

# 단백질 2-DE 이미지 분석에서 정확한 스팟 매칭 패턴 검색을 위한 효과적인 방법

(An Efficient Method to Find Accurate Spot-matching Patterns in Protein 2-DE Image Analysis)

김 연 화 <sup>†</sup>      이 원 석 <sup>\*\*</sup>  
(Yanhua Jin)      (Won Suk Lee)

**요 약** 단백질 2-DE 이미지 분석에서 단백질 자체가 가지고 있는 불안정성과 2-DE 실험이 가지고 있는 근본적인 문제점으로 인하여 이미지 스팟 매칭 분석의 정확도가 낮아지게 된다. 이 논문에서는 다중 참조이미지를 사용하여, 스팟 매칭 패턴의 정확도에 큰 영향을 주는 이미지 찌그러짐을 보완하고, 그에 따른 노이즈 스팟 제거와 참조 이미지 품질에 의한 정확도 저하를 최소화하는 방법을 제안하였다. 또한 2-DE 이미지의 데이터 특성에 의하여 이미지 수가 증가할 때 성능이 급격히 떨어지는 문제를 해결하기 위하여, 다중 참조이미지를 사용하여 구축한 스팟 매칭 데이터 베이스를 이미지의 생물학적 특성에 의하여 “분할 및 확장” 방법을 사용하여, 정확도를 향상시키는 동시에 패턴 길이를 보장하는 스팟 매칭 패턴을 효과적으로 생성하였다. 실험에서는 실제 인간 2-DE 이미지 데이터를 사용하여 제안한 방법의 타당성을 보여준다.

**키워드** : 2-DE 이미지, 스팟 매칭, 이미지 분석, 파티션, 빈발 패턴

· 이 논문은 2010년도 정부(교육과학기술부)의 재원으로 한국과학재단의 국가 지정연구실사업(No.2010-0008007) 및 특정기초연구(No.2008-0052335) 지원으로 수행된 연구임

· 이 논문은 제36회 추계학술발표회에서 ‘단백질 2-DE 이미지 분석에서 데이터베이스 분할을 통한 최대 길이 스팟 매칭 패턴 검색’의 제목으로 발표된 논문을 확장한 것임

<sup>†</sup> 학생회원 : 연세대학교 컴퓨터과학과  
yienah@database.yonsei.ac.kr

<sup>\*\*</sup> 종신회원 : 연세대학교 컴퓨터과학과 교수  
leewo@database.yonsei.ac.kr

논문접수 : 2009년 12월 24일  
심사완료 : 2010년 2월 12일

Copyright©2010 한국정보과학회 : 개인 목적이거나 교육 목적인 경우, 이 저작물의 전체 또는 일부에 대한 복사본 혹은 디지털 사본의 제작을 허가합니다. 이 때, 사본은 상업적 수단으로 사용할 수 없으며 첫 페이지에 본 문구와 출처를 반드시 명시해야 합니다. 이 외의 목적으로 복제, 배포, 출판, 전송 등 모든 유형의 사용행위를 하는 경우에 대하여는 사전에 허가를 얻고 비용을 지불해야 합니다.

정보과학회논문지: 컴퓨팅의 실제 및 레터 제16권 제5호(2010.5)

**Abstract** In protein 2-DE image analysis, the accuracy of spot-matching operation which identifies the spot of the same protein in each 2-DE gel image is intensively influenced by the errors caused by the various experimental conditions. This paper proposes an efficient method to find more accurate spot-matching patterns based on multiple reference gel images in spot-matching pattern analysis in protein 2-DE image analysis. Additionally, in order to improve the reduce the execution time which is increased exponentially along with the increasing number of gel images, a “partition then extension” framework is used to find spot-matching pattern of long length and of higher accuracy. In the experiments on real 2-DE images of human liver tissue are used to confirm the accuracy and the efficiency of the proposed algorithm.

