한국생물공학회지 제20권 제2호
Korean J. Biotechnol. Bioeng.
Vol. 20, No. 2, 71--75(2005)

2D-Gel 이미지의 정렬 및 클러스터링

허 원
강원대학교 바이오산업공학부
(접수 : 2005. 3. 20., 게재승인 : 2005. 4. 23.)

Clustering of 2D-Gel Images

Won Hur
School of Biotechnology and Bioengineering, Kangwon National University, Chunchon, Kangwon 200-701, Korea
(Received : 2005. 3. 20., Accepted : 2005. 4. 23.)

Alignment of 2D-gel images of biological samples can visualize the difference of expression profiles and also inform us candidates of protein spots to be further analyzed. However, comparison of two proteome images between the case and control does not always successfully identify differentially expressed proteins because of sample-to-sample variation, poor reproducibility of 2D-gel electrophoresis and inconsistent electrophoresis conditions. Multiple alignment of 2D-gel image must be preceded before visualizing the difference of expression profiles or clustering proteome images. Thus, a software for the alignment of multiple 2D-Gel images and their clustering was developed by applying various algorithms and statistical methods. Microsoft Visual C++ was used to implement the algorithms in this work. Multiresolution-multilevel algorithm was found out to be suitable for fast alignment and for largely distorted images. Clustering of 10 different proteome images of Fetal Alcohol Syndrome, was carried out by implementing a k-means algorithm and it gave a phylogenetic tree of proteomic distance map of the samples. However, the phylogenetic tree does not discriminate the case and control. The whole image clustering shows that the proteomic distance is more dependent to age and sex.

Key Words : Proteomics, 2D-Gel image, Clustering, Image registration

서 론

프로테옴 분석은 주로 생명체의 개체나 혹은 일부 조직의 단백질을 2차원 폴리 아크릴아마이드 젤 전기영동(2D-Gel)으로 분리하여 2차원 이미지를 만들고 각각의 단백질 스팟을 질량분석 장치로 분석하여 이미 알고 있는 단백질 서열 데이터베이스의 도움을 받아 단백질을 하나 하나 규명해 나가는 작업을 말한다(1). 특히 전사 프로필과 실제 단백질의 발현 프로필이 매우 다르며 동시에 복잡한 것으로 밝혀져 프로테옴 분석의 중요성이 더욱 강조되고 있다(2, 3). 최근 들어 질량분석기술의 발달과 다양한 프로테옴 분석기술의 개발로 프로테옴의 분석의 정확도와 속도가 가속화되고 자동화되어 2D-Gel상의 소량의 단백질 스팟까지 신속하게 분석되고 있다(4).

† Corresponding Author : School of Biotechnology and Bioengineering, Kangwon National University, Chunchon, Kangwon 200-701, Korea
Tel: +82-33-250-6276, Fax : +82-33-243-6350
E-mail : wonhur@kangwon.ac.kr

프로테옴 연구의 다른 한 부류는 프로테옴 비교분석 연구이다. 이것은 비교유전체학과 마찬가지로 정상적인 세포와 질환 혹은 결함 상태에 있는 세포의 프로테옴 2D-Gel 이미지를 비교하여 그 원인 단백질 혹은 질환이나 결함 상태와 관련이 있는 단백질을 찾아내는 것이다(5). 이미 병원성균과 비병원성균의 프로테옴(6, 7)의 비교분석 및 암세포와 정상세포의 프로테옴(8, 9)의 차이점을 성공적으로 분석되어 보고되고 있다. 비교 프로테옴 분석에 있어서 단 2장의 2D-Gel 이미지를 통해서도 control과 case의 차이를 명확하게 알 수 있는 경우도 있으나 대부분의 경우에는 많은 숫자의 프로테인 발현 패턴의 차이를 보여 의미 있는 다른 점 (meaningful difference)을 찾아내기가 용이하지 않다(10). 그러나 현재까지의 기술 수준에서 프로테옴 비교분석은 가장 강력한 연구방법으로 질병의 메커니즘을 분자수준에서 추적할 수 있으며, 의약품의 투여에 대한 단백질의 발현이 변화, 생물체의 분류 심지어는 배양이 어려운 미생물 컨소시움을 진단하는 다양한 분야에서 사용될 것으로 기대되고 있다(11, 12).

