전산생물학을 이용한 마이크로어레이의 유전자 발현 데이터 분석 및 유형 분류 기법

Analysis and Subclass Classification of Microarray Gene Expression Data Using Computational Biology

유 창 규*, 이 민 영, 김 영 황, 이 인 범
(ChangKyoo Yoo, Min-Young Lee, YoungHwang Kim, and In-Beum Lee)

Abstract : Application of microarray technologies which monitor simultaneously the expression pattern of thousands of individual genes in different biological systems results in a tremendous increase of the amount of available gene expression data and have provided new insights into gene expression during drug development, within disease processes, and across species. There is a great need of data mining methods allowing straightforward interpretation, visualization and analysis of the relevant information contained in gene expression profiles. Specially, classifying biological samples into known classes or phenotypes is an important practical application for microarray gene expression profiles. Gene expression profiles obtained from tissue samples of patients thus allow cancer classification. In this research, molecular classification of microarray gene expression data is applied for multi-class cancer using computational biology such gene selection, principal component analysis and fuzzy clustering. The proposed method was applied to microarray data from leukemia patients; specifically, it was used to interpret the gene expression pattern and analyze the leukemia subtype whose expression profiles correlated with four cases of acute leukemia gene expression. A basic understanding of the microarray data analysis is also introduced.

Keywords : bioinformatics, computational biology, fuzzy clustering, gene expression analysis, leukemia, microarray, molecular biology, multi-class cancer

I. 서론

최근 DNA칩(마이크로어레이) 과 같은 고효율 어레이(high-throughput array) 기술의 개발로 방대한 양의 데이터가 생산되어 생명현상을 게놈 수준에서 해석할 수 있게 되었다. 수천에서 수만 개에 이르는 유전자의 발현 정보를 이용하여 특정 질병에 관여하는 유전자를 찾아내거나 질병의 유형을 분류하는 등, 유용한 정보를 추출하기 위한 연구가 진행되었으며, 이러한 연구는 전산생물학(computational biology) 또는 생명정보학(bioinformatics)이라는 새로운 학문 분야를 탄생시켰다. 데이터 분석을 근간으로 한 생명정보학은 생명현상의 생리학적, 병리학적 메커니즘을 밝혀내고 이해하는데 큰 기여를 하고 있다. 특히 암의 진행, 전이와 관련된 메커니즘이나 신약 후보물질을 찾아내는 일에서 실험적으로는 불가능한 부분이 마이크로어레이를 이용한 연구로 규명되어 왔다. 이에 따라 유전자 칩으로부터 얻어지는 정량적인 생체 유전자정보를 고도의 데이터 처리 기법으로 분석하여 특정 암 (혹은 암의 유형)과 관련성이 유의한 유전자들을 찾아내고, 선택된 유전자를 이용하여 암의 진단 및 유형을 분류하는 연구가 활발히 진행되고 있다. 분석된 유전자 정보와 환자의 분류 결과는 추후 각종 암에 핵심적으로 작용하는 표적유전자(drug target)의 확보에 유용한 정보를 제공하여 항암제 및 항암치료 요법의 개발에 기여할 수 있다. 이러한 생명정보학은 예방, 진단, 처방, 신약 개발 등 전반적인 의료 제약산업에 반드시 필요한 핵심 기술로 자리잡게 될 것이다[1-8].

마이크로어레이 기술에 기반한 유전자 발현 패턴 분석은 전형적인 데이터 분석의 범위를 벗어난다는 점에서 특이성을 갖는다. 일반적인 데이터가 변수의 개수보다 샘플의 개수가 많은데 비해서 유전자 발현 데이터는 변수의 개수, 즉 유전자의 개수가 수천에서 수만에 이르고, 그에 반해 샘플의 개수, 즉 환자 수는 수십 명을 넘지 않는 고차원 데이터(high-dimensional data)로 알려져 왔다. 이러한 고차원 데이터는 curse of dimensionality, overfitting, unstable inverse같은 문제들 때문에 일반적인 데이터 분석 기법으로는 신뢰할만한 분석 결과를 얻을 수 없다. 따라서 유전자 발현 패턴을 분석하기 위한 새로운 분석 알고리즘을 개발해야 한다[1-4,9-20].

