단백질 2DE 이미지에서 참조 이미지에 의한 유사도 기반 에러 스팟 필터링 기법

김연화*, 심정은, 이원석
연세대학교 컴퓨터과학과
e-mail :{yienah*, jjuggeuni, leewo }@database.yonsei.ac.kr

Error Spot Filtering Based on Similarity of Reference Image In Protein 2DE Image

Yan Hua Jin, Jung Eun Shim, Won Suk Lee
Dept. of Computer Science, Yonsei University

요 약

단백질 2DE 이미지 분석의 주요작업은 스팟 매칭에 의한 동일한 종류의 단백질 그룹인 패어링 클래스를 구성하는 것으로서 단백질간의 상호 작용, 질병에 관련한 단백질의 변화 등을 관찰할 수 있다. 하지만 2DE 실험의 여러 가지 문제점으로 인하여 패어링 클래스는 먼지, 공기방울 등 에러를 포함하게 되며 이런 에러들은 왜곡된 분석결과를 초래한다. 따라서 본 논문에서는 동일한 조직에서 같은 종류의 단백질은 발현량이 비슷하다는 특성을 이용하여 패어링 클래스의 개개의 스팟을 참조 스팟 속성으로 나눈 값을 유사도로 정의하고, 스팟의 유사도가 사용자에 의하여 선택되는 필터링 배수에 의한 범위를 벗어날 때 에러 스팟으로 간주하여 제거되는 에러 필터링 기법을 제안한다. 실험에서는 정확도(Precision), 재현율(Recall) 및 조화평균(Harmonic-mean) 값을 사용하여 제안된 필터링 기법의 타당성을 보여준다.

1. 서론

현재 유전체학 분야에서 DNA 에 대한 연구가 발전함에 따라 각종 질병에 직접적으로 관련되는 단백질에 대한 분석이 중요한 이슈가 되고 있다. 그러나 DNA 와는 달리 단백질은 샘플 보관상태, 실험환경 등 여러 가지 요인에 따라 동적으로 변화되는 것으로 단백질에 관한 연구는 훨씬 더 많은 어려움이 존재한다.

단백질체학(Proteomics)분야에서 2DE(two dimensional Electrophoresis)[1]는 조직내의 수천 개의 단백질을 동시에 분리해 내는 기법으로 단백질 분석을 위한 주요한 기술이다. 수천 개의 단백질 분석을 자동으로 수행하기 위하여 2DE 이미지분석 소프트웨어를 사용하며, 조직내의 모든 단백질은 2DE 이미지에서 개개의 스팟으로 나타난다.

단백질체학[4][6]에서 상이한 단백질들 사이의 상호작용을 밝히는 것이 중요한 이슈의 하나이다. 따라서 동일한 조직의 2DE 이미지들 가운데서 동일한 종류의 단백질 스팟을 찾아내는 스팟 매칭 작업은[7][8] 2DE 이미지 분석에서의 주요 작업이다. 또한 조직 내 단백질의 수가 많고 분석하려고 하는 2DE 이미지 수가 많기 때문에 스팟 매칭 작업은 자동으로 처리되어야 한다.

2DE 이미지를 생성하기 위한 2 차원 전기영동실험은 인위적인 과정이기 때문에 실험환경으로 인하여 먼지, 공기 방울 등 이물질이 들어갈 수 있으며 이미지 제작시의 찌그러짐, 흐림 등으로 하여 이미지 품질에 의한 오류가 생길 수 있다[2][3]. 이런 오류들은 이미지분석 소프트웨어에 의하여 전부 복구될 수 있는 것이 아니다. 따라서 이런 오류가 포함된 2DE 이미지들을 대상으로 스팟 매칭을 진행하여 생성된 동일한 종류의 단백질 그룹에는 먼지, 공기방울 등의 오류와 이미지 품질로 인한 질량, 크기 등이 다른 상이한 종류의 단백질 스팟이 포함될 수 있다. 이러한 오류들은 단백질 발현량 분석결과에 치명적인 영향을 준다[5]. 따라서 본 논문에서는 이런 오류를 제거하기 위한 방법을 제시하고자 한다.

2. 관련연구

현재 2DE 이미지 분석을 위하여 Melanie[7], Progenesis[8] 등 많은 상용 소프트웨어들이 활용되고 있으며 이 소프트웨어들의 주요 기능은 스팟 감지(Spot detection) 작업과 스팟 매칭(Spot matching) 작업이 있다.