**Key words** : 2-DE image, spot-matching, image analysis, partition, frequent pattern

## 1. 서 론

단백질체학(Proteomics)[1]은 유전체에 의해 발견되는 단백질들을 다루는 학문으로 단백질은 생명 활동의 핵심 역할을 맡고 있기 때문에 질병 진단이나 치료제 개발의 궁극적 목표는 특정 단백질에서 찾을 수 있다. 대표적인 단백질분석 기술로 2차원 전기 영동(2-DE)[2]과 비-2차원 전기 영동 분석 방법[3,4]이 널리 사용되고 있다. 최근에 기술의 발전으로 인하여 “액체 크로마토그래프 질량분석장치(LC/MS)”나 “동위원소 기반 친화 태그(ICAT)”거나 “분자량 기반 질량 태그(MCAT)”를 사용하는 비-2차원 전기 영동 기술이 개발되어 사용되고 있지만 장비 비용, 분석에 소요되는 시간 등 문제로 인해 아직 많이 사용되고 있지 않다. 반면 2-DE분석 방법은 저렴한 비용으로 수천 개에 달하는 단백질을 동시에 효과적으로 분리하고 분석자의 경험에 의하여 실험에 자유롭게 개입할 수 있다는 등 장점을 가지고 있기 때문에 오랜 시간동안, 널리 사용되고 있는 기술이다. 따라서 최근 십여 년 동안 2-DE 이미지 분석을 위한 다양한 기술과 소프트웨어가 지속적으로 개발되고 있다[5,6].

단백질 2-DE 기술은 각 샘플에 들어있는 단백질들을 용해시킨 수용액을 pH값과 분자량에 의하여 각각 수평과 수직으로 분리시킨다[2]. 따라서 2-DE 이미지에서 스팟은 개개의 단백질을 의미하며 서로 다른 2-DE 이미지 상에 존재하는 같은 단백질을 효과적으로 찾아내는 것이 2-DE 이미지 분석 기술의 목표이다. 2-DE 이미지 분석 소프트웨어는 크게 “스팟 우선 검출” 방법[7,8]과 “이미지 변형 기반” 방법[9,10]으로 나누며 전자는 검출된 스팟 정보를 사용하여 참조이미지와 나머지 이미지들을 매핑시키고 후자는 이미지들을 변형시켜 동일

한 스팟을 가지는 하나의 가상이미지를 생성하는 것으로 동일한 단백질을 찾아낸다. 두 방법 모두 장단점이 있다. 전자의 경우 단백질을 찾기 위한 “스팟 검출”이 먼저 진행되고 그 정보를 기반으로 동일한 단백질을 연결시키기 위한 “스팟 매칭”이 진행된다[11]. 때문에 참조 이미지의 선택 및 이미지의 왜곡과 검출된 스팟에 포함된 노이즈에 따른 영향이 크다. 후자의 경우, 스팟 정보와 이미지 정보를 분리하지 않고 두 과정을 동시에 처리하기 때문에 “이미지 변형” 과정이 상당한 비용이 들고 분석 결과에 막대한 영향을 준다[12].

이러한 문제점을 해결하기 위하여 이 논문에서는 “스팟 우선 검출” 방법에서 추출된 단백질 스팟 정보를 사용하여 1) 객관적인 이미지 품질 척도를 지정하며, 이미지 품질 척도에 의하여 선택된 다중의 참조 이미지들을 사용하여 구축된 스팟 매칭 데이터베이스에서 2) 빈발한 스팟 매칭 패턴을 탐색함으로써 단백질 2-DE 이미지 분석의 정확도를 향상시키려고 한다.

2. 문제 정의

단백질 스팟 매칭 패턴 검색에서 문제에 대한 정의를 진행하기 위하여 표 1과 같은 기호를 사용한다.

표 1 기호 정의

스팟 집합	$S = \{s_1, s_2 \dots s_n\}$
이미지 집합	$G = \{g_1, g_2 \dots g_m\}, g_i = \{s_a, s_b \dots s_k\}$

이미지 분석에 사용된 전체 이미지 스팟집합을 S라고 하며, S는 모든 이미지의 스팟을 포함하고 있다. 이미지 집합을 G라고 하며, G에 포함된 임의의 이미지  $g_i$ 는 스팟 집합 S에 있는 한 개 이상의 스팟을 포함한다. 또한 S에 있는 한 스팟은 동시에 서로 다른 이미지에 포함될 수 없다. 즉,  $g_i \cap g_j = \emptyset (i \neq j \text{ \& } 1 \leq i, j \leq m)$ 의 조건을 만족해야 한다.

