Proceedings of KFIS Fall Conference, 2003. 9. 27

암진단을 위한 2차원 단백질 전기영동 젤 해석

Analysis of 2D Electrophoresis For Cancer Classification

김재민
홍익대학교 전자전기공학부

Jaemin Kim
School of Electronic and Engineering, Hongik University
E-mail : jaemin@hongik.ac.kr

ABSTRACT

유전자에 대한 정보를 획득하는 기술적인 문제가 해결되면서, 질병 진단을 위한 새로운 접근 방법으로 혈액 속에 있는 모든 단백질(proteome)의 구성을 분석하는 프로테오믹스(proteomics)에 대한 연구가 최근 들어 활발하게 진행되고 있다. 본 논문은 암 진단을 위하여 혈액 중의 단백질의 구성을 측정한 2차원 전기영동 (2D electrophoresis) 젤 데이터를 해석하는 새로운 방법을 제시하였다. 우선 측정된 많은 단백질 스팟(spot) 중에서 T-statistics 방법으로 단백질 스팟들을 선택하였다. 선택된 단백질 스팟들로 이루어진 암 환자와 정상인 두 샘플들의 확률 분포를 각 집단에 따로 적용된 PCA 영역에서 계산하였다. 최종적으로 조건부 확률의 차이에 근거한 베이즈 분류(Bayes classification) 이론을 적용하여 암 진단을 하였다.

Key words : 암진단, 2차원 전기영동 젤, 주성분 분석, 베이스 분류, T-statistics

I. 서 론

사람의 각각의 세포는 동일한 게놈을 가지고 있으나, 단백질의 구성(proteome)은 사람에 따라 다르다. 각 세포의 프로테움은 시간과 사람의 몸 상태에 따라 변화하는 특성을 가지고 있다. 노화, 명상, 운동, 식이요법 등도 프로테움을 변화시킨다. 또한 특히 특정 질병에 감염되었을 때, 질병에 대항하기 위하여 프로테움이 변화하는 특성을 가지고 있다. 이러한 프로테움에 대하여 연구하는 프로테오믹스 (proteomics)는 유전자에 대한 정보를 획득하는 기술적인 문제가 해결되면서 활발하게 진행되고 있으며, 최근에는 질병 진단을 위한 새로운 접근 방법으로 연구되고 있다 [1].

프로테움을 측정하는 방법으로는 2차원 전기영동(2D electrophoresis)이 널리 사용되고 있다. 2차원 전기영동은 모든 단백질은 고유의 등전점(iso-electric point)과 질량을 가지고 있음을 이용하여 2차원 평면 상의 특징 위치에 특정 단백질에 놓여지도록 하고 있다. 2차원 전기영동을 통하여 2차원 평면 상에 놓여진 각각의 단백질 스팟(spot)은 silver 혹은 fluorscent 방법으로 염색을 하여 정성적 혹은 정량적인 분석을 하게 된다. 염색 방법의 발전과 염색된 단백질 스팟의 측정하는 방법의 발전으로부터 정량적인 값을 얻는 방법이 꾸준히 개선되고 있으나, 현재 가장 정확한 방법은 질량 분광 식별(mass spectrometric identification)으로 알려져 있다. 질량 분광 식별은 많은 처리 과정과 처리 시간을 필요로 하고 있다.

Wu[2] 등은 질량 분광식별로 얻은 데이터로부터 난소암을 진단하기 위하여 다양한 통계적인 방법을 적용하였다. 우선 T-statistics와 random forest [3] 방법을 사용하여 많은 스팟으로부터 15~25개의 주요 스팟(key spot)을 선택하였다. 다음에는 LDA (linear discriminant analysis), QDA (quadratic discri-

minat analysis), K-NN (k-nearest neighbor), B&B (Bagging and Boosting), SVM (support vector machine)을 적용하였다 [4,5].

본 논문은 질량 분광 식별 법을 사용하지 않고, PD-Quest라는 소프트웨어를 이용하여 2D 젤에서 측정한 스팟 데이터를 해석하여 폐암을 진단하는 방법을 설명한다. 주요 스팟의 선택은 T-statistics 방법을 사용하였고, 분류 방법으로는 얼굴인식 등에서 성공적으로 사용된 확률에 의한 주성분 분석(probabilistic PCA) 방법 [6,7]을 적용하였다. 실험에서는 제안된 방법과 SVM, K-NN 방법을 비교 분석하였다.

