 표현한다. OD와 Vol에 대해 이러한 과정에 의해 보정된 값이 $\%OD, \%Vol$이다.

[정의 1] 스팟의 속성

n개의 이미지 집합 M = { $m_1, ..., m_n$ }에 대해 이미지 m_i의 스팟 집합 M_i = { $x \mid x$는 이미지 m_i에 존재하는 스팟 }이고 이미지 m_i의 스팟 s의 위치 정보인 좌표값 (x,y)를 ν로 표현할 경우 각각의 스팟 s는 속성 값 $property(s) = (\nu, \%OD, \%Vol)$을 갖는다.

이미지 상에서 위치 정보 ν의 좌표 값 x,y를 구하는 함수 $getx(s)$, $gety(s)$와 스팟 s의 농도 정보인 $\%OD, \%Vol$을 구하는 함수 $getod(s), getvol(s)$는

- $getx(s)$: 스팟 s의 속성 x
- $gety(s)$: 스팟 s의 속성 y
- $getod(s)$: 스팟 s의 속성 $\%OD$
- $getvol(s)$: 스팟 s의 속성 $\%Vol$

이고 스팟 s의 속성 중 임의의 한 f의 값을 구하는 함수는 $getf(s)$라고 정의한다. □

[정의 2] 기준점 스팟 집합

서로 다른 r개의 기준점 P = { $p_1, ..., p_r$ }와 이미지 집합 M = { $m_1, ..., m_n$ }이 주어졌을 때, 이미지 m_i의 기준점 p^k에 해당하는 스팟을 s_i^k라고 하면, 이미지 집합 M에서 기준점 p^k에 해당하는 스팟들의 집합 $P^k = \{x \mid x = s_i^k, 1 \leq i \leq n\}$이다. □

한 이미지는 r개의 기준점을 모두 포함하여야 하며 결국 기준점 p^k는 각 이미지에서 상대적 위치가 동일한 스팟들이다. 이미지 m_i는 r개의 기준점을 갖으며 r개의 기준점에 대한 특징을 프로파일로 요약하기 위해 각 이미지에 나타난 특정 기준점 p^k에 해당되는 스팟 s_i^k에 대한 속성 $property(s_i^k)$의 평균과 표준편차, 범위를 다음과 같이 계산한다.

$property(s_i^k)$ 중 $f \in (\nu, \%OD, \%Vol)$일 때,

기준점 p^k의 평균 : $\mu(f,p^k) = \frac{1}{n}\sum_{i=1}^{n} getf(s_i^k)$ (식 1)

기준점 p^k의 표준 편차 :

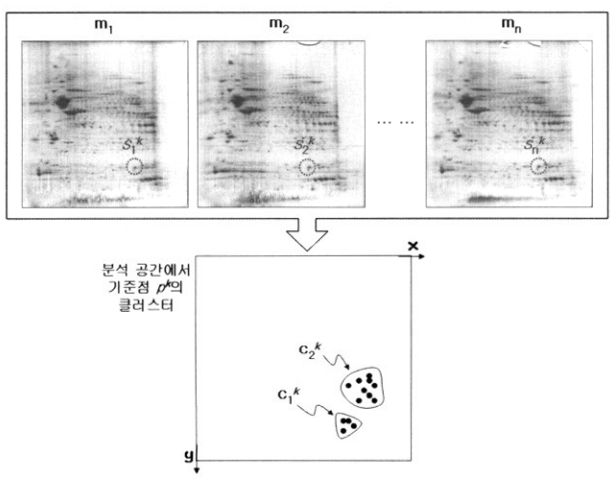
$$\sigma(f,p^k) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\{getf(s_i^k) - \mu(f,p^k)\}^2} \quad \text{(식 2)}$$

3.2 기준점 클러스터링

단백질 이미지는 실험기기 및 실험실 조건에 따라 평행이동, 회전, 확대, 축소의 이미지 변형이 일어날 수 있으며

이로 인해 이미지 내의 스팟 변화가 일어날 수 있다. 그러나 이 이미지 변형은 실험실과 실험자에 의해 어느 정도 정형화되어 있다. 이러한 2DE 이미지 실험 시 발생하는 이미지 변형을 모델링하기 위해 학습용 데이터의 클러스터링을 수행한다. 본 논문에서는 DBSCAN[11, 12]방법을 적용하였다. DBSCAN 알고리즘은 두 가지의 변수 Eps와 MinPts에 기반을 두어 클러스터를 찾으며 Eps는 클러스터를 형성하기 위한 데이터간의 사정거리이며 MinPts는 클러스터를 형성하기 위한 최소 데이터의 개수이다.

학습용 이미지 집합 $M_L=\{m_1, m_2, \ldots, m_n\}$이 주어졌을 때, (그림 1)은 n개의 이미지 각각에서 기준점 p^k의 스팟들을 좌표 값에 의해 클러스터링 하여 두개의 클러스터를 형성한 예이다. 각 이미지에서 기준점 p^k인 스팟 s_i^k가 n개 존재하며 각 스팟은 속성 v를 갖는다. 클러스터링 알고리즘을 통하여 n개의 스팟을 군집화 하였을 때, 유사한 좌표 값을 갖는 스팟이 서로 모여 하나의 클러스터를 형성한다. n개의 이미지에서 기준점 p^k는 클러스터링에서 좌표 값을 차원으로 할 때 각 클러스터는 비슷한 좌표 값을 갖기 때문에 전반적인 유사한 이미지 변형이 이루어졌다는 특징을 갖는다. 반면 기준점은 비교적 강한 농도 값을 갖는다는 특징이 있기 때문에 농도속성에 의한 클러스터링은 적절치 않다. 따라서 클러스터링은 위치 정보에 의해서 수행되며, 농도 정보는 생성된 클러스터 내의 스팟의 농도 정보를 대상으로 통계치를 계산한다.