이외에도 프로테옴 현재까지 수집된 많은 임상 시료를 프로테옴 이미지로 전환시키고 이들을 데이터베이스로 구

성하여 데이터 마이닝을 이미지 자체를 비교분석하여 질환을 예측하는 방법이 제안되고 있다(13). 그러나 프로테옴 이미지에서 의미있는 차이점을 신뢰도 수준 이상으로 찾아내기 위해서는 시료의 숫자 즉 case의 숫자를 증가시켜야 하는데 case의 수를 조금 증가시켜 성공적으로 해결되기도 하지만 많은 량의 case의 수를 필요로 하는 경우가 대부분이다. 따라서 다량의 시료 샘플로부터 2D-Gel 이미지를 얻어야 프로테옴 비교분석을 통한 의미있는 데이터를 얻을 수 있는 가능성이 높아진다.

그러나 많은 수의 시료로부터 프로테옴 2D-Gel 이미지를 얻어도 이들을 서로 비교 분석하는 것은 매우 어려운 일이다. 그 이유는 시료의 양을 정확하게 조정하고 동일한 조건에서 electofocusing과 electrophoresis를 수행하고 발색과 탈색을 표준화시켜도 여전히 매 실험마다 동일한 시료의 2D-Gel 이미지도 상당한 변화를 보인다.

따라서 본 논문에서는 이미지를 처리하기 위하여 개발된 알고리즘을 사용하여 2D-Gel 이미지를 서로 비교할 수 있도록 정렬시키는 소프트웨어를 개발하고 정렬된 이미지들이 서로 얼마나 유사한지 비교하기 위하여 클러스터링할 수 있는 기능을 추가하였다. 이렇게 개발된 소프트웨어를 이용하여 인터넷상으로 공개되어 있는 Fetal Alcohol Syndrome의 2G-Gel 프로테옴 이미지(14)를 클러스터링하여 데이터 마이닝을 시도하였다.

재료 및 방법

소프트웨어 개발

2D-Gel 이미지를 처리하여 정렬시키고 서로 비교하는 소프트웨어는 Visual C++ Ver 6.0 (Microsoft, USA)을 사용하여 개발하였다. 이미지를 정렬하기 위하여 Multi-resolution and multi-level 알고리즘을 적용하였고 2차원 이미지 상관계수를 사용하여 이미지를 클러스터링하였고 Phylogenetic tree로 나타내었다. 개발된 소프트웨어는 펜티엄III 933 MHz, 256 MByte RAM의 개인용 컴퓨터와 마이크로 소프트 윈도우 2000 환경에서 테스트 하였으며 인터넷상의 http://hurwon.net의 자료실에 공개되어 있다.

2D-Gel 이미지

개발된 소프트웨어에 사용된 2D-Gel 이미지는 인터넷에서 일련의 시료에 대한 다수의 이미지를 제공하는 Flicker 사이트(14)에서 다운로드 받아 사용하였다. 클러스터링을 위해서는 10개의 Fetal Alcohol Syndrome의 case와 control의 프로테옴 2D-Gel 이미지를 사용하였다.