II. 본 론

2.1 PCA 영역에서 확률분포

N개의 측정 값으로 이루어진 M 개의 샘플 $\mathbf{x} = [x_1, \Lambda, x_N]$은 폐암 환자와 정상인 두 집단 $\{\Omega_1, \Omega_2\}$으로 구성되어 있다. 두 집단을 분류하기 위하여 각 샘플 집단의 확률 분포를 학습하고, 분류할 샘플이 두 집단에서 발생할 확률의 크기를 비교하여 샘플을 분류하고자 한다. 각 샘플은 큰 상관관계를 가지고 있기 때문에 정확한 확률 분포를 구하기 위해서는 각 집단에서 샘플간의 상관 행렬(correlation matrix)을 구하여야 한다. 문제점은 각 집단의 샘플 수가 충분하지 않을 경우에는 상관 행렬이 불량조건(ill-condition) 특성을 가지고 있어 정확한 확률을 구하기 어렵다. 이러한 문제점을 해결하기 위하여 각 집단에 PCA를 적용하여 기저 벡터(basis vector)를 형성하고, 기저 벡터상에서 확률 분포를 구한다. 각 집단에 소속된 샘플의 평균값이 $\{\boldsymbol{\mu}_1, \boldsymbol{\mu}_2\}$일 때, 기저벡터는 다음과 같이 형성한다.

$$\boldsymbol{\Sigma} = \tilde{\mathbf{X}}\tilde{\mathbf{X}},\ \tilde{\mathbf{X}} = [\mathbf{x}_1 - \boldsymbol{\mu}_i \quad \mathrm{K} \quad \mathbf{x}_{Mi} - \boldsymbol{\mu}_i],\ i = 1,2, \tag{1}$$

$$\boldsymbol{\Sigma}\mathbf{v}_k = \lambda_k \mathbf{v}_k. \tag{2}$$

위 식에서 $\boldsymbol{\Sigma}$는 상관 행렬이고, λ_k와 $\mathbf{v}_k$는 각각 고유 값과 고유 벡터이다. $\{\mathbf{v}_k\}$는 기저 벡터를 형성한다. 상위 F 개의 기저 벡터만을 사용하여 작은 오차를 가지고 모든 샘플 $\mathbf{x}$를 표현할 수 있기 때문에, 기저 벡터의 수 F는 두 집단의 샘플 수 M_1과 M_2 보다 작은 값을 선택한다.

샘플 $\mathbf{x}$는 집단 $\{\Omega_i\}$에 소속된다는 가정 하에 기저 벡터 $\{\mathbf{v}_k\}$의 선형조합으로 아래와 같이 표현된다.

$$\mathbf{x} - \boldsymbol{\mu}_i = \sum_{k=1}^{F} \alpha_k \mathbf{v}_k,\ \ \alpha_k = \langle \mathbf{x} - \boldsymbol{\mu}_i, \mathbf{v}_k \rangle. \tag{3}$$

위 식에서 $\langle \cdot, \cdot \rangle$는 내적(inner product)을 나타낸다. α_k는 평균값이 0이고, 분산 값이 λ_k이며, 임의의 두 α_k, α_l은 상호 독립이다. α_k가 정규분포를 이룬다는 가정 하에 샘플 $\mathbf{x}$의 조건부확률은 다음과 같이 표현된다.

$$P(\mathbf{x} \mid \Omega_i) = \prod_{k=1}^{F} \frac{1}{\sqrt{2\pi\lambda_k}} \exp\left(-\frac{\alpha_k^2}{2\lambda_k}\right). \tag{4}$$

고유값 λ_k는 k 값이 증가할수록 0에 가까운 값을 가지게 되고, 샘플의 수가 적은 경우 λ_k의 작은 오차는 조건부 확률 값에 많은 영향을 미치게 된다. 이러한 영향을 줄이기 위하여 λ_k에 작은 값 ε을 더하여 수치계산 오차로 인한 영향을 줄였다.