(그림 1) 기준점 p^k에 대한 클러스터의 형성 예

[정의 3] 클러스터의 스팟 집합과 요소

e^k개의 클러스터를 갖는 기준점 p^k에서 클러스터 u를 c_u^k라 정의하고 c_u^k의 스팟 집합을 C_u^k라 할 때, $C_u^k = \{ s | s = s_i^k, s_i^k \in P^k$ and s_i^k는 학습용 이미지 $m_i (\in M_L)$의 스팟이며, 동시에 클러스터 c_u^k에 존재하는 스팟$\}$이고 클러스터 c_u^k의 속성 $property(c_u^k)$=(범위, 평균, 표준편차, 지지도)이다. □

클러스터 c_u^k의 요소인 범위는 클러스터 내의 데이터의 최소값과 최대값에 의해 결정되며 클러스터의 범위에 따라 기준점 후보 데이터를 결정하게 된다.

$$min(f, c_u^k) = min(f | f = property(s_i^k \in C_u^k)) \quad \text{(식 3)}$$

$$max(f, c_u^k) = max(f | f = property(s_i^k \in C_u^k)) \quad \text{(식 4)}$$

클러스터 c_u^k의 요소 중 평균값은 클러스터 내에 존재하는 모든 기준점 s_i^k의 평균 f값이다. 클러스터의 평균은 기준점의 평균과는 의미가 다르다. 기준점의 평균은 전체 기준점의 산술적 계산에 의해 나온 값이기 때문에 중심에는 데이터가 존재하지 않을 수 있다. 그러나 클러스터의 평균은 클러스터 자체가 유사한 속성을 갖고 모여 있는 데이터의 집합이므로 클러스터의 평균값이 실제 데이터의 대표값이다.

클러스터의 평균 :

$$\mu(f, c_u^k) = \frac{1}{|C_u^k|} \sum getf(s_i^k), \quad \forall s_i^k \in C_u^k \quad \text{(식 5)}$$

클러스터 c_u^k의 요소인 표준편차는 클러스터의 밀집된 정도를 의미한다. 클러스터의 표준편차가 작을수록 데이터의 유사도가 크고 강한 속성을 갖는 클러스터임을 의미한다.

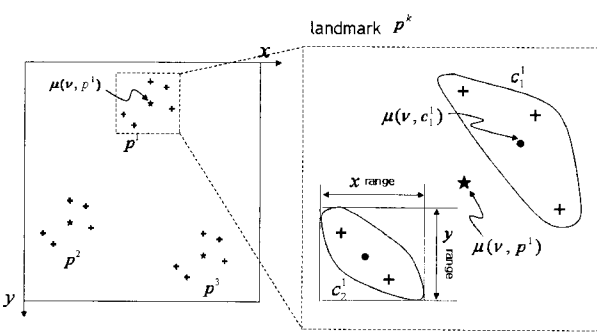
클러스터의 표준편차 :

$$\sigma(f, c_u^k) = \sqrt{\frac{1}{|C_u^k|} \sum \{getf(s_i^k) - \mu(f, c_u^k)\}^2}, \quad \forall s_i^k \in C_u^k \quad \text{(식 6)}$$

클러스터 c_u^k의 요소 중 지지도는 전체 학습 이미지 n개 중 기준점 p^k가 해당 클러스터에 속하는 이미지의 비율을 의미하여 (식 7)과 같이 계산된다.

$$\text{클러스터의 지지도} : Support(c_u^k) = \frac{|C_u^k|}{n} \quad \text{(식 7)}$$

(그림 2)는 n개의 이미지를 통합한 분석 공간에서의 기준점과 기준점의 클러스터의 개념을 예를 들어 설명하고 있다.



(그림 2) 분석 공간에서의 기준점과 기준점 클러스터의 요소들

〈표 1〉 클러스터 프로파일 요소

속성	I.D	범위	평균	표준편차	지지도
항목	p^k, c_u^k	$\min(x, c_u^k)$, $\max(x, c_u^k)$ $\min(y, c_u^k)$, $\max(y, c_u^k)$ $\min(od, c_u^k)$, $\max(od, c_u^k)$ $\min(vol, c_u^k)$, $\max(vol, c_u^k)$	$\mu(x, c_u^k)$, $\mu(y, c_u^k)$, $\mu(od, c_u^k)$, $\mu(vol, c_u^k)$	$\sigma(x, c_u^k)$, $\sigma(y, c_u^k)$, $\sigma(od, c_u^k)$, $\sigma(vol, c_u^k)$	$support(c_u^k)$

생성된 클러스터는 클러스터 요소인 클러스터의 범위, 평균값과 표준편차, 지지도를 통해 각 클러스터의 특징을 추출한다. 추출된 특징은 <표 1>의 형식으로 프로파일에 기록된다. 그리고 기록된 프로파일은 기준점 추출 과정에서 후보 데이터 집합을 선정하는 기준으로 사용된다. 클러스터에 포함되는 스팟의 최소값과 최대값은 클러스터의 범위이다.