스팟 감지과정에서는 사용자가 입력한 각종 파라메터에 의하여 2DE 이미지의 스팟을 자동으로 감지한다. 이때 모든 스팟은 좌표 값 (x, y), 면적에 관련된 area, 체적 발현량과 관련된 Vol, %Vol, Od, %Od 등 속성값을 부여받게 된다.

스팟 매칭 과정에서는 감지된 스팟을 대상으로 동일한 종류의 단백질 스팟 그룹을 생성한다. 스팟 매칭을 위하여 사용자는 하나의 이미지를 참조 이미지로 지정해 주는데 소프트웨어는 참조 이미지의 모든 스팟을 기준으로 하여 대상 이미지에서 같은 종류의 단백질 스팟을 선택하여 하나의 그룹을 생성한다.

위의 과정을 거쳐 생성된 동일한 종류의 단백질 스팟 그룹은 2DE 실험의 에러, 이미지분석 소프트웨어에 의한 스팟 감지 에러 및 스팟 매칭 과정의 에러들에 의한 많은 문제점을 포함할 수 있다. 에러를 포함한 동일한 단백질 스팟 그룹은 단백질 발현량 분석, 단백질 특성 분석, 각종 속성의 통계 등 많은 면에서 왜곡된 수치를 제공한다. 예를 들면 정상 조직에서 나타나는 특정 단백질과 비정상 조직에서 나타나는 단백질 발현량의 차이를 분석하기 위하여 두 그룹 내 이미지들의 스팟 매칭을 진행한다. 스팟 매칭에 의하여 생성된 정상군과 비정상군의 특정 단백질 스팟 그룹에 대해 각종 통계를 수행하게 되는데, 이때 각 이미지에 포함된 오류는 통계 결과에 심각한 영향을 줄 수 있다. 그러나 상용 이미지분석 소프트웨어 자체는 동일한 단백질 그룹의 에러를 제거하는 방법을 제공하지 않는다. 따라서 사용자는 보다 정확한 결과를 얻기 위하여 수천 개의 단백질 그룹에 대하여 직접 수동으로 확인하는 작업을 반복하여야 한다.

본 논문에서는 수동으로 진행되어야 하는 불필요한 반복작업을 해결하기 위하여 참조 이미지를 기준으로 사용자의 파라메터 설정에 의하여 동일한 종류의 단백질 그룹에 포함되어 있는 에러를 자동적으로 제거하는 방법을 제안한다.

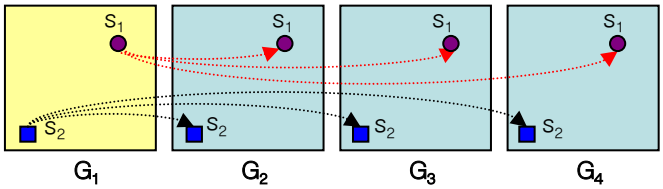
3. 패어링 클래스에서 유사도에 의한 에러 제거 방법

2DE 이미지는 이미지분석 소프트웨어를 통하여 조직상의 수천 개의 단백질을 스팟으로 자동 검출한다. 이런 스팟들은 이미지상의 좌표 축에 의하여 고유한 번호인 spotID 가 부여되며 스팟 크기, 농도, 모양에 의하여 area, %Vol, %Od, Vol, Od 등 스팟 속성이 부여된다.

정의 1. 2DE 이미지와 스팟의 속성

2DE 이미지가 n 개 있을 때 이미지 집합 G 는 $G=\{G_1, G_2, \ldots, G_n\}$이고 스팟 집합 G_i는 $G_i=\{ G_i.s_t \mid G_i.s_t$는 이미지 G_i에 존재하는 스팟 $(1\le t\le m,\ m$ 은 이미지 G_i 내의 스팟 수)$\}$ 이다. 스팟 $G_i.s_t$ 의 k 번째 속성은 $G_i.s_t(att_k)$로 나타낸다. □

2DE 이미지 분석에서 스팟 매칭 과정은 이미지 집합에서 선택된 하나의 참조 이미지 내에 존재하는 모든 참조 스팟들에 대하여, 각 대상 이미지에서 대응되는 스팟들을 찾음으로써 동일한 단백질 스팟 그룹을 생성하며 생성과정은 그림 1 과 같다.