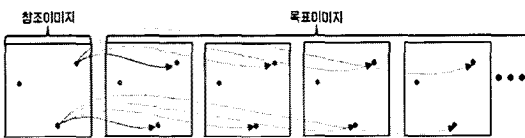


그림 1 단백질 이미지 “스팟 매칭” 과정

그림 1은 “스팟 우선 검출”방법에서 진행되는 “스팟 매칭”과정을 보여주는 것으로 이미지 상의 각 점은 단백질 스팟을 나타낸다. 사용자는 스팟 매칭을 진행하기 위하여 이미지집합에서 단 하나의 참조이미지를 선택하며, 선택된 참조이미지의 각 스팟과 그에 대응하는 목표 이미지의 한 스팟은 “매칭 페어”를 형성하며 다음과 같이 정의한다.

정의 1 매칭 페어

이미지 집합  $G = \{g_1, g_2 \dots g_m\}$ 에서  $g_i$ 를 참조이미지로 하였을 때  $g_i$ 의 스팟  $s_x$ 가 목표이미지  $g_j$ 의 스팟  $s_y$ 와 대응하는 관계를 가진다면 매칭 페어( $s_x, s_y$ )를 형성하며 모든 매칭 페어의 집합을  $P(g_i, g_j)$ 로 표시한다.

매칭 페어에서 두 스팟은 서로 다른 이미지에서 온 것이며, 첫번째 스팟은 참조이미지의 스팟이고 두번째 스팟은 목표이미지의 스팟을 의미한다. 또한 한 참조 이미지 스팟은 목표이미지에서 단 한 개의 스팟과 매칭 페어를 형성하며, 해당 참조이미지 스팟과 매칭 페어 스팟의 집합은 정의 2와 같이 매칭클래스를 구성한다.

정의 2 매칭클래스

매칭 페어의 집합을  $P(g_i, g_j)$ 에서 스팟  $s_x$ 을 포함하는 모든 매칭 페어의 스팟 집합을 “매칭클래스”라고 하며  $C(s_x) = \{s_x | (s_x, s_y) \in P(g_i, g_j)\}$ 로 표시한다.

매칭클래스에 포함된 모든 스팟은 동일한 단백질을 나타내며, 따라서 정확도가 높은 매칭클래스를 제공하는 것이 단백질 2-DE 이미지 분석의 핵심 목표라고 할 수 있다. 그러나 2-DE 실험은 환경에 민감하다. 특히 단백질 자체의 불안정성, 인위적이거나 주위환경으로 인하여 이미지가 왜곡될 수 있고, 또한 스팟 검출과 스팟 매칭을 분리하여 각각 독자적으로 진행하기 때문에 검출된 스팟에 포함된 노이즈 때문에 정확도가 높은 스팟 매칭 결과를 생성하는데 어려움이 존재한다. 이러한 문제점을 최소화하기 위하여 다중 참조 이미지를 사용함으로써 다음의 세 가지 목적에 의하여 이미지 분석에 미치는 영향을 감소시키거나 최소화 시키려고 한다.

- 1) 참조이미지 선택에 따른 영향: 다중의 참조 이미지를 선택함으로써 실험자의 주관적인 선택이 실험에 주는 영향을 최소화한다.
- 2) 이미지 왜곡에 의한 영향: 품질이 높은 참조이미지라고 하더라도 각 부분별로 서로 다르게 왜곡되어 있기 때문에, 왜곡된 부위를 서로 다른 참조 이미지들에 의하여 상호 보완하게 한다.
- 3) 스팟 검출 시 포함된 노이즈에 의한 영향: 먼지, 공기 방울과 같은 노이즈는 이미지 별로 서로 다르게 나타나기 때문에 매칭클래스의 실제 단백질보다 낮은 확률로 나타난다.

3. 최대 길이 스팟 매칭 패턴 탐색 방법

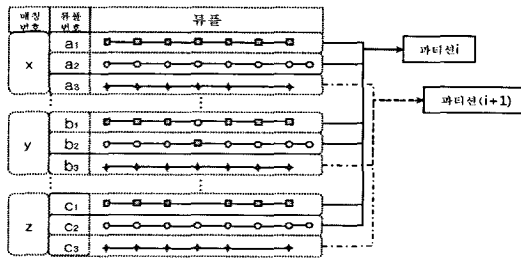
정확도가 높은 스팟 매칭 패턴 탐색을 위하여 우선 선택된 다중 참조이미지들을 사용하여 이미지 분석 SW를 통하여 이미지 별 스팟 매칭 정보를 수집하여 스팟