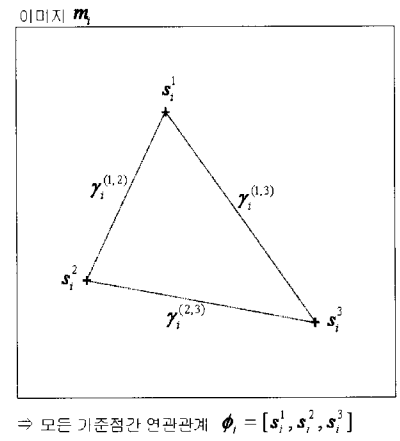
[정의 4] 한 이미지 내의 기준점 스팟들의 연관관계

학습 이미지 $m_i \in M_L$에 존재하는 기준점 중 두 점간에는 거리와 각도를 갖는 연관관계가 존재한다. 이 경우 이미지 m_i에서 두 기준점 p^k, p^l에 대한 스팟 s_i^k, s_i^l간의 연관관계는 $\gamma_i^{(k,l)} = (s_i^k, s_i^l)$이고 $\gamma_i^{(k,l)}$의 속성 $property(\gamma_i^{(k,l)})$=(두 점 사이의 거리 δ, 두 점 사이의 각도 θ)이다. 이 때, 이미지 m_i의 모든 기준점 r개 전체를 포함하는 연관관계 $\phi_i = (s_i^1, ..., s_i^r)$이고 ϕ_i의 속성 $property(\phi_i)$ = (거리합, 각도합)이다. □

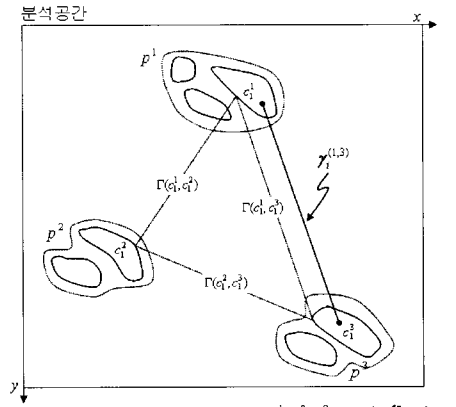
[정의 5] 분석 공간에서의 기준점 클러스터간의 연관관계

n개의 이미지를 통합한 분석 공간에서 두 클러스터 c_u^k, c_v^l $(k \neq l)$의 연관관계 $\Gamma(c_u^k, c_v^l) = \{x | x = \gamma_i^{(k,l)} = (s_i^k, s_i^l), s_i^k \in C_u^k \text{ and } s_i^l \in C_v^l, 1 \leq i \leq n\}$이며 $\Gamma(c_u^k, c_v^l)$의 속성 $property(\Gamma(c_u^k, c_v^l))$ = (거리의 평균, 각도의 평균, 거리의 표준편차, 각도의 표준편차)이다. 이 때, 이미지에 존재하는 r개의 기준점 각각에 대한 모든 클러스터의 연관관계 $\Phi(c_u^1, .., c_v^r) = \{x | x = \phi_i = (s_i^1, ..., s_i^r), s_i^1 \in C_u^1, .., s_i^r \in C_v^r\}$이고 $\Phi(c_u^1, ..., c_v^r)$의 속성 $property(\Phi(c_u^1, ..., c_v^r))$ = (거리합의 평균, 각도합의 평균, 거리합의 표준편차, 각도합의 표준편차)로 정의된다. □

(그림 3) (a)는 이미지 내의 각 기준점에 대해 두 점의 관계 $\gamma_i^{(k,l)}$과 모든 기준점의 관계 ϕ_i를 나타내고 있다. 두 점 간에는 거리와 각도가 존재하며 이를 확장하여 각 기준점의 클러스터를 형성하는 분석공간에서 두 클러스터와의 관계는 (그림 3) (b)에서 보는 바와 같이 $\Gamma(c_u^k, c_v^l)$로 정의된다. 앞서 언급한 바와 같이 단백질 2DE 이미지는 절대적인 좌표값이 중요한 의미를 갖고 있지 않다. 한 스팟은 그 스



(a) 이미지 m_i에서의 스팟간 연관관계



(b)분석 공간에서의 클러스터간 연관관계

(그림 3) 스팟들과 클러스터들의 상호 연관관계

팟의 절대적 좌표 ν보다는 주변의 다른 스팟과의 상대적인 거리와 방향으로 표현할 때 더욱 의미가 있다. 따라서 기준점을 추출하는 과정에 있어서 스팟과 클러스터는 절대적 위치보다는 상대적인 위치, 즉 두 스팟 또는 클러스터 상호간의 거리와 방향으로 표현되어야 한다. 두 스팟 $s_i^k(x_1, y_1)$과 $s_i^l(x_2, y_2)$의 거리 δ와 각도 θ는 (식 8)과 (식 9)와 같이 계산된다.

$i = 1, ...n, \quad k \neq l$ 에 대해서

$$\delta(i, c_u^k, c_v^l) = \begin{cases} \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} & (if, s_i^k \in C_u^k \text{ and } s_i^l \in C_v^l) \\ 0 & (otherwise) \end{cases} \quad \text{(식 8)}$$

$$\theta(i, c_u^k, c_v^l) = \begin{cases} tan^{-1}(\frac{y_2 - y_1}{x_2 - x_1}) & (if, s_i^k \in C_u^k \text{ and } s_i^l \in C_v^l) \\ 0 & (otherwise) \end{cases} \quad \text{(식 9)}$$

$\delta(i, c_u^k, c_v^l)$는 이미지 m_i에의 두 기준점 p^k (s_i^k)과 p^l (s_i^l)가 각각 두 클러스터 c_u^k, c_v^l에 포함될 경우 두 스팟간의 거리를 계산하여 표현하고 그렇지 않으면 거리를 고려하지 않는다. $\theta(i, c_u^k, c_v^l)$ 역시 마찬가지로 동일 이미지에 존재하는 두 스팟이 클러스터 c_u^k, c_v^l에 포함될 경우 두 스팟간의 각도를 역탄젠트로 계산하여 두 스팟을 연결한 벡터의 방향을 표현한다. 두 스팟의 거리와 각도가 계산되면 두 클러스터 c_u^k, c_v^l의 관계 $\Gamma(c_u^k, c_v^l)$의 요소들인 거리와 각도에 대한 범위, 평균, 표준편차 그리고 $\Gamma(c_u^k, c_v^l)$의 신뢰도를 계산한다. 평균 거리와 평균 각도는 한 이미지에 두 기준점 p^k, p^l이 각각 두 클러스터 c_u^k, c_v^l에 포함되는 모든 $\gamma_i^{(k,l)}$들의 거리와 각도에 대한 평균을 계산한다. 거리 $\delta(i, c_u^k, c_v^l)$와 각도 $\theta(i, c_u^k, c_v^l)$를 이용하여 전체의 평균을 구할 수 있다. 두 클러스터 c_u^k, c_v^l의 평균 거리를 계산하는 방법은 다음과 같다.