(그림1) 참조 이미지 G_1에 의한 스팟 매칭 과정

그림1의 스팟 매칭 과정에서 참조 이미지의 스팟 $G_1.s_1$을 참조 스팟으로 하여 각각의 대상 이미지에서 스팟 $G_2.s_1, G_3.s_1, G_4.s_1$이 선택된다. 따라서 동일한 종류의 단백질 스팟 그룹 $\{G_1.s_1, G_2.s_1, G_3.s_1, G_4.s_1\}$이 생성된다. 이렇게 스팟 매칭을 통하여 하나의 그룹으로 형성되는 동일한 종류의 단백질 스팟의 집합을 패어링 클래스(Pairing Class)라고 하며 다음과 같이 정의한다.

정의 2. 패어링 클래스(Pairing Class)

N 개의 이미지 집합 G 에서 $G_r \in G$ 을 참조 이미지로 하였을 때 참조 스팟 $G_r.s_j$ 에 의하여 형성된 패어링 클래스는 $PG_r.s_j= \{G_1.s_{t1}, G_2.s_{t2}, \ldots, G_r.s_j, \ldots, G_n.s_{tn}\}$이며 패어링 클래스 $PG_r.s_j$의 길이는 집합 $PG_r.s_j$ 의 원소 개수이다. □

스팟 매칭에서 한 스팟은 동시에 참조 이미지의 서로 다른 스팟에 부합될 수 없기 때문에 패어링 클래스들의 교집합은 공집합이다($PG_r.s_j \cap PG_r.s_k = \Phi$).

한 패어링 클래스에 속한 스팟들은 서로 다른 이미지로부터 온 것으로서 이미지분석 소프트웨어에 의하여 참조 이미지의 기준 단백질 스팟과 같은 종류의 단백질로 판단된 스팟들의 집합이다. 만일 한 패어링 클래스가 포함하고 있는 모든 스팟이 참조 이미지 스팟과 동일한 종류의 단백질이라고 하면 그들의 각종 속성은 비슷한 양상을 나타낼 것이다. 같은 조직의 동일한 종류의 단백질 덩어리는 생물 객체의 차이, 실험 환경의 차이 등으로 하여 미세한 차이는 있지만 대체적으로 비슷한 속성을 나타내기 때문이다.

단백질 2DE 이미지 분석에서 생성된 패어링 클래스에는 참조 스팟의 단백질과 다른 먼지, 공기방울 등 에러 스팟이 포함되어 있다. 이런 에러 스팟의 각종 속성은 참조 이미지 스팟의 속성과 확연히 구별될 것이다. 따라서 본 논문에서는 에러 스팟의 속성이 참조 스팟의 속성과 다르다는 점을 이용하여 패어링 클래스의 에러 스팟 필터링 기법을 제안한다.

패어링 클래스 정의에서도 알 수 있듯이 모든 패어

링 클래스는 참조 스팟이 존재한다. 참조 스팟은 참조 이미지의 스팟으로서 스팟 매칭과정에 이 스팟을 기준으로 하여 대상 이미지로부터 같은 스팟을 선택한다. 따라서 한 패어링 클래스에서 어떤 스팟이 그 패어링 클래스에 포함되기 위해서는 그 패어링 클래스의 참조 스팟의 속성과 비슷해야 한다. 따라서 각 패어링 클래스 내의 모든 스팟 원소에 대하여 그 패어링 클래스에 얼마나 적합한지를 나타내는 값인 유사도를 부여한다.

정의 3. 유사도 (Similarity Value)

패어링 클래스 $PG_r.s_j = \{G_1.s_{t1}, G_2.s_{t2}, \ldots, G_r.s_j, \ldots, G_n.s_{tn}\}$ 에서 스팟 $G_n.s_{tn}$ 의 유사도는 $SV(G_n.s_{tn}) = G_n.s_{tn}(att_k) / G_r.s_j(att_k)$이다. □

패어링 클래스에서 어떤 스팟의 유사도 값이 1 에 가까울수록 정확한 스팟일 가능성이 크다. 따라서 유사도 값이 지나치게 크거나 작을 경우 해당 스팟을 에러로 판단한다. 즉 유사도 값의 로그 스케일에 의한 정규분포에서 변두리에 있는 스팟은 에러 스팟일 가능성이 크기 때문에 이런 스팟을 제거함으로써 패어링 클래스의 에러 스팟을 제거할 수 있다.