매칭 데이터베이스 D를 생성한다. 스팟 매칭 데이터베이스 D에서 전체 항목집합은 스팟 집합 S에 해당하며, S에 포함된 각 스팟은 항목에 해당한다. 하나의 스팟 매칭을 통하여 생성된 모든 매칭클래스는 튜플에 해당하며, 튜플은 유일한 튜플번호를 가지며 같은 스팟 매칭에 의하여 생성된 튜플은 동일한 매칭번호를 가진다. 각 튜플에서 참조이미지 스팟에 해당하는 항목을 참조항목이라고 하고 목표이미지 스팟에 해당하는 항목을 목표항목이라고 한다. 스팟 매칭 데이터베이스 D에서 빈발하게 발생하는 최대 항목집합이 바로 정확도가 높은 최대 길이 스팟 매칭 패턴에 해당한다. 스팟 매칭 데이터베이스 D에서 최대길이의 패턴을 탐색하기 위하여 다음과 같은 두 단계로 진행한다.

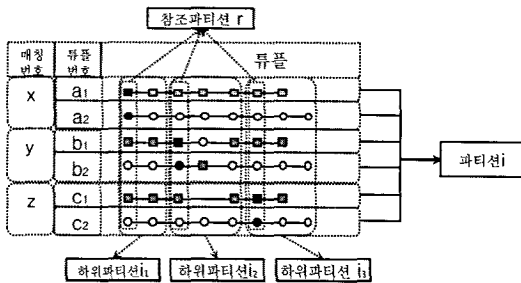
**1단계: 데이터베이스 분할**

• 데이터베이스의 수평 분할

스팟 매칭 데이터베이스에서 각 트랜잭션은 매칭클래스에 해당하기 때문에 같은 매칭번호를 가지는 트랜잭션들은 서로 같은 항목을 포함할 수 없다. 따라서 한 항목은 서로 다른 매칭번호를 가지는 트랜잭션 그룹에서 오직 하나의 대응하는 튜플을 가지며, 한 튜플은 그것이 포함하고 있는 모든 항목에 대하여 매칭번호별 대응하는 튜플들을 선택한다. 이렇게 선택된 튜플들로 구성된 데이터베이스의 각 파티션은 최소 하나의 매칭클래스로 구성된 튜플들의 집합이다. 그림 2는 이러한 방식에 의하여 데이터베이스를 분할하는 과정을 개념적으로 보여준다.



(a) 데이터베이스의 수평 분할



(b) 파티션의 수직 분할  
그림 2 데이터베이스 분할

그림 2(a)에서 모양이 같은 스팟은 동일한 단백질을 나타낸다. 튜플 a1과 a2는 매칭번호가 x인 트랜잭션그룹에서 서로 다른 매칭클래스임에도 불구하고 모두 동일한 “파티션i”로 할당되었다. 그것은 매칭번호가 y인 트랜잭션 그룹에서 b1(혹은 b2)은 스팟 매칭에서 발행한 문제로 인하여 서로 대상 단백질에 속하는 스팟을 교차하여 포함하고 있기 때문이다.

• 파티션의 수직 분할 및 참조파티션 생성

스팟 매칭 데이터베이스에서 트랜잭션의 길이를 감소시키기 위하여 수평 분할에 의하여 생성된 각 파티션을 참조이미지 스팟에 해당하는 항목을 기준으로 수직으로 분할하여 하위파티션을 생성한다. 각 하위파티션은 모두 같은 수의 참조 항목집합과 대상 항목집합을 포함하게 된다. 한 파티션이 k개의 하위파티션으로 분할되었을 경우, 튜플의 평균 길이는 원래 길이의 1/k로 감소된다. 또한 참조항목들로부터 구성된 참조파티션도 생성한다.

그림 2(b)는 그림 2(a)의 “파티션i”를 세 개의 하위파티션과 하나의 참조파티션으로 분할하는 것을 보여준다. 검은색의 항목들은 참조이미지 스팟에 해당하는 참조항목을 나타내며 각 하위파티션들은 모두 참조이미지 스팟에 해당하는 항목들을 균일하게 포함하고 있음을 보여준다. 이때 “참조파티션”은 모두 참조항목들로부터 구성된 파티션임을 알 수 있다. 기존의 빈발 패턴을 찾기 위한 알고리즘들은 패턴의 길이가 증가할 때, 수행 시간이 급격히 증가한다. 따라서 패턴의 길이를 최소화하는 것이 그 성능을 높이기 위한 유일한 방법이다. 또한 수평분할에서 모든 파티션이 “파티션(i+1)”과 같이 동일한 단백질로 구성된 항목집합들로 구성되는 것이 아니라, 실제로는 스팟 매칭의 문제점에 의하여 “파티션 i”와 같이 여러 매칭클래스에 속하는 파티션들로 구성될 수밖에 없다. 따라서 수직 분할은 꼭 필요한 과정이다.