이미지 정렬

2개 이상의 2D-Gel 이미지를 비교하기 위해서는 먼저 기준이 될 2D-Gel 이미지를 결정하고 여기에 다른 이미지를 종축 및 횡축으로 변형시켜가면서 왜곡시켜 두 개의 이미지가 가장 같아지도록 한다. 두 개의 이미지의 유사성은 2차원 이미지 상관계수를 사용하여 수치화할 수 있다. 상관계수는 2개의 이미지의 각 좌표상의 색상의 강도를 각각 $I_1(x,y)$ 및 $I_2(x,y)$로 나타내었을 때 이 값의 곱의 표준편차를 각각의 이미지의 표준편차로 나눈 값으로 정의 된다. 따라서 두 이미지가 동일하다면 2차원 이미지의 상관계수는 1이 된다. 두 이미지간의 유사성이 낮을수록 상관계수는 감소한다, 상관계수 (corr)은 다음과 같다.

$$corr = \frac{\sigma_{image\ 1 \cdot image\ 2}}{\sigma_{image\ 1} \cdot \sigma_{image\ 2}} = \frac{n\sum_{(x,y)} I_1(x,y)\cdot I_2(x,y) - \sum_{(x,y)} I_1(x,y)\cdot \sum_{(x,y)} I_2(x,y)}{\sqrt{\left(n\sum_{(x,y)} I_1(x,y)^2 - \left(\sum_{(x,y)} I_1(x,y)\right)^2\right)}\cdot\sqrt{\left(n\sum_{(x,y)} I_2(x,y)^2 - \left(\sum_{(x,y)} I_2(x,y)\right)^2\right)}}$$

비교하려는 이미지가 심하게 왜곡되거나 비선형적으로 서로 다를 수 있다. 따라서 이미지를 N × N 개의 이미지 조각으로 나누어서 각각의 부분 이미지를 종축 및 횡축으로 줄이거나 늘여서 2개의 이미지를 비교하는 elastic image registration 법을 사용하였다. 그 중에서도 특히 이미지의 해상도를 낮추어 먼저 이미지를 정렬시키고 점차 원래의 이미지 상태로 해상도를 증가시키면서 계속 이미지를 정렬시키는 multiresolution image registration 알고리즘을 사용하였다(15). 기준 이미지에 대하여 다른 이미지들은 2 × 2, 4 × 4, 8 × 8 및 16 × 16으로 이미지를 분할시켜 정렬을 하고 그리드 정보만을 기록하여 보관하여 원시 이미지 정보를 가능한 변화시키지 않도록 하였다.

이미지 클러스터링

프로테옴간의 거리 즉 단백질의 발현 측면에서의 유사성은 2D-Gel 이미지의 유사성과 같다고 가정할 수 있다. 따라서 2D-Gel의 이미지를 비교하기 위하여 각각의 이미지 사이의 상관계수를 구하고 그 값이 큰 것 즉 1에 가까운 것 끼리 묶어내는 것이 바로 클러스터링이다. 먼저 유사성이 높은 이미지끼리 서로 묶는 클러스터링은 k-means clustering 알고리즘(16)을 사용하였다. 그리고 서로 유사성이 높은 이미지를 한꺼번에 묶어서 보여주는 방법은 이미지간의 거리를 상관계수로 환산하여 Phylogenetic tree(17)로 나타내는 것이다. 이를 위하여 n개의 이미지가 있으면 n(n-1)/2 번의 상관계수를 계산하여 표로 나타내고 이들 중 그 값이 1에 가까운 순서로 phylogenetic tree 형태로 표현하는 소프트웨어를 개발하였다.

결과 및 고찰

2D-Gel 이미지 분석 소프트웨어 개발

프로테옴 2D-Gel 이미지를 서로 비교하고 클러스터링하기 위한 전제 조건으로 먼저 2D-Gel 이미지를 읽고 처리하는 소프트웨어가 필요하다. 이를 위하여 이미지 파일의 확장자가 gif, bmp 및 tiff인 프로테옴 이미지 파일을 읽어서 회색조로 변화시키고 이미지상의 특정 위치에서의 값을 읽을 수 있고 이미지를 저장하거나 다시 불러오는 기능을 가진 소프트웨어를 개발하였다. 이렇게 개발된 소프트웨어가 Human breast ductorial carcinoma 세포의 프로테옴 2D-Gel 이미지를 읽은 상태에서 컴퓨터 화면을 캡처한 것을 Fig. 1에 나타내었다. 전기영동으로 2D-Gel을 제조하는 과정에서 사용된 분자량 마커에 해당하는 분자량 및 pI 값을 입력할 수 있는 기능을 추가하였으며 커서가 위치