$$\mu_\delta(c_u^k, c_v^l) = \frac{1}{|\Gamma(c_u^k, c_v^l)|} \sum_{i=1}^{n} \delta(i, c_u^k, c_v^l) \quad \text{(식 10)}$$

두 클러스터의 평균 각도 역시 평균 거리와 동일한 방법으로 계산한다.

$$\mu_\theta(c_u^k, c_v^l) = \frac{1}{|\Gamma(c_u^k, c_v^l)|} \sum_{i=1}^{n} \theta(i, c_u^k, c_v^l) \quad \text{(식 11)}$$

학습 데이터는 검증된 기준점 데이터이다. 이 기준점은 기준점 p_k의 한 클러스터에 반드시 속하기 때문에 기준점 데이터간의 관계를 통하여 두 클러스터간의 $\Gamma(c_u^k, c_v^l)$을 거리와 각도의 평균, 표준편차, 범위 그리고 신뢰도를 통하여 설명하였다. 추출된 $\Gamma(c_u^k, c_v^l)$의 요소들은 <표 2>의 형식으로 프로파일에 기록한다.

이미지 m_i에서 r개의 모든 기준점을 통합한 벡터를 ϕ_i라 정의하였다. 모든 이미지는 r개의 기준점을 모두 갖고 있으며 이들은 다른 ϕ_j $(i \neq j)$들과의 거리와 각도의 차이에 따른 거리와 각도에 관한 에러값과 통합 에러값을 갖는다. 그리고 모든 이미지를 통합하여 분석하는 분석 공간에서는 각 이미지의 ϕ_i에 알맞은 클러스터들로 구성된 $\Phi(c_u^k, ..., c_v^l)$에 포함되게 되며 각 $\Phi(c_u^k, ..., c_v^l)$는 ϕ_i의 요소인 에러값에 대한 범위, 평균, 표준편차, 지지도를 요소로 갖고 있다.

〈표 2〉 두 기준점 클러스터간 $\Gamma(c_u^k, c_v^l)$의 프로파일 요소

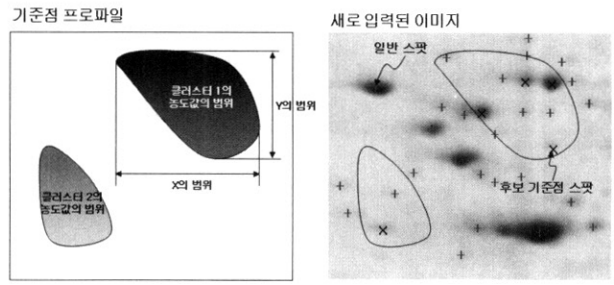
속성	I.D	평균	표준편자
항목	$p^k, c_u^k,$ p^l, c_v^l	$\mu_\delta(c_u^k, c_v^l)$, $\mu_\theta(c_u^k, c_v^l)$	$\sigma_\delta(c_u^k, c_v^l)$, $\sigma_\theta(c_u^k, c_v^l)$

4. 기준점 추출 알고리즘

기준점 추출 과정에서는 생성된 프로파일을 기반으로 신생 이미지 데이터에서 기준점이 될 수 있는 후보 데이터를 먼저 선택하고 이들과의 상대적 위치를 비교하여 프로파일과 각 기준점에 대해 가장 오차가 적은 기준점 후보 스팟들을 추출하는 과정이다. 기준점 추출 방법에는 이미지에 존재하는 모든 스팟들을 후보로 탐색하는 방법과 A^* 알고리즘에 기반 하여 탐색을 최소화하여 최적의 후보 스팟들을 찾는 방법이 있다.

4.1 기준점 후보 집합 선정

새로 입력된 한 대상 이미지 m_t에 대해서 후보 집합은 학습 과정을 통해 생성된 클러스터의 절대 위치 (x, y)와 농도 값$(\%OD, \%Vol)$의 범위에 의해 선정된다. $x, y, \%OD, \%Vol$으로 이뤄진 4차원 공간에서 클러스터의 속성 값 범위 내의 스팟들을 후보 집합으로 선정한다.

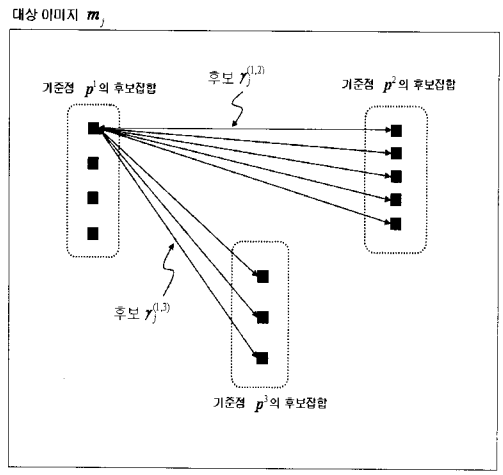


(그림 4) 기준점 후보 스팟의 선정

(그림 4)는 기준점 프로파일이 주어졌을 때, 새로 입력된 이미지에서 기준점 후보 스팟이 어떻게 선정되는지를 보여준다. 3장에서 설명한 기준점 클러스터 프로파일의 X,Y, %OD,%Vol의 범위가 그림과 같다고 할 때, 새로 입력된 이미지 내의 스팟중에서 클러스터의 영역 내에 존재하면서 클러스터의 농도 값 범위 내에 포함되는 스팟들이 기준점 후보 스팟으로 선정되며, 그림에서는 'X'로 표시되었다.