본 논문에서 제안하는 참조 스팟에 의한 패어링 클래스의 에러 스팟 검출 방법은 아래와 같은 순서로 진행된다.

Step1. 패어링 클래스 생성

2DE 이미지분석 소프트웨어에 의하여 감지된 스팟을 대상으로 이미지 그룹 G 에서 참조 이미지 G_r 의 모든 스팟을 참조 스팟으로 하여 패어링 클래스를 생성한다. 참조 이미지는 이미지 그룹에서 품질이 좋은 이미지로서 스팟들의 각종 속성(Circularity, %Vol, area 등)의 평균, 표준편차 등에 의하여 선택된다.

Step2. 스팟 속성의 선택

단백질 발현량과 관련된 스팟의 속성에서 area, Vol, Od 은 2DE 이미지의 절대적인 속성이고 %Vol, %Od 등 속성은 정규화(Normalization)된 상대적인 속성이다. 본 논문에서는 동일한 조직에서 같은 종류의 단백질은 발현량이 비슷하다는 것에 초점을 두기 때문에 단백질 발현량과 관련된 상대적인 속성 %Vol, %Od 등을 선택하여 스팟의 유사도 값을 구하는 속성으로 사용한다.

Step 3. 스팟의 유사도 값 계산

Step2 에서 선택한 스팟 속성에 의하여 정의 3 과 같이 모든 패어링 클래스의 각 원소에 대하여 유사도 값을 계산한다. 유사도는 참조 이미지의 선택과 직접적인 연관을 가지기 때문에 Step1 에서 품질이 낮은 참조 이미지가 선택되었을 때 필터링 결과의 정확도는 낮게 나올 수도 있다.

Step 4. 필터링 배수에 의한 에러 스팟 필터링

필터링 배수란 사용자에 의하여 선택되는 값으로서 한 패어링 클래스에서 어떤 스팟의 유사도 값이 필터링 배수 F 에 의한 범위 [1/F, F]에 포함 되지 않을 때 해당 스팟은 에러 스팟으로 간주되어 패어링 클래스에서 제거된다. 필터링 배수 F 는 너무 크거나 작게 주어서는 안 된다. F 를 크게 주면 참조 스팟의 속성값과 차이가 큰 에러 스팟을 검출 할 수 없고 작게 주게 되면 객체의 차이에 의한 유사도 값의 차이를 에러로 인식하기 때문에 정확한 스팟이 제거된다.

위의 4 개 단계에 의하여 패어링 클래스의 에러 스팟을 제거할 수 있다. 필터링 배수 F 와 필터링 정확도, 재현율, 조화평균의 관계에 대한 그래프는 실험에서 보여준다.

4. 실험

본 논문에서 사용한 실험 데이터는 인간 간암 조직에 대한 2DE 이미지 53 쌍(정상-비정상)을 Melanie III 소프트웨어에 의하여 스팟 매칭을 진행하여 얻은 패어링 클래스 데이터이다. 2DE 이미지에 대한 스팟 감지 작업 및 스팟 매칭을 위한 기준점 입력 과정[7]은 생물학자들에 의하여 진행되었기 때문에 비교적 신뢰도가 높은 데이터라고 할 수 있다. 하지만 이렇게 전처리가 잘 된 데이터에도 그것들의 패어링 클래스는 에러 스팟을 포함하고 있다. 따라서 실험에서는 필터링 배수를 설정하여 패어링 클래스에 숨어 있는 에러 스팟을 제거하고 그 정확도를 그래프로 나타낸다.