현재까지 고차원 데이타의문제점을 해결하기 위하여 분석하는 마이크로어레이의 유전자의 갯수를 줄이는 방법이 가장 타당하다고 알려져 있다. 일반적으로 유전자 발현 데이터 분석은 마이크로어레이에 사용되는 수천, 수만 개의 유전자 중에서 특정 질병과 관련되어 이상 발현 양상을 보이는 유전자는 소수에 불과하고, 대부분의 유전자(housekeeping gene)는 거의 일정한 발현 양상을 보인다는 사실로부터 출발하므로 이러한 housekeeping gene을 제거하는 방법이 관건이 된다. 또한 데이터 분석에 있어서 변수의 개수가 적을수록 일반화와 예측에 유리하므로 housekeeping gene의 제거뿐만 아니라 특정 발현 패턴과 관련이 높은 소수의 유전자를 선택하는 일이 중요하게 된다[11-15,21-22].

본 논문에서는 마이크로어레이의 데이터 분석에 관한 기본적인 개념을 소개하고 전산생물학 (생명정보학)을 이용한 백혈병 환자의 마이크로어레이 데이터로부터 특이한 발현 양상을 보이는 유전자를 선택하는 방법과 환자의 유형을 분

* 책임저자(Corresponding Author)
논문접수 : 2005. 8. 25., 채택확정 : 2005. 9. 12.
유창규, 이민영, 김영황, 이인범 : 포항공과대학교 화학공학과
(ckyoo@postech.ac.kr/orange@postech.ac.kr/ograng@postech.ac.kr/iblee@postech.ac.kr)
※ 본 연구는 Brain Korea 21프로젝트의 지원을 받아 수행되었음.

류하는 방법, 특히 두 개 이상의 카테고리에 적용할 수 있는 다유형 암분류 방법을 제시한다.

II. 본론

1. 전산생물학을 이용한 마이크로어레이 유전자 발현 데이터 분석

마이크로어레이 유전자 발현 데이터의 분석방법은 크게 3가지로 나눌 수 있다[4,6] (i) 유형 발견, 예를 들어 유전자 발현 데이터로부터 백혈병의 유형을 찾아내는 것; (ii) 유형 예측, 분류되지 않은 샘플이 어느 유형에 속하는지 예측하는 것; (iii) 발현 양상이 다른 유전자 검출. 위와 같은 데이터 분석에는 다양한 기계학습법과 통계적 방법이 사용된다. 기본적으로 마이크로어레이 데이터 분석에는 (i) 정상샘플과 암 샘플을 병리학적으로 잘못 분류해서는 안되며 (ii) 통계학적 분석방법에 대하여 적절한 학습데이터를 사용하여야 하고, (iii) 각 유형마다 충분한 수의 샘플을 정확하게 측정하는 것이 요구된다[4]. 또한 모든 문제에서 (i) 알맞은 분석방법을 선택하는 것; (ii) 다양한 샘플로 테스트 하는 것; (iii) 오버피팅(overfitting)를 피하는 것; (iv) 정확히 예측하는 것; (v) 유형 예측을 하고자 하는 샘플이 존재하는 분류의 유형 중의 하나에 속해야 된다는 것을 고려해야 된다.

마이크로어레이 분석 방법은 크게 감독학습법(supervised learning)과 비감독학습법(unsupervised learning)으로 나눌 수 있다. 비감독학습법은 유전자 발현 데이터 외에 다른 정보는 사용 하지 않는다. 따라서 비감독학습법은 이전에 발견하지 못한 새로운 정보를 추출할 수 있으며, 결과는 다른 정보에 의해 편향되지 않는다. 반면, 감독학습법은 유전자, 샘플, 질병 등에 대한 추가적인 정보를 사용한다[2].