한 부분을 확대하여 오른쪽에 따로 보여주고 단백질 스팟의 강도의 변화를 하단과 오른쪽에 동시에 볼 수 있도록 고안되었다. 동시에 오른쪽 하단에는 자동으로 단백질의 스팟을 찾아내어 이를 데이터베이스로 나타내는 기능을 추가하였다.

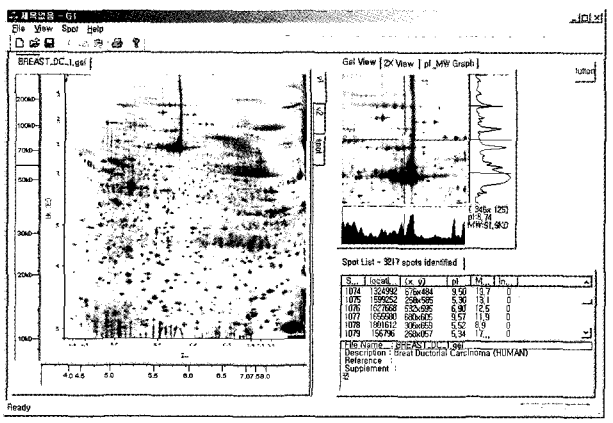

Figure 1. A screen capture of the software for 2D-Gel image processing.

2D-Gel 이미지를 읽어 이미지를 비교하거나 스팟을 자동으로 찾아내는 기능을 가진 소프트웨어로서는 공개 소프트웨어인 Flicker(18)를 비롯하여 상업적으로 판매되는 Melanie PDQuest 등이 있다. 이와 같이 상업적으로 판매되는 프로테옴 분석 소프트웨어에서 일반적으로 제공되는 자동 스팟 인식 기능, 이미지를 중첩시키고 이미지상 상이한 부분의 색을 다르게 보여 주는 기능 등도 추가하였다. 그러나 이미지의 분석을 위하여 다양한 기능을 제공하는 이러한 소프트웨어와 달리 본 연구에서 개발된 소프트웨어는 주로 2D-Gel 이미지의 정렬과 클러스터링을 소프트웨어적으로 구현하기 위한 중간 단계로 작성되었다.

다수의 이미지를 정렬시키기 위한 기능을 추가하였다. 먼저 기준이 되는 이미지를 읽고 이후 정렬할 다른 이미지들을 순차적으로 읽어 기준이미지에 대하여 여러 이미지를 순차적으로 정렬시키는 방법을 사용하였다. 그리고 이미지로부터 단백질 스팟을 먼저 찾아 내지 않고 원시 이미지를 사용하여 여러 단계의 해상도에서 보조 그리드 좌표를 활용하여 순차적으로 이미지를 정렬시키는 multi-resolution-multilevel algorithm을 활용하여 Visual C++로 메모리가 허용하는 한 다수의 이미지를 정렬시키는 기능을 추가하였다.

이후 2D-Gel 이미지를 Alignment를 통하여 하나의 기준 2D-Gel 이미지에 정렬된 다른 2차원 이미지 상관계수를 구하여 Phylogenetic tree를 생성시켜 직관적으로 프로테옴 이미지의 클러스터링을 할 수 있는 소프트웨어를 구현하였다.