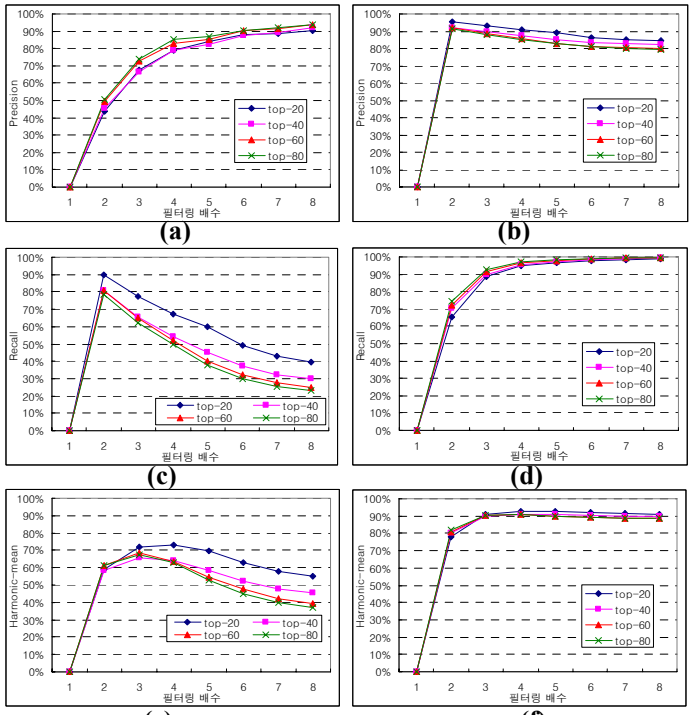
2DE 이미지에는 수천 개의 단백질 스팟이 포함되기 때문에 패어링 클래스도 수천 개가 생성 될 수 있다. 실험에서 필터링 결과의 정확도를 계산하기 위하여 해당 패어링 클래스에 포함된 스팟의 정확 여부를 알아야 한다. 때문에 패어링 클래스의 모든 스팟을 일일이 수작업으로 확인해야 하는데 수천 개의 패어링 클래스를 스팟 마다 일일이 확인하는 것은 불가능하다. 따라서 본 논문에서는 랜덤 샘플링 방법으로 패어링 클래스들을 선택하여 그들이 포함하고 있는 스팟들의 정확도 여부를 수작업으로 확인한다.

단백질 스팟 속성 중 특히 %Vol 은 단백질의 발현량을 나타내는 상대적인 속성으로서 이미지의 품질에 상관없이 단백질의 발현량을 상대적인 수치로 표현이 가능하기 때문에 실험에서는 스팟의 유사도를 계산할 때 %Vol 속성을 사용한다.

실험결과의 정확도를 나타내기 위하여 정확도(Precision), 재현율(Recall) 및 조화평균(Harmonic-mean) 값을 사용한다. 패어링 클래스 $PG_r.s_j$ 는 V 개의 에러 스팟과 W 개의 정확한 스팟을 포함하고 있을 때, 필터링 배수 F 에 의하여 v 개의 에러 스팟과 w 개의 정확한 스팟이 필터링 된다. 이때 필터링 된 스팟에서 에러 스팟에 대해서 보면, 정확도는 v/(v+w), 재현율은 v/V 이다. 반대로 필터링에서 제거되지 않고 남은 스팟에서 정확한 스팟에 대해서 보면 정확도는 (W-w)/(V-v+W-w), 재현율은 (W-w)/W 이다.

실험에서는 필터링 배수 F 에 의하여 패어링 클래스에서 유사도가 [1/F, F] 범위에 있지 않은 스팟을 에러 스팟으로 필터링 하기 때문에 F 의 값이 클수록 많은 에러 스팟이 제거되는 동시에 정확한 스팟도 같이 제거 된다. 즉 필터링 된 에러 스팟의 정확도가 증가할 때 재현율은 감소하게 된다. 따라서 필터링의 정확도를 확인하기 위하여 정확도와 재현율 값으로 표

현되는 조화평균 값을 사용하며 조화평균은 2/(1/정확도+1/재현율) 이다. 에러 스팟 필터링을 통해서 얻은 조화평균 값이 클 때 정확도와 재현율은 합리하며 따라서 선택된 필터링 배수도 적합하다는 것을 알 수 있다.



(그림 2) 에러 필터링에 의한 정확도 그래프

실험 그래프에서 x 축은 필터링 배수, y 축은 정확도이고 계열 top-n 은 각 패어링 클래스 스팟의 %Vol 값의 표준편차가 큰 클래스부터 n 개 선택한 것이다.

실험 1. 필터링 배수에 의하여 에러를 제거할 경우 제거된 스팟에서 에러 스팟의 정확도

패어링 클래스에서 필터링의 정확도를 확인 위하여 필터링 된 스팟에서 에러 스팟의 정확도를 본다. 그림 2 에서 (a)는 필터링 배수 값이 증가할 때 필터링 된 스팟에서 에러 스팟의 정확도, (c)는 재현율을 나타낸다. 필터링 배수가 증가할 때 정확도는 증가하지만 재현율은 감소한다. 그것은 필터링 배수가 커질 때 필터링 되는 스팟에서 에러 스팟의 비율이 크다는 것을 나타낸다. (e)는 에러 스팟의 조화평균을 보여주며 필터링 배수가 4 일 때 제일 좋은 필터링 효과를 가진다. 모든 정확도가 계열 top-20 에서 제일 높게 나오는데 그것은 패어링 클래스에서 스팟 속성의 편차가 클수록 에러 필터링이 잘 된다는 것을 알 수 있다.