4.2 기준점 후보 스팟 집합의 에러값 계산

이미지 m_j에서 후보 집합으로 선정된 스팟들은 두 기준점 s_j^k, s_j^l의 후보 스팟들간의 $\gamma_j^{(k,l)}$을 만들며 각 후보 $\gamma_j^{(k,l)}$는 두 후보 스팟이 속한 클러스터로 구성된 $\Gamma(c_u^k, c_v^l)$의 프로파일 요소를 통해 에러값을 계산한다. (그림 5)는 기준점별 후보 스팟들이 $\gamma_j^{(k,l)}$을 생성하는 과정이다.



(그림 5) 후보 $\gamma_j^{(k,l)}$ 생성 방법

그림과 같이 기준점 p^1에 총 4개의 후보 데이터가 있다고 가정하면 그 중 첫 번째 데이터에 대해 p^1이 아닌 다른 기준점의 후보 집합들과의 관계를 생성할 수 있다. 만일 각 기준점 p^k의 후보 데이터 집합이 $can(p_k)$이라 하면, 선정된 후보 집합에서 생성 가능한 후보 벡터의 개수는 $\sum_{i=1}^{r-1}\sum_{j=i+1}^{r}|can(p^i)|\cdot|can(p^j)|$이다. 한 후보 $\gamma_j^{(k,l)}$의 오차값은 크게 거리에 대한 오차 ϵ_δ과 각도에 대한 오차 ϵ_θ로 나눌 수 있다. 오차값은 과거에 학습된 데이터에 비해 새로운 이미지상의 후보 기준점들로 생성된 후보 $\gamma_j^{(k,l)}$이 얼마나 큰 차이를 갖는지를 수치화시킨 값이다. 두 기준점 p^k, p^l에 해당하는 후보 스팟간의 관계 $\gamma_j^{(k,l)}$의 거리와 각도에 관한 오차값 $\epsilon_{\delta_{kl}}, \epsilon_{\theta_{kl}}$는 다음과 같이 계산된다.

$$\epsilon_{\delta_{kl}} = \frac{\delta(j, c_u^k, c_v^l) - \mu_\delta(c_u^k, c_v^l)}{\sigma_\delta(c_u^k, c_u^l)} \quad \text{(식 12)}$$

$$\epsilon_{\theta_{kl}} = \frac{\theta(j, c_u^k, c_v^l) - \mu_\theta(c_u^k, c_v^l)}{\sigma_\theta(c_u^k, c_v^l)} \quad \text{(식 13)}$$

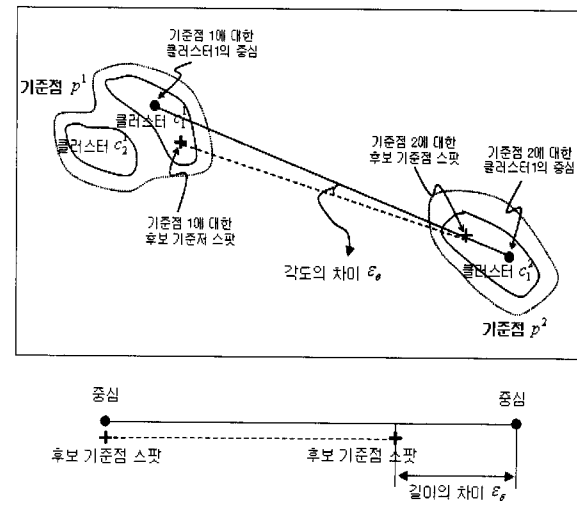
μ_δ와 σ_δ는 $\Gamma(c_u^k, c_v^l)$내에 포함된 각 스팟들 사이의 벡터들의 거리의 평균과 표준편차를 나타낸다. 후보 $\gamma_j^{(k,l)}$는 후보 집합의 모든 조합으로 생성된 것이다. 후보 $\gamma_j^{(k,l)}$가 생성되면 각 기준점의 후보 데이터를 통하여 후보 ϕ_j를 생성한다. 기준점 p_k의 후보 데이터의 개수가 $can(p^k)$이므로 생성할 수 있는 후보 ϕ_j의 개수는 총$\prod_{k=1}^{r} can(p^k)$개이다. 이 중 가장 최적의 후보 ϕ_j을 찾기 위해서 ϕ_j를 구성하는 기준점 간의 $\gamma_j^{(k,l)}$의 거리와 각도에 관한 오차값을 계산하여야 한다.

후보 ϕ_j를 구성하는 각 $\gamma_j^{(k,l)}$의 오차값의 총 합계로 ϕ_j의 오차값을 계산할 수 있다. ϕ_j의 오차값 ϵ을 계산하는 식은 다음과 같다.

$$\epsilon_{kl} = \epsilon_{\delta_{kl}} + \epsilon_{\theta_{kl}} \quad \text{(식 14)}$$

$$\epsilon = \sum_{k=1}^{r-1}\sum_{l=k+1}^{r} \epsilon_{kl} \quad \text{(식 15)}$$

여기서 ϵ_{kl}은 두 기준점 p^k, p^l의 후보 스팟간의 관계 $\gamma_j^{(k,l)}$의 오차값을 의미하며 거리의 오차값과 각도의 오차값에 동등한 가중치를 부여한다. 동등한 가중치를 부여한다는 것은 각도의 변화 거리의 변화의 중요도에 차이를 두지 않는다는 의미이다. (그림 6)은 학습 과정을 통해 생성된 프로파일과 새로운 이미지 m_j를 비교하는 과정을 설명한다. 기존 학습된 프로파일에서의 두 클러스터간의 평균 거리와 새로운 이미지의 후보 스팟들간의 거리의 표준편차는 $\epsilon_{\delta_{kl}}$이고 각도의 차이는 $\epsilon_{\theta_{kl}}$이다. 오차값 계산을 통해 새로운 이미지에서 가능한 후보 ϕ_j의 오차값을 계산하고 최소 오차값을 갖는 ϕ_j를 기준점 집합으로 정의한다.