2D-Gel 이미지 정렬

개발된 소프트웨어를 이용하여 서로 다른 2개의 Fatal Alcohol Syndrome (FAS)의 프로테옴 이미지를 읽고 정렬시켰다. Fig. 2의 왼쪽은 2개의 서로 다른 혈장 시료로부터 얻어진 1024 × 1024 pixel의 크기의 2개의 이미지를 중첩시킨 것이다. 그러나 전기영동 시간의 차이 및 isoelectro-focusing에서의 오차로 인하여 스팟들은 서로 중첩되지 않음을 관찰할 수 있다. 특히 혈장 시료에서 관찰되는 가장 큰 2개의 스팟은 혈청 내에서 가장 양이 많은 알부민과 면역글로브린이다. 이들은 pI값이 조금씩 다른 단백질의 스팟들이 수평으로 길게 연이어 보이는 것으로 서로 다른 곳에 위치하고 있는 것이 분명하게 관찰된다.

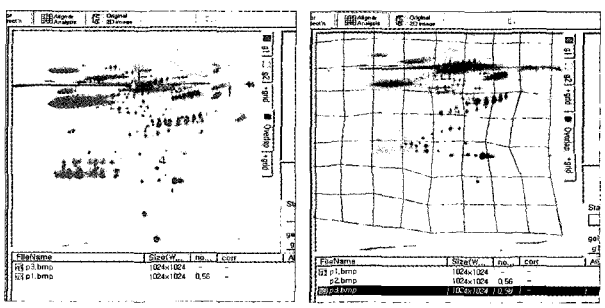

Figure 2. Alignment of 2D-Gel images showing a grid guiding local distortion.

이미지 정렬 기능을 사용하여 2개의 이미지를 정렬시켰다. 1024 × 1024 pixel 크기의 이미지를 정렬시키는데 약 30초의 시간이 소요되었다. 정렬 전에는 2차원 상관계수는 0.56이었으나 정렬 후에는 상관계수는 0.98로 증가하였다. Fig. 2의 오른쪽 이미지에서는 정렬된 2개의 이미지를 중첩시켜 보여 주고 있으며 서로 다른 곳에 위치하였던 알부민과 면역글로브린의 스팟들이 중첩하고 있음을 관찰할 수 있다. 동시에 정렬시키려는 이미지가 어떻게 변형되었는가를 보여주는 그리드가 같이 표시되어 있다. 그리드는 이미지를 8 × 8의 이미지 조각으로 나누는 경계점의 좌표를 연결한 것이다. 각각의 이미지 조각이 어떻게 변형되어 기준 이미지의 조각과 가장 잘 일치도록 좌표가 변형된 것을 나타낸다.

이와 같이 이미지를 비선형적으로 변형시켜 유사한 이미지를 중첩시키는 elastic image registration 기술에 바탕을 둔 프로테옴의 이미지를 정렬시키는 기술은 Phoretics사의 프로테옴 이미지 분석 소프트웨어인 Z3에 적용된 image registration 방법(19)과 본 연구에 적용된 알고리즘과 같은 multi-grid and multi-resolution registration 기술(15)이 개발되어 있다. Multi-resolution 방식의 image registration은 본 연구에서 사용된 것과 비슷한 크기의 이미지를 정렬시키는데 10초 이하의 시간이 소요된다고 보고하고 있다. 같은 알고리즘을 사용하였으나 프로그래밍상의 기술적인 문제로 계산 시간의 차이가 발생한 것으로 판단된다.

2D-Gel 이미지의 클러스터링

인터넷으로 공개된 10개의 Fetal Alcohol Syndrome (FAS)의 case와 control의 프로테옴 이미지를 전술한 방법으로 정렬시킨 후 클러스터링을 시도하였다(Fig. 3). 각각의 프로테옴 이미지간의 계산된 2차원 상관계수는 최고 0.94에서 최저 0.80의 값을 나타내었다. 이 값들을 기준으로 각각의 프로테옴 이미지를 phylogenetic tree로 나타내어졌다.