**2단계: 최대 길이 패턴 탐색**

• 하위파티션 별 빈발항목 탐색

1단계에서 생성된 참조파티션과 각 하위파티션은 각각 기존의 빈발항목 탐색 알고리즘을 사용하여 최대 빈발 항목집합을 탐색한다.

• 최대 빈발항목집합 확장

빈발항목 탐색을 통하여 각 하위파티션은 상호 배타적인 최대 빈발항목집합을 생성한다. 따라서 상호 배타적인 최대항목집합을 하나로 확장하기 위한 연결고리로 참조파티션에서 생성된 빈발항목집합을 사용한다.

그림 3은 그림 2(b)의 하위파티션에서 탐색된 빈발항목 집합을 참조파티션의 빈발항목집합을 기초로 확장하는 과정을 보여준다. 그림에서 검은 색 항목집합은 참조파티션에서 생성된 빈발항목 집합이다.

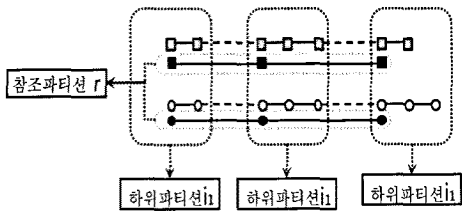


그림 3 참조파티션 기반 최대 반발항목 확장

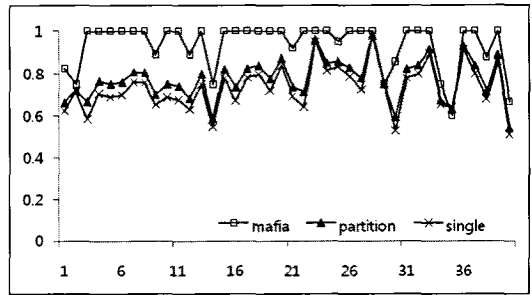
스팟 매칭의 특성에 의하여 데이터베이스를 분할함으로써 반발한 항목집합들을 탐색함으로써 연관성이 없는 튜플들을 따로 처리할 수 있기 때문에 후보 반발항목집합 계산을 위한 저장공간과 시간을 절약할 수 있다. 만일 첫 번째 단계에서 데이터베이스를 분할하지 않는다면, 반발항목집합 계산의 중간 과정에 생성되는 후보 반발항목집합은 그 크기가 엄청날 것이며 특히 사용되는 이미지의 개수에 따라서 기하급수적으로 증가할 것이다.

두 번째 단계에서 최대 반발항목 합을 탐색하는 이유는 2-DE 이미지 스팟 매칭에서는 최대한 길이가 길면서 정확도가 높은 매칭 패턴을 찾는 것이 추후의 생물학적인 분석에 더 유용하게 사용된다. 또한 최대반발항목을 찾음으로써 모든 반발한 항목집합을 모두 유지할 필요가 없기 때문에 실행 속도를 향상시킬 수 있다.