2.2 베이즈 분류(Bayes Classification)

최대 우도(maximum likelihood) 방법에 의한 샘플 $\mathbf{x}$의 분류는 다음과 같이 표현된다.

$$\mathbf{x} \in \Omega_1 \ \text{ if } \ \log P(\mathbf{x} \mid \Omega_1) - \log P(\mathbf{x} \mid \Omega_2) > \delta, \tag{5}$$
$$\mathbf{x} \in \Omega_2 \ \text{ otherwise.}$$

암 환자를 정상인으로 분류하는 penalty는 정상인을 암 환자로 분류하는 penalty 보다 크기 때문에 δ 값을 사용하였다. $\log P(\mathbf{x} \mid \Omega_i)$는 다음과 같이 표시된다.

$$\log P(\mathbf{x} \mid \Omega_i) = C - \frac{1}{2}\log\lambda_k - \frac{1}{2}\frac{\alpha_k^2}{2\lambda_k}. \tag{6}$$

III. 실험

3.1 실험 데이터

46명의 암 환자와 49명의 정상인에서 채취한 혈액으로부터 구한 2차원 전기영동 젤을 형성하고, 2차원 젤 상의 각 스팟들의 값을 정량적으로 구한다. 분류하고자 하는 샘플들은 스팟들의 값으로 이루어져 있다. 그림 1은 생성된 2차원 젤을 silver로 염색한 영상을 보여주고 있다.

3.2 전처리

특정 위치의 스팟은 특정 단백질을 나타내나, 해상도가 낮아 몇 개의 변형된 단백질들이 겹쳐 표시되기도 한다. 본 연구에서는 특정 위치에서 주변과 분리되어 같은 밝기(bright intensity)로 표시되는 스팟은 하나의 단백질로 간주한다. 각 스팟의 정량적인 값은 스팟의 면적과 스팟의 밝기를 곱하여 계산한다. 이러한 과정은 PD-Quest 소프트웨어를 사용하여 처리하였다. 모든 샘플에서 각 스팟들의 값을 구한 후, 각 스팟들의 분산 값이 1이 되도록 정규화 하였다.

3.3 주요 스팟의 선택

정규화된 각 스팟들 중에서 Student T-test 방법을 이용하여 두 집단의 평균값의 차이가 큰 29개의 스팟을 선택하였다.

3.4 암환자 및 정상인 샘플 집단의 특성

두 샘플 집단을 PCA로 분석하였을 때 특성은 다음과 같다.

- 각 집단의 첫번째 고유 값은 두번째 고유 값에 비하여 약 4배정도 크다. 이는 각 샘플들 사이의 상관관계가 작지 않음을 보여주고 있다.
- 암 환자 집단의 고유 값은 정상인 집단의 고유 값에 비하여 약 2배정도 큰 값을 가지고있다. 이는 암 환자들의 프로테옴은 정상인과 다름을 나타내고 있다.
- 암 환자 집단의 첫번째 고유벡터와 정상인 집단의 첫번째 고유벡터 사이의 각은 약 65도를 이루고 있다. 이는 암이 발생시 특정 단백질은 변화정도가 다름을 보여주고 있다.

그림 2는 이러한 특성은 시각적으로 보여주고 있다.

3.4 각 샘플들의 조건부 확률 차이

그림 3은 각 샘플에서 구한 조건부 로그 확률의 차 $\log P(\mathbf{x}|\Omega_1) - \log P(\mathbf{x}|\Omega_2)$를 보여 주고 있다. 위 줄부터 좌에서 우로 46개의 샘플은 암 환자 샘플을 나타내고 있고, 아래의 49개 샘플은 정상인 샘플을 나타내고 있다. 암 환자의 경우 모두 양수 값을 가지고 있으며, 정상인의 경우 모두 매우 큰 음수 값을 가지고 있다. 두 집단은 매우 큰 마진을 가지면서 분리됨을 보여주고 있다.

3.5 다른 분류 방법과의 비교 및 분석

폐암진단을 위하여 본 논문에서 제안한 방법과 암 진단을 위하여 기존 논문에서 사용한 SVM, K-NN 방법과 비교하였다. 실험 방법은 암 환자 집단과 정상인 집단에서 난수를 발생시켜 각각 35개의 샘플을 선택하고, 선택된 샘플을 이용하여 분류기(classifier)를 학습시켰다. 학습된 분류기는 95개의 모든 샘플을 분류하도록 시도하였다. 100만번의 반복실험을 통하여 3가지 분류 방법을 비교하였다. 제안된 방법은 100% 성공률을 가지면서 정확하게 모든 샘플을 분류하였다. 반면에 K-NN과 SVM은 97.8와 98.5의 성공률을 보여주고 있다.

그림 3에서 옅은 색으로 표시된 부분은 SVM에서 support vector들로 선택된 샘플들을 보여주고 있으며, 밑줄 친 값은 SVM에서 잘 못 분류하고 있는 샘플들을 보여 주고 있다.