(그림 6) 거리와 각도에 대한 에러값의 의미

4.3 A^*알고리즘 기반 기준점 추출 방법

4.2절에서 설명한 기준점 추출 방법은 $\prod_{k=1}^{r}|can(p^k)|$개의 조합이 생성되며 이미지당 스팟의 수가 많아지고 후보 스팟 집합의 범위가 커지고 기준점의 개수가 많아질수록 $can(p^k)$의 크기가 커지므로 조합의 수는 기하급수적으로 늘어나며 최적의 후보 ϕ_j을 찾기 위한 검색 횟수도 상당히 많아진다. 따라서 효율적으로 ϕ_j을 탐색할 수 있는 탐색 알고리즘이 필요하다. 본 논문에서는 A^* 알고리즘[13, 14]의 트리 확장 기법을 이용하여 검색 시간을 감소시키는 알고리즘을 제안한다. A^* 탐색 알고리즘에서 각각의 노드는 ϕ_j을 구성하는 스팟의 집합과 검색한 기준점의 개수, 그리고 ϕ_j의 오차값을 함께 저장한다. A^* 탐색 트리의 레벨별 확장 순서를

결정하기 위해서 각 기준점의 상대 정형도 τ^k 값을 다음과 같이 계산한다.

$$\tau^k = \frac{|P^k|}{\frac{1}{|P^k|}\Sigma R}, \quad R = \sqrt{(getx(s)-\mu(v.x,p^k))^2+(gety(s)-\mu(v.y,p^k))^2}, \quad \forall s \in P^k \qquad \text{(식 16)}$$

상대 정형도란 학습용 기준점 데이터가 유사한 위치에 많이 발현되면 앞으로 찾아야 할 기준점 역시 유사한 위치에 발현될 것을 기대하여 트리 확장에서 먼저 우선순위를 주는 것이다. R은 기준점 p^k에 속하는 스팟들과 p^k의 중심과의 거리이다. 또한 각 레벨에서는 각 후보 스팟이 속한 클러스터의 지지도에 따라 지지도가 높은 클러스터에서 선정된 후보 스팟을 우선적으로 탐색한다.

각 이미지는 프로파일로 모델링된 모든 기준점을 갖고 있음을 보장하지 못한다. 따라서 임의의 오픈 노드를 n에 대하여 확장할 때, 만일 비용이 지나치게 증가할 경우 해당 기준점을 ϕ_j에서 제외시킨다. 이 값을 사용자가 허용하는 최대 비용 증가값 *MaxDelta*라 정의한다. 비용은 노드의 오차값의 합으로 계산하였다. 만일 오픈 노드 중 최소의 오차값을 갖는 노드가 전체 기준점을 모두 검색하였다면 이 노드의 구성 스팟들이 최적의 기준점 스팟으로 결정된다.

5. 실험 및 결과 분석

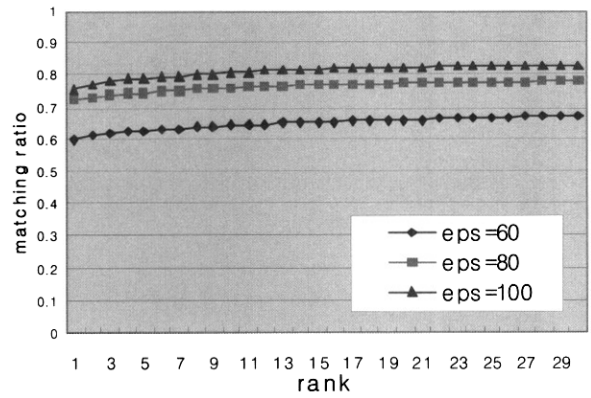
본 장에서는 이미 검증된 기준점을 포함하는 이미지 데이터를 대상으로 기준점 추출 알고리즘의 정확도와 시간을 측정하여 학습 데이터의 양과 클러스터의 범위에 따른 결과의 차이를 실험을 통하여 분석하였다. 실험에 사용한 데이터는 인간의 간 조직 단백질 이미지 데이터 88개이며 <표 3>과 같은 특성을 갖고 모든 이미지는 이미 생물학자의 실험적 검증을 통해 기준점을 알고 있는 데이터이다.

〈표 3〉 데이터 특성

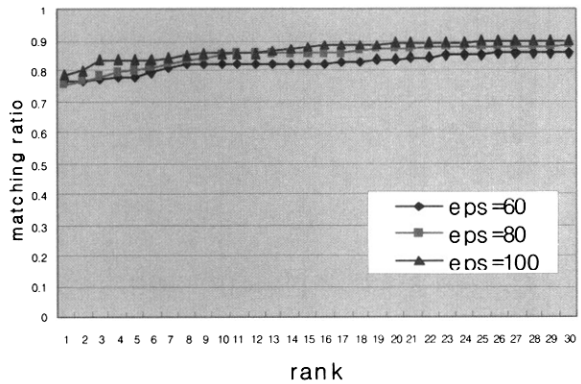
총 이미지 개수	88개 (사람의 간 조직)
이미지별 평균 스팟수	898개
이미지당 최대 기준점의 개수	14개
이미지당 평균 기준점의 개수	10개

본 실험은 88개의 이미지 데이터 중 일부 이미지를 학습을 통한 프로파일 생성에 사용할 것이며, 학습 데이터를 포함한 전체 88개의 이미지를 모두 테스트 데이터로 사용할 때 정확도를 측정해보았다. 학습의 양에 따라 기준점 클러스터 영역은 차이가 있고 학습을 많이 할수록 대부분의 데이터를 프로파일이 포함할 수 있기 때문에 보다 정확도가 높아질 수 있다.