그림 1은 마이크로어레이 데이터 분석에서 비감독학습에 의한 유형 발견(class discovery)과 감독학습에 의한 유형 분류(class distinction)를 나타낸다. (a) 비감독학습법은 유전자 발현 양상이 비슷한 샘플들을 클러스터링 하는 것이다. 클러스터링의 결과를 시각화하기 위해 heat maps 을 이용한다. 유전자 발현 수준이 높으면 붉은색, 낮으면 녹색으로 표시하기 때문에 유형간의 유전자 발현 차이를 명확히 보여준다. 각 샘플을 다차원 스케일링(multidimensional scaling)이나 주성분 분석을 이용해 각 유전자를 3차원 공간상에 점으로 내어 볼 수도 있다. (b) 감독학습법은 각 유형에 속하는 샘플을 학습데이터로 사용해 분류기를 구성하고 이를 기존에 알려지지 않은 샘플의 유형을 예측하는데 사용한다.

2. 최적 유전자 선택 및 퍼지 클러스터링

마이크로어레이 데이터 분석에서 가장 먼저 수행하여야 하는 것이 주요 유전자 선택이다. 본 논 문지에서는 DPLS (Discriminant Partial Least Square)를 유전자 선별 방법으로 사용하였다[10,14,15]. DPLS는 독립변수 X와 종속변수 Y 사이의 공분산 (covariance)을 최대로 만들며 차원을 감소시키는 방법이다. 마이크로어레이 데이터분석에서 DPLS의 의존변수는 암 종류이고 독립변수는 유전자로, 유전자와 암 종류 사이의 공분산을 최대로 만드는 모델을 수립한다.

$$\mathbf{w_k} = \arg\max \ \mathrm{Cov}(\mathbf{Xw}, \mathbf{y}) \tag{1}$$

$$\mathbf{w}'\mathbf{S}\mathbf{w}_j = 0 \text{ for all } 1 \le j \le k \tag{2}$$

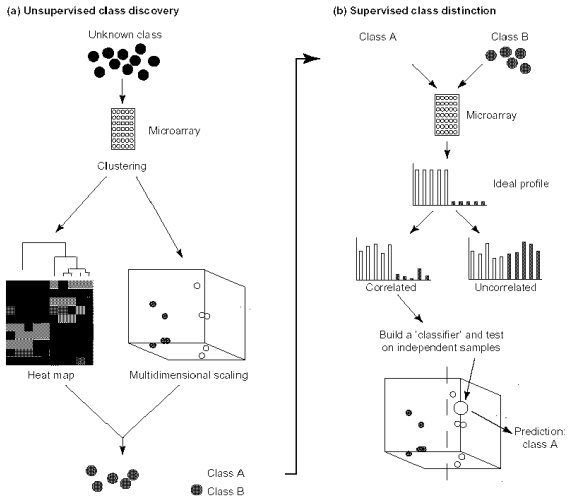


그림 1. 마이크로어레이 데이터 분석에서 비감독 유형 발견과 감독 유형 분류[1].

Fig. 1. Unsupervised class discovery and supervised class distinction in the microarray data analysis .