그림 1. 전기영동으로 생성한 2차원 젤

특정 위치의 스팟들은 특정 단백질을 나타낸다. 상단 중단 부근에 좌우로 넓게 엷게 퍼진 검정색은 혈액에 포함된 알부민을 나타낸다.

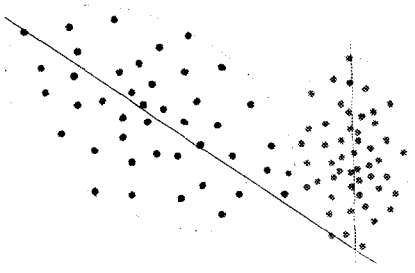

그림 2. 암 환자와 정상인 샘플의 분포

두 집단에서 분산이 가장 큰 방향 벡터를 각각 구하고, 두 방향 벡터로 이루어진 평면에 투영한 두 집단의 샘플들의 분포

0	1	2	3	4	5	6	7	8
177.8	182.2	89.8	1342.5	115.9	165.0	325.2	320.7	273.5
343.9	284.4	39.1	269.2	188.5	217.2	688.1	230.3	499.4
757.6	1158.4	182.7	570.4	188.7	190.7	515.3	651.5	245.0
199.1	136.4	747.5	334.9	268.9	323.1	267.3	40.1	32.8
1495.6	618.9	62.3	402.5	365.8	162.9			
						-309.8	-271.7	-151.3
-249.4	-275.7	-246.7	-407.3	-300.9	-301.0	-278.5	-306.1	-263.4
-276.7	-181.2	-233.5	-272.7	-372.7	-302.3	-253.0	-302.0	-255.4
-359.0	-379.0	-197.3	-275.5	-289.7	-402.0	-478.9	-265.9	-275.8
-411.6	-302.1	-173.0	-209.0	-359.5	-279.4	-262.8	-236.7	-247.0
-213.7	-182.6	-195.3	-205.5	-190.6				

그림 3. 각 샘플의 조건부 로그 확률 차

위에서 아래로, 왼쪽에서 오른쪽으로 상위 46개는 암 환자의 조건부 로그 확률 차를 나타내며, 하위 49개는 정상인 조건부 로그 확률 차를 나타내고 있다. 엷은 회색으로 표시한 값은 SVM의 support vector로 선택된 샘플들을 나타내며, 밑줄 친 값은 SVM에서 잘못 인식된 샘플들을 나타낸다.

III. 결 론

본 논문은 암 진단을 위하여 혈액 중의 단백질의 구성을 측정한 2차원 전기영동 (2D electrophoresis) 젤 데이터를 해석하는 새로운 방법을 제시하였다. 우선 측정된 많은 단백질 스팟(spot) 중에서 T-statistics 방법으로 단백질 스팟들을 선택하였다. 선택된 단백질 스팟들로 이루어진 암 환자와 정상인 두 샘플들의 확률 분포를 각 집단에 따로 적용된 PCA 영역에서 계산하였다. 최종적으로 조건부 확률의 차이에 근거한 베이즈 분류(Bayes classification) 이론을 적용하여 암 진단을 하였다. 제안된 방법은 SVM과 K-NN에 비하여 폐암진단에 효과적임을 실험을 통하여 입증하였다.

감사의 글 : 본 연구는 ㈜바이오인프라 지원으로 수행되었습니다.

IV. 참고문헌

[1] A. M. Campbell and L. J. Heyer, Discovering Genomics, Proteomics, and Bioinformatics, Addison-Wesley, 2003.

[2] B. Wu, T. Abbott, D. Fishman, W. McMurray, G. Mor, K. Stone, D. Ward, K. Williams, H. Zhao, "Comparison of statistical methods for classification of ovarian cancer using a proteomics dataset", *Bioinformatics, in press.*

[3] L. Breiman, randomforest, Technical Report, Statistics Dept. UCB, 2001.

[4] B. D. Ripley, *Pattern recognition and neural network*, Cambridge University Press, 1996.

[5] C. J. C. Burges, "A tutorial on support vector machines for pattern recognition", Data Mininng and Knowledge Discovery, Vol. 2, No. 2, 121-167, 1998.

[6] B. Moghaddam, "Principal Manifold and Probabilistic Subsapces for Visual Recognition", IEEE Trans. on PAMI, Vol. 24, No. 6, June 2002.

[7] B. Moghaddam, and A. Pentland "Probabilistic Visual Learning for Objct Representation", IEEE Trans. on PAMI, Vol. 19, No. 7, July 1997.