(그림 7)은 학습 데이터를 제외한 비학습 데이터의 정확도(matching ratio)를 보여준다. 여기서 x축이 의미하는 순위(rank)는 A^* 탐색 알고리즘에 의해 추천되는 기준점 집합의 순위를 의미한다. 정확도는 총 M개의 기준점 정답 중 몇 개를 순위 내에서 정확히 추출했는가를 계산하였다. 정확도를 살펴보면 그림 (a)의 50% 학습에서는 최대 82%의 정확도를 보이는데 비해 (그림 7) (b)의 75% 학습에서는 최대 90%의 높은 정확도를 보인다. 따라서 학습을 많이 하면 할수록 기준점 추출에 있어서의 정확도는 향상됨을 유추할 수 있다. 다음은 학습 데이터 집합의 크기와 클러스터 범위에 따른 수행 시간을 비교한 실험이다. 학습 데이터 집합의 크기가 커지고 클러스터의 범위가 커질수록 후보 집합의 데이터 크기 역시 커진다. 후보 집합의 크기가 커지면 수많은 후보 ϕ_j가 생성되며 최적의 ϕ_j를 검색하기에는 많은 시간이 소요된다. 그러나 A^*탐색 트리를 사용하여 전체 탐색을 하지 않음으로써 기준점 추출의 수행시간을 단축하였다. 시간은 초 단위로 계산하였으며, 매 시도에서 학습을 하고 88개의 전체 이미지의 기준점을 모두 추출하는데 걸린 전체 시간을 측정하였다. (그림 8)은 50% 학습한 실험과 75% 학습한 실험의 수행 시간을 보여준다.

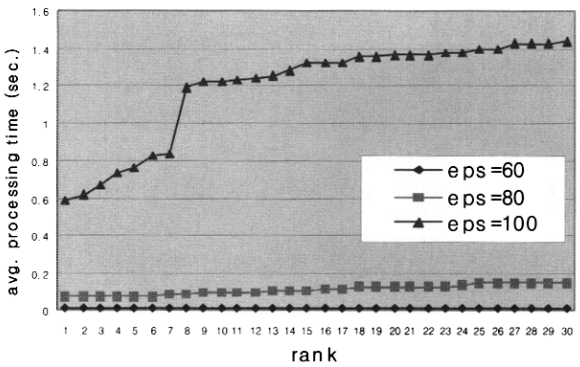


(a) 50% 학습의 정확도

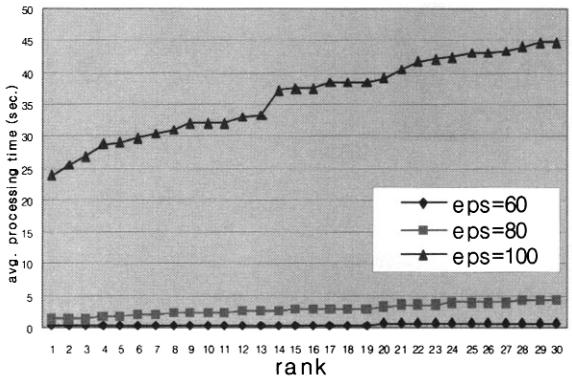


(b) 75% 학습의 정확도

(그림 7) 클러스터의 범위와 학습량에 따른 정확도



(a) 50% 학습의 수행시간



(b) 75% 학습의 수행시간

(그림 8) 클러스터의 범위와 학습량에 따른 수행 시간

(그림 8)의 x축은 (그림 7)의 순위와 동일하며, y축은 하나의 이미지에서 평균적으로 소요되는 기준점 추출 시간이다. 이 수행 시간은 학습 시간을 제외하고 기준점을 추출하는데 걸리는 시간만을 측정하였다. 평균 수행 시간(average processing time) 정확도 실험과 마찬가지로 5번의 반복 시행을 수행하였다. (그림 8) (a)는 50% 학습을 수행했을 경우의 수행시간으로 각 이미지 내의 기준점은 1초 내외에 추출되었다. 또한 30번째 순위의 기준점 후보 집합을 제시하는데 걸리는 시간은 Eps=100에서 1.4초 만에 수행되었다. 반면, (그림 8) (b)는 88개의 이미지 중 66를 랜덤으로 선택하여 학습한 75%의 학습시의 수행시간을 보여준다. 이 경우, Eps가 80 이하인 경우는 수행시간이 5초 이내로 비교적 빠른 수행시간을 보였으나, Eps를 100으로 설정한 경우, 수행 속도는 최대 45초까지 증가하였다. (그림 7)에서 75%의 정확도는 각 Eps별로 큰 차이가 없으나, 수행시간은 급격히 상승하였다. 따라서 Eps의 값을 지나치게 크게 잡을 경우 전체 알고리즘의 성능을 저하시키는 요인이 된다.