여기서 $\mathbf{S}' = \mathbf{X}'\mathbf{X}$는 공분산이고 i번째 PLS성분은 원래 유전자들의 $(\mathbf{Xw}_i)$ 선형조합으로 이루어진다. DPLS의 최대 장점은 유전자의 선형조합으로 이루어진 스코어(latent variables)의 직교성에 있다. DPLS모델링을 이용하여 직교화된 스코어 값은 유전자사이의 관계를 독립으로 만들기 때문에 유전자간의 상관관계를 더 이상 고려할 필요가 없게 된다. 이 DPLS모델링 정보는 유용한 정보를 지닌 유전자를 선별하기 위해 사용되었다. 여기서는 종속변수(암 분류)에 대한 모든 독립변수(유전자)의 영향을 계량화한 변수인 변수 가중치(Variable Importance in the Projection, VIP)값이 유전자 선택 기준으로 사용되었다. VIP는 DPLS 모델의 가중치 벡터와 모델의 차원에 의해 설명되는 정도로부터 계산할 수 있다[15].

$$VIP = \sum_{\mathbf{a}} (\mathbf{w}_{ak})^2 \tag{3}$$

일반적으로 VIP 값은 각 유전자가 그 환자의 암 발병에 기여한 정도를 나타내는 값으로 유전자의 VIP값이 클수록 그 유전자가 해당 암유형 분류에 더 중요하다고 할 수 있다. 본 논문에서는 암 분류에 유용한 정보를 지닌 최적의 유전자 집합을 선별하는 기준으로 VIP값이 사용되었다.

클러스터링은 데이터를 그 속성에 따라 비슷한 것들끼리 묶어 분류하는 방법으로 실제 마이크로어레이 데이터의 경우 그 분포가 복잡하여 명확한 경계를 형성하지 않는 특성을 나타낸다. 퍼지 클러스터링(Fuzzy C-Means, FCM)은 다분류 클러스터링 방법 중 하나로서 일반적인 클러스터링 방법과는 달리 하나의 샘플이 다수의 집단에 속할 수 있으며 그 속하는 정도를 소속행렬로 표현하는 방법이다. 한 샘플이 동시에 몇 개의 클러스터의 멤버일 수 있다는 생각에서 출발하였고 클러스터 분할 분석을 보다 유연하게 한다[15,22]. 퍼지 클러스터링은 각 분류기 내의 거리들에 멤버쉽 값을 곱한 가중치를 목적함수로 두고 이를 반복 계산을 통하여 최소로 만드는 방법을 이용한다.

$$J_m(C,m)=\sum_{i=1}^{C}\sum_{k=1}^{N}(u_{k,i})^m d_{k,i}^2 \tag{4}$$

여기서 C는 전체 클러스터 개수, N은 데이터 개수, d_{ki} 는 각 클러스터 센터 i 와 데이터 k 사이의 거리, 그리고 u_k는 이때의 멤버쉽 함수이다. 반복 계산에 의해 각 클러스터 중심 (v_i)은 다음과 같은 퍼지 가중 평균으로 계산된다.

$$v_i=\frac{\sum_{k=1}^{N}(u_{k,i})^m x_k}{\sum_{k=1}^{N}(u_{k,i})^m},\quad \forall i \tag{5}$$

한편 퍼지 분류기에 사용되는 새로운 샘플들의 각 클러스터에서 거리를 나타내는 멤버쉽 값은 다음과 같이 표현된다.

$$u_{N+1,i}=1\Big/\sum_{j=1}^{C}\left(\frac{d_{k,i}^2}{d_{k,j}^2}\right)^{\frac{2}{(m-1)}} \tag{6}$$

퍼지 분류기는 각 샘플이나 유전자들이 하나 이상의 암 유형에 속할 수 있다는 생물학적 특성에 더 적합하다. 그림 2는 사용된 유전자 선택, 차원감소 및 퍼지분류기의 순서도를 나타낸다[15]. 첫째, DPLS의 VIP를 기준으로 각 특정 질병에 관련 있는 유전자들을 선별한다. 둘째, 주성분분석을 이용하여 차원을 축소하여 마이크로어레이 데이터를 분석한다. 셋째, 주성분 분석에 의해 차원이 축소된 스코어(score)를 기준으로 다분류 퍼지 클러스터링을 수행한다. 이를 유전자 발현 데이터 패턴 해석, 특정 암 유형(subclass)의 판별분석, 암환자의 임상결과 예측에 사용한다.