실험 2. 필터링 배수에 의하여 에러를 제거한 후 남은 패어링 클래스의 정확도

그림 2 에서 (b), (d), (f) 은 필터링을 통해 남은 패어링 클래스에서 정확한 스팟의 정확도, 재현율, 조화평균을 보여준다. 그래프에서 필터링 배수가 증가할 때 정확도 값은 증가하고 재현율은 감소한다. 그것은 필터링 배수가 커질 때 제거되는 정확한 스팟의 비중이 감소한다는 것을 알 수 있다. 정확도와 재현율에 의한 조화평균은 필터링 배수가 4 일 때 가장 높게 나왔다. 계열로 보면 조화평균과 재현율은 top-20 에서 제일 높게 나왔기 때문에 배수에 의한 필터링에서 패어링 클래스의 스팟 속성의 편차가 클수록 더 좋은 필터링 결과를 가진다는 것을 알 수 있다.

5. 결론 및 향후 연구

본 논문에서는 단백질 2DE 이미지 분석에서 패어링 클래스에 포함된 에러를 참조 이미지의 스팟에 의한 유사도를 기반하여 필터링하는 방법을 제안한다.

패어링 클래스는 참조 스팟을 기준으로 하여 생성된 동일한 종류의 단백질 스팟 그룹으로서 그들의 속성은 참조 스팟의 속성과 비슷해야 한다. 대상 이미지 스팟 속성을 참조 스팟 속성으로 나눈 유사도는 그것들이 동일한 종류의 스팟일 경우에는 1 에 가까울 것이다. 어떤 스팟의 유사도가 필터링 배수 F 에 의한 범위 [1/F, F]에 속하지 않으면 에러 스팟으로 판단하여 제거함으로써 패어링 클래스의 정확도를 보장한다.

본 논문에서는 실험을 통하여 제안한 에러 필터링 기법의 타당성을 증명하였다. 실험에서 사용한 패어링 클래스 데이터는 하나의 참조 이미지에 의하여 생성된 데이터이기 때문에 유사도 계산에서 다소 오차를 포함 할 수 있다. 이런 오차를 극복하기 위해 향후 연구에서는 여러 개의 참조 이미지를 사용하여 유사도를 계산하는 기법을 고려할 것이다.

본 논문에서 제안한 패어링 클래스의 에러 스팟 필터링 기법을 향후 추진할 자동화 단백질분석시스템의 구축에서 사용자의 수요에 의하여 데이터 정제 작업을 할 수 있는 전처리 모듈로 추가함으로써 생물학자에게 자동화된 높은 품질의 데이터를 보장하는 시스템을 구축하고자 한다.

참고문헌

[1]Takahashi K., Nakazawa M., “DNAinsight: An Image Processing System for 2-D Gel Electrophoresis of Genomic DNA”, Proc. of Genome Informatics Workshop, 8: 135-146, 1997

[2]Frank H., Klaus K., Carola W., “Matching 2D Patterns of Protein Spots”, In Proceedings SoCG’98, 231-239, 1998

[3]Stefan Veeser, Michael J. Dunn, Guang-zhong Yang, “Multiresolution image registration for two-dimensional gel electrophoresis”, Proteomics, 1, 856-870, 2001

[4]Drs Scott D Patterson, Terence E Ryan, “Proteomics-A new Toll for Biology”, Business Briefing Pharmatech, 96-101, 2002

[5]S.Y. Cho, K.-S. Park, J.E. Shim, M.-S. Kwon, K.H. Joo, W.S. lee, J.Chang, H.kim, H.C.Chung, H.O.Kim, Y.-K.Paik, “An integrated proteome database for two-dimensinal electrophoreses data analysis and laboratory information management system”, Progeomics, 2, 1104-1113, 2002

[6]Rabilloud, T., “Proteomics”, 2 , 3-10, 2002

[7]Gene Bio, Melanie Ⅲ Image Analysis Program Manual

[8]NLD, Progenesis Image Analysis Program manual