6. 결론 및 향후 연구

본 논문에서는 단백질 2DE 이미지의 기준점을 추출하는 알고리즘을 통해 기존의 이미지 분석 프로그램에서의 다량의 이미지 분석의 어려움이었던 기준점 입력 과정을 자동화하여 클러스터링을 통한 학습을 하고 프로파일을 생성하여 신생 이미지에서의 기준점을 추출하는 방법을 제안하였다. 기준점 데이터의 클러스터링을 통해 프로파일을 생성하고, 생성된 프로파일을 기반으로 신생 이미지에서의 기준점 후보 스팟 집합을 생성하고 후보 집합의 오차값을 계산을 통해 최소 오차값을 갖는 후보 스팟 집합을 기준점으로 판단한다. 본 알고리즘의 성능을 비교하기 위해서 88개의 사람의 간 조직 이미지에서 알려진 기준점들을 대상으로 기준점 추출 알고리즘을 수행하여 보았다. 그러나 학습 과정은 반복적으로 수행되어야 하고 새로 찾아진 기준점은 프로파일에 재반영되어야 한다. 따라서, IncrementalDBSCAN[15]와 같은 기준점의 점진적 클러스터링을 수행해야 하며 초기 원천 데이터가 없더라고 프로파일 정보만을 통해 프로파일을 갱신할 수 있는 점진적 기준점 추출 알고리즘의 연구가 수행되어야 한다.

참 고 문 헌

[1] GÖrg, A., Obermaier, C., Boguth, G., Harder, A., Scheibe, B., Wildgruber, R. and Weiss, W., The Current State of Two-Dimensional Electrophoresis with Immobilized pH Gradients, Electrophoresis, Vol.21, No.6, pp.1037-53, 2000.

[2] Corbett, J., Dunn, MJ., Posch, A. and GÖrg, A., Positional Reproducibility of Protein Spots in Two-Dimensional Polyacrylamide Gel Electrophoresis Using Immobilized pH Gradient Iso-Electric Focusing in The First Dimension : An Interlaboratory Comparison, Electrophoresis, Vol.15, No.8-9, pp.1205-11, 1994.

[3] Gygi, SP, Rist, B., Gerber, SA, Turecek, F., Gelb, MH and Aebersold, R., Quantitative Analysis of Complex Protein Mixtures Using Isotope-Coded Affinity Tags, Nat.Biotech, Vol.17, No.10, pp.994-9, 1999.

[4] Cagney, G. and Emili, A., De Novo Peptide Sequencing and Quantitative Profiling of Complex Protein Mixtures Using Mass-Coded Abundance Tagging, Nat Biotech., Vol.20, No. 2, 163-70, 2002.

[5] http://www.genebio.com/products/2d_image.html

[6] http://www.nonlinear.com/products/2d/progenesis/overview.asp

[7] Takahashi, K., Nakazawa, M., Watanabe, Y. and Konagaya, A., Automated Processing of 2-D Gel Electrophoretograms of Genomic DNA for Hunting Pathogenic DNA Molecular Changes, Genome Informatics 1999, 121-132, Tokyo, Japan, December, 1999.

[8] Toussaint, G.T., The Relative Neighborhood Graph of Finite Planar Set, Pattern Recognition, Vol.12, No.4, pp.261-8, 1980.

[9] Hayashizaki, Y., Hirotsune, S., Okazaki, Y., Hatada, I., Shibata, H., Kawai, J., Hirose, K.,Watanabe, S., Fushiki, S. and Wada, S., Restriction Landmark Genomic Scanning

Method and Its Various Applications, Electrophoresis, Vol. 14, No.4, 251-8, 1993.

[10] Hoffmann, F., Kriegel, K. and Wenk, C., Matching 2D Patterns of Protein Spots, Proceedings of the 14th Annual Symposium on Computational Geometry, 231-9, Minnesota, U.S.A., June, 1998.

[11] Ester, M., Kriegel, H.-P., Sander, J. and Xu, X., A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise, Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, 226-31, Portland, U.S.A., August, 1996.

[12] Jiawei Han and Micheline Kamber, DataMining: Concepts and Techniques, 2000.

[13] Nils J. Nilsson and Morgan, K., Artificial Intelligence: A New Synthesis, 1998.

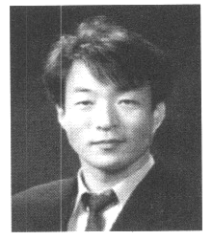
[14] Labio, W.J., Quass, D. and Adelberg, B., Physical Database Design For Data Warehouses, Proceedings of the 13th International Conference on Data Engineering, 277-88, Birmingham, U.K., April, 1997.

[15] Ester, M., Kriegel, H.-P., Sander, J., Wimmer, M. and Xu X., Incremental Clustering for Mining in a Data Warehousing Environment, Proceedings of the 24th International Conference on Very Large Data Bases, 323-33, New York City, U.S.A., August, 1998.

심 정 은

e-mail : jjuggeuni@database.yonsei.ac.kr
2001년 인천대학교 전자계산학과(공학사)
2003년 연세대학교 컴퓨터과학과 (공학석사)
2003년~현재 연세대학교 컴퓨터과학과 박사과정
관심분야 : 생물정보학, 데이터웨어하우스, 데이터마이닝

이 원 석



e-mail : leewo@database.yonsei.ac.kr
1985년 미국 보스턴대학교 컴퓨터공학과(공학사)
1987년 미국 퍼듀대학교 컴퓨터공학과 (공학석사)
1990년 미국 퍼듀대학교 컴퓨터공학과 (공학박사)
1990년~1992년 삼성전자 선임연구원
1993년~1999년 연세대학교 컴퓨터과학과 조교수
1999년~2004년 연세대학교 컴퓨터과학과 부교수
2004년~현재 연세대학교 컴퓨터과학과 교수
관심분야 : 분산데이터베이스, 미디에이터시스템, 데이터마이닝, 데이터스트림