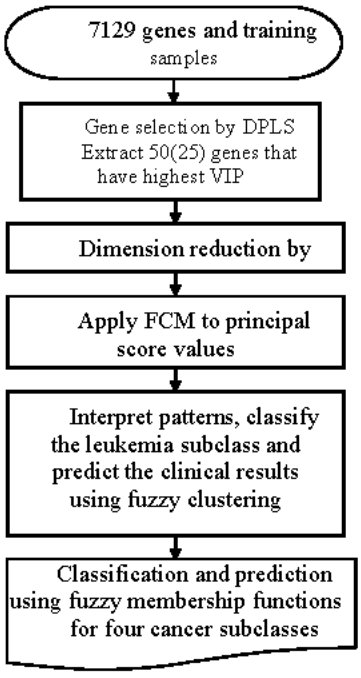


그림 2. 유전자 발현 데이터를 위한 다분류 퍼지 분석법.

Fig. 2. Schematic flow diagram of the fuzzy-based data analysis algorithm for gene expression data.

3. 백혈병 유전자 발현 데이터를 이용한 암 진단

제시된 방법은 유명한 Golub *et al.* [6] 논문에서 제시된 백혈병(leukemia) 데이터에 적용되었다. 이 데이터는 oligonucleotide microarrays(Affymetrix)를 사용한 것으로 7129 유전자와 72 샘플로 구성되어 있다. 데이터는 http://www.genome.wi.mit.edu/MPR에서 받을 수 있다. 백혈병 데이터(acute leukemia)는 크게 급성 림프구성 백혈병(Acute Lymphoblastic Leukemia, ALL)과 급성 골수성 백혈병(Acute Myeloid Leukemia, AML)으로 나뉘고, ALL은 다시 영향을 받는 림프구 세포의 형태에 따라 T-cell ALL 과 B-cell ALL로 분류된다. 72개 백혈병 데이터 중 47 샘플은 ALL (38 B-ALL과 9 T-ALL)이고 25 샘플은 AML이다. 전체적으로 학습데이터(training data) 는 38샘플로 27 샘플은 ALL 환자 (19 B-ALL와 8 T-ALL) 그리고 11 샘플은 AML 환자에서 수집되었다. 테스트 샘플(test sample)은 20 ALL 환자와 14 AML 환자로 전체 34 샘플로 구성 되었다. 통계적 처리에 앞서 데이터에 log10을 취하고 마이크로어레이내에서 실험오차 및 노이즈를 줄이기 위해 평균이 0이고 분산은 1이 되도록 정규화하였다[23]. 제안된 다분류 퍼지 접근법은 백혈병 데이터를 1) 급성 림프구성 백혈병(ALL) 과 급성 골수성 백혈병(AML), 2) ALL subtype (T-cell or B-cell), 3) AML subtype (M1, M2, M4, or M5), 4) AML subtype (success or failure clinical outcome)으로 유형별로 적용하였다. 이러한 다양한 유형의 백혈병은 유사한 조직화학적 특성을 지니지만 각 유형에 대한 처방은 다르다. 따라서 임상학적인 관점에서 특정 유형의 백혈병에서만 발현되는 유전자를 찾고 새로운 환자를 적절히 진단하기 위해 환자가 어떤 범주에 속하는지 판별하는 작업이 중요하다.