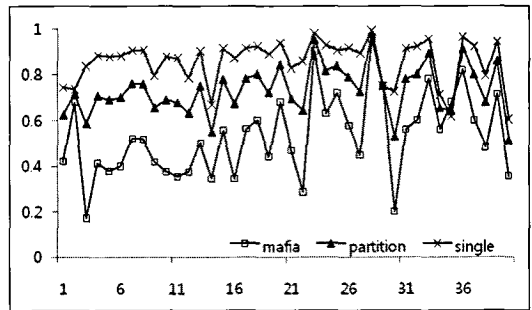
4. 실험

실험에서는 50개의 인간간세포 샘플로 생성한 2-DE 이미지를 사용하였다. 각 이미지가 포함하고 있는 스팟 수는 1000개에서 3000개 사이로 편차가 매우 크며 각 이미지는 서로 다른 종류의 에러와 다른 패턴으로 왜곡되어 있다. 실험에서는 참조 이미지선택의 주관성을 배제하기 위하여 각 이미지가 포함하고 있는 스팟 수를 이미지 품질 판단을 위한 척도로 사용하였으며, 품질이 좋은 상위 10개의 이미지를 선택하여 10번의 반복적인 스팟 매칭을 통하여 스팟 매칭 데이터베이스를 구축하였다. 따라서 참조파티션의 각 튜플은 최대 10개의 참조항목으로만 구성되며 각 하위파티션은 최대 2개의 참조항목과 8개의 목표항목을 포함하게 되며 한 파티션은 5개의 균일한 하위파티션으로 분할되었다.

정확도 비교 실험에서는 PDB[14]에 등록된 40개의 단백질에 해당하는 매칭클래스를 선택하였으며, 생성된 매칭클래스의 정확도(precision)와 재현율(recall)을 사용하여 성능을 비교하였다. 정확도는 생성된 매칭클래스에 포함된 정확한 단백질 스팟의 비율을 나타내고, 재현율은 실제 매칭클래스 스팟 수 대비 생성된 매칭클래스에 포함된 정확한 스팟의 비율을 나타낸다.



(a) 정확도

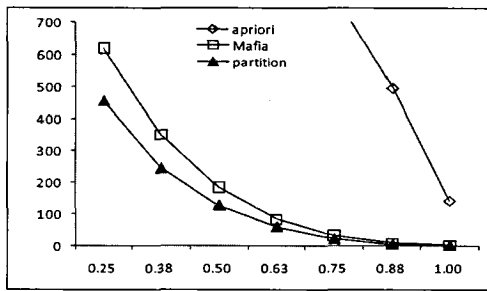


(b) 재현율

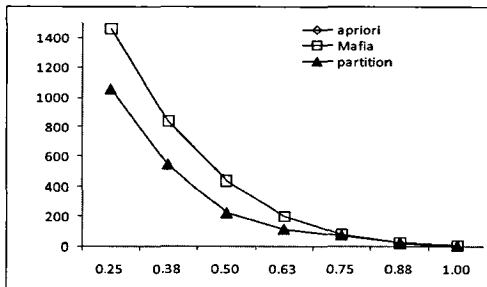
그림 4 매칭클래스 정확도 비교

그림 4(a)에서 “mafia”[13]알고리즘을 사용하여(최소 지지도=0.5) 생성된 최대길이의 반발패턴을 말하며, “partition”은 이 논문에서 제안한 방법으로 생성한 스팟 매칭 패턴이다(하위파티션 최소지지도=0.5). 그리고 “single”은 한 참조이미지를 사용하여 생성된 매칭클래스를 나타낸다. x축은 선택된 샘플 매칭클래스의 번호를 보여주며 y축은 각각 정확도와 재현율을 나타낸다. 그래프에서 알 수 있듯이, 제안한 방법의 정확도는 “single”보다 높게 나왔지만 “mafia”보다는 낮다. 이와 반대로 그림 4(b)에서는 제안한 방법의 재현율은 “mafia” 알고리즘보다 높고 “single”보다는 낮다. 그것은 각 하위파티션에서 최소지지도 이하의 항목들이 제거되었기 때문이다. 따라서 제안한 방법은 재현율과 정확도를 적절하게 조절함으로써 길이를 유지하면서 동시에 정확도를 증가시킬 수 있다.

그림 5에서는 제안한 방법과 두 개의 반발항목 패턴을 찾는 알고리즘 “mafia” 및 “apriori”의 실행 시간을 비교하였다. 두 그래프의 x축은 최소지지도를 나타내고(제안한 방법에서는 하위파티션 최소지지도), y축은 실행 시간(단위:초)을 나타낸다. 제안한 “partition”방법은 그림 5(a)와 그림 5(b)에서 모두 “mafia” 및 “apriori”보다 좋은 성능을 나타냄을 알 수 있다. 그것은 제안한 방법은 데이터베이스를 분할하여 하위파티션 내에서 최대



(a) 이미지 수 = 24



(b) 이미지 수 = 36

그림 5 실행 시간 비교

항목을 찾기 때문에 최대항목의 길이의 영향, 즉 이미지 수의 영향을 덜 받기 때문이다. 하지만 반대로 각 하위 파티션에서 생성된 빈발항목을 다시 확장하는데 걸리는 시간 때문에 아쉽게도 실행시간은 큰 폭으로 줄어들지는 않지만 "mafia"보다는 항상 좋은 성능을 나타낸다.