상관계수 0.94의 값을 가지는 가장 유사한 프로테옴 이미지는 9세 여아의 FAS의 case와 9세 여아의 control의 것으로 나타났다. 다음으로 유사한 이미지는 10세 여아의 FAS의 case의 것으로 나타났다. 트리의 아랫부분에 클러스터링된 이미지중 유사한 것은 10세 11세의 남아의 프로테옴 이미지이고 그 다음으로 FAS의 case/control과는 무관하게 남아의 이미지가 유사한 것으로 클러스터 되었다. 따라서 FAS 프로테옴 이미지의 클러스터링 결과는 case와 control로 구분되기 보다는 나이 및 성별로 클러스터링 되는 경향을 보였으나 나이 및 성별도 분명한 기준은 아닌 것으로 나타났다. 그 이유는 FAS의 경우 2D-Gel 이미지상의 소수의 단백질의 발현의 변화가 있을 수 있으나 성별 혹은 나이에 따른 단백질의 발현의 차이가 더 뚜렷하기 때문인 것으로 판단된다. 결론적으로 2D-Gel 프로테옴 전체의 이미지를 비교하여 유사한 정도에 따라 모으는 클러스터링은 FAS 시료의 경우 case와 control보다는 시료원의 외연적인 특징인 나이 혹은 성별에 더 의하여 의존하는 것으로 판단할 수 있다.

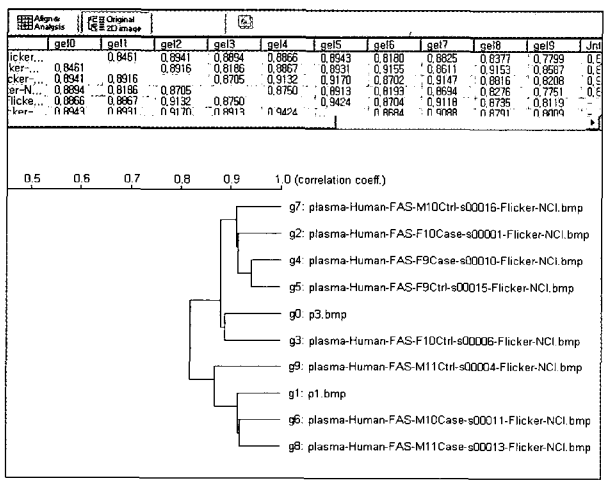

Figure 3. A phylogenetic tree showing the proteomic distance between the plasma samples of the Fetal Alcohol Syndrome cases and controls.

요 약

2D-Gel 이미지간의 유사성을 기준으로 생물학적인 시료가 프로테옴 수준에서 유사성의 정도와 서로 다른 단백질 스팟을 파악해 낼 수 있다. 그러나 생물학적인 시료는 개체간 변화가 크고 2차원 전기영동장치의 재현성의 한계로 인하여 비교가 어려운 경우가 많고 의미 없는 차이점만 발견되는 경우 또한 비일비재하다. 이를 극복하기 위해서는 프로테옴 이미지간의 정렬을 통하여 정확한 비교가 가능하게 하여야한다. 본 연구에서는 이미지상의 단백질 스팟을 일일이 찾지 않고 여러 개의 원시 이미지를 동시에 정렬시키는 multiresolution-multilevel algorithm을 활용하여 소프트웨어를 개발하였다. 또 이렇게 정렬된 이미지들이 서로 얼마나 유사한지 보여주는 Phylogenetic tree를 자동으로 생성시키는 소프트웨어를 개발하였다. 이 방법을 이용하여 Fetal Alcohol Syndrome의 case와 control의 10개의 프로테옴 이미지에 대하여 클러스터링을 시도하였다. 이와 같이 2D-Gel 프로테옴 전체의 이미지를 비교하여 유사한 정도에 따라 모으는 클러스터링은 FAS 시료의 경우 case와 control 보다는 시료원의 외연적인 특징인 나이 혹은 성별에 더 의하여 의존하는 것으로 나타났다.

감 사

이 논문은 2001년도 강원대학교 기성회교수 국외파견연구지원에 의하여 연구되었습니다.