## 5. 결론

이 논문에서는 단백질 2-DE 이미지 분석에서 한 참조 이미지를 사용한 스팟 매칭에 수반되는 문제점을 해결하고자 다중의 참조 이미지를 사용함으로써 1) 스팟 검출과정에서 생성되는 문제에 의한 오류를 최소화하고, 2) 이미지 왜곡에 의한 영향을 상호 보완하도록 하였으며, 3) 참조 이미지 선택에 의한 영향을 최소화 하도록 하여 보다 정확도가 높은 매칭 패턴을 제공하였다. 또한, 4) "분할 및 확장"방법을 사용함으로써 이미지 수에 따른 실행 시간 증가폭을 효과적으로 감소시켰다. 또한 실험에서는 실제 인간 샘플의 2-DE 이미지를 사용하여 제안한 방법의 타당성을 검증하였다.

현재 "스팟 우선 검출"을 기반으로 하는 패키지들은 100%의 스팟 매칭(모든 이미지에서 해당 스팟을 검출하는 것)을 목표로 하고 있다. 따라서 기존에 찾은 항목 집합을 핵심으로 하여 확장함으로써 길이가 더 길면서 정확도가 사용자 요구사항에 만족하는 스팟 매칭 패턴을 찾는 연구를 계속하여 진행할 것이다.

## 참고 문헌

- [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*, 21, pp.1037-1053, 2000.
- [2] O'Farrell PH, "High resolution two-dimensional electrophoresis of proteins, *Journal of Biology and Chemistry*," 250, pp.4007-4021, 1975.
- [3] Cagney G. and Emili A., "De novo peptide sequencing and quantitative profiling of complex protein mixtures using mass-coded abundance tagging," *Nat. Biotech.*, 20, pp.163-170, 2002.
- [4] Mann M, Hendrickson RC and Pandey A., "Analysis of proteins and proteomes by mass spectrometry," *Biochem.*, 70, pp.437-443, 2001.
- [5] Senthilkumar D. and Richard AR, "Minimizing variability in two-dimensional electrophoresis gel image analysis," *OMICS: A Journal of Integrative Biology*, 11(2), pp.225-230, 2007.
- [6] Spyros G., Gert L. and Michael F., "Limitations of current proteomics technologies," *Journal of Chromatography A*, 1077, pp.1-18, 2005.
- [7] Piers M. and Paul D.: Quantitative and reproducible two-dimensional gel analysis using Phoretix 2D Full, *ELECTROPHORESIS*, 22(10), pp.2075-2085, Jun 2001.
- [8] Appel RD, Palagi PM, Walter D, Vargas JR, Sanchez J, Ravier PC and Hochstrasser DF, "Melanie II -- a third generation software package for analysis of two-dimensional electrophoresis images: I: Features and user interface," *Electrophoresis*, 18(15), pp.2735-2748, 1997.
- [9] Arsi TR, Jussi MS, Tero A, Jan W, Riitta L, Tuula AN and Olli SN, "Comparison of PDQuest and Progenesis software packages in the analysis of two-dimensional electrophoresis gels," *PROTEOMICS*, 3(10), pp.1936-1946, Jul 2003.
- [10] Karp NA, Feret R, Rubtsov DV and Lilley KS: Comparison of DIGE and post-stained gel electrophoresis with both traditional and SameSpots analysis for quantitative proteomics, *Proteomics*, 8, pp.948-960, 2008.
- [11] Angelika G, Walter W and Michael JD, "Current two-dimensional electrophoresis technology for proteomics," *Proteomics*, 4, pp.3665-3685, 2004.
- [12] Clark BN and Gutstein HB, "The myth of automated, high-throughput two-dimensional gel analysis," *Proteomics*, 8(6), pp.1197-1203, Mar 2008.
- [13] Burdick D., Calimlim M., Flannick J., Gehrke J., Yiu T., "MAFIA: a maximal frequent itemset algorithm," *IEEE Transactions on Knowledge and Data Engineering*, 17(11), pp.1490-1504, 2005.
- [14] <http://www.pdb.org/> (accessed on Sep 1th, 2009)