REFERENCES

1. Jensen, O. N., M. R. Larsen, and P. Roepstorff (1998) Mass spectrometric identification and microcharacterization of proteins from electrophoretic gels: Strategies and applications, *Proteins*, **Supplement 2**, 74-89.
2. Lopez, M. F. (2000), Better approaches to finding the needle in a haystack: Optimizing Proteome analysis through automation, *Electrophoresis* **21,** 1082-1093.
3. Haynes, P. A. and J. R. Yates (2000), Proteome profiling - pitfalls and progress, *Yeast* **17,** 81-87.
4. Harry, J. L., M. R. Wilkins, B. R. Herbert, N. H. Packer, A. A. Gooley, and K. L. Williams (2000), Proteomics: Capacity versus utility, *Electrophoresis* **21,** 1071-1081.
5. Michener, C. M., A. M. Ardekani, E. F. 3rd Petricoin, L. A. Liotta and E. C. Kohn (2002), Genomics and proteomics: application of novel technology to early detection and prevention of cancer, *Cancer Detect Prev.* **26,** 249-55.
6. Cordwell, S. J., A. S. Nouwens, and B. J. Walsh (2001), Comparative proteomics of bacterial pathogens, *Proteomics* **1,** 461-72.
7. Jungblut, P. R., D. Bumann, G. Haas, U. Zimny-Arndt, P. Holland, S. Lamer, F. Siejak, A. Aebischer, and T. F. Meyer (2000), Comparative proteome analysis of *Helicobacter pylori*, *Molecular Microbiology* **36,** 710-725.
8. Celis, J. E., M. Ostergaard, H. H. Rasmussen, P. Gromov, I. Gromova, H. Varmark, H. Palsdottir, N. Magnusson, I. Andersen, B. Basse, J. B. Lauridsen, G. Ratz, H. Wolf, T. F. Orntoft, P. Celis, and A. Celis (1999), A comprehensive protein resource for the study of bladder cancer, *Electrophoresis* **20,** 300-309.
9. Tomlinson, A. J., M. Hincapie, G. E. Morris, and R. M. Chicz (2002), Global proteome analysis of a human gastric carcinoma, *Electrophoresis* **23,** 3233-3240.
10. Humpherysmith, I., S. J. Cordwell, and W. P. Blackstock (1997), Proteome research - Complementarity and limitations with respect to the RNA and DNA worlds, *Electrophoresis* **18,** 304-318.
11. Hanash, S. M. and D. Teichroew (1988), Mining the human proteome - Eeperience with the human lymphoid protein database, *Electrophoresis* **19,** 301-309.
12. Haynes, P. A., S. P. Gygi, D. Figeys, and R. Aebersold (1999), Proteome analysis - Biological assay or data archive, *Electrophoresis* **19,** 1403-1421.
13. Veenstra, T. D. and T. P. Conrads (2003), Serum protein fingerprinting, *Curr. Opin. Mol. Ther.* **5,** 584-93.
14. Robinson M. K., J. E. Myrick, L. O. Henderson, C. D. Coles, M. K. Powell, G. A. Orr, and P. F. Lemkin (1995), Two-dimensional protein electrophoresis and multiple hypothesis testing to detect potential serum protein biomarkers in children with fetal alcohol syndrome, *Electrophoresis* **16,** 1176-1183.

15. Veeser, S., M. J. Dunn, and G. Z. Yang (2001), Multiresolution image registration for two-dimensional gel electrophoresis, *Proteomics* **1**, 856-870.
16. Han, J. and M. Kamber (2001), Data Mining: Concepts and Techniques, p314, Academic Press, San Diego.
17. Everitt, B. S., S. Landau and M. Leese (2001), Cluster Analysis, p11, Oxford University Press, New York.
18. Andrew, L. (1997), A book in methods in molecular biology, Vol. 112, Humana Press, Totowa, NJ, pp 339-410.
19. Smilansky, Z. (2001), Automatic registration of images of two-dimensional protein gels, *Electrophoresis* **22**, 1616-1626.