3.1 AML과 ALL 분류

본 단락에서는 퍼지 클러스터링을 백혈병 중 가장 대표적인 분류인 ALL 과 AML 의 유전자 발현 해석 및 분류에 응용하였다. DPLS 모델이 사용되었고 VIP 순위에 따라 백혈병의 7129 유전자중 50개를 주요 유전자로 선별하였다. 각 선별 유전자의 특성은 NCBI LocusLink (http://www.ncbi.nlm.nih.gov/LocusLink)를 이용하여 파악하였다. 본 연구에서 선택된 유전자는 Golub *et al.*[6] 의 선별 유전자와 비교하여 볼 때, zyxin, leukotriene, leptin, CD33 antigen, FAH, and myeloperoxidase (MPO)등에 높은 순위를 부여하였다. 이렇게 선택된 유전자

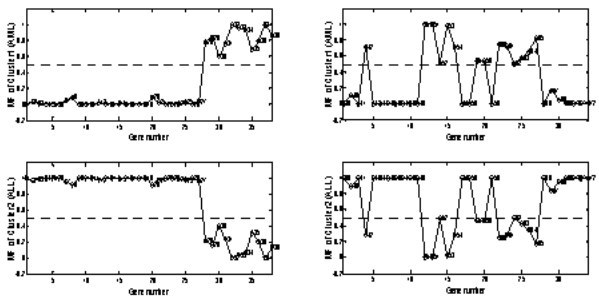


그림 3. 퍼지 분류기를 이용한 백혈병의 학습 및 테스트 샘플들의 예측 결과 (클러스터1: AML, 클러스터2: ALL).

Fig. 3. Prediction result of membership values of FCM for training (left) and test samples (right) with cluster 1 (AML, upper) and cluster 2 (ALL, lower).

들은 AML과 ALL사이에 큰 발현 차이를 보이며, 또한 발현 수준이 상당히 높음을 알 수 있었다. 기존의 백혈병에 관한 많은 연구 결과와 비교해 볼 때 DPLS모델에 의해 선택된 유전자들이 대부분 이전 연구결과와 일치함을 확인할 수 있었다[6,23-25].

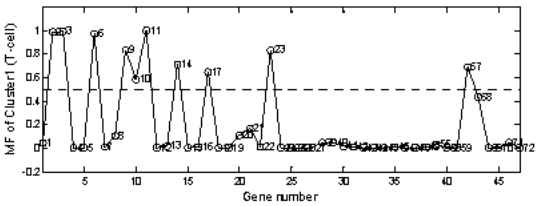
선택된 주요 유전자를 주성분 분석을 통하여 차원을 축약하고 그에 대한 스코어 값을 퍼지분류기로 분석하였다. 그림 3는 백혈병의 AML/ALL에 대한 퍼지 분류기의 멤버쉽 함수를 나타낸다. 학습 및 테스트 샘플들의 예측 결과, 38개의 학습 샘플의 경우 AML과 ALL을 완벽히 분류하였고 34개의 테스트 샘플의 경우 2개(#42,66)의 분류에러가 나타났다.

3.2 ALL subclass(T-cell 과 B-cell) 분류

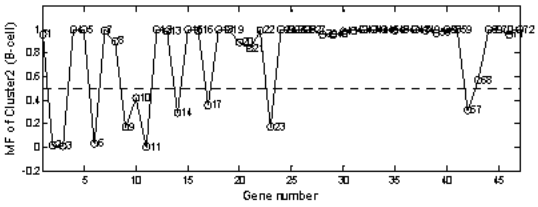
본 단락에서는 ALL 환자들의 유형의 해석 및 분석에 퍼지 클러스터링을 이용하였다. 일반적으로 ALL은 T-cell 과 B-cell로 분류되고 임상학적으로 B-cell이 T-cell보다는 치료하기가 훨씬 쉽다고 알려져 왔다[7]. 따라서 환자가 ALL로 분류된 경우, 다시 T-cell 과 B-cell로 분류하는 것 역시 중요한 문제이다. 47 ALL 백혈병 환자 중 27개 샘플을 학습샘플(19 B-cell ALL and 8 T-cell ALL)로 선택하여 이를 바탕으로 25개의 유전자를 선택하였다. 선택된 유전자중 T-cell antigen receptor (X03934, 9p56), TCRB (X00437, 7q34), CD47 (X69398 3q13), CD7 (D00749, 7q34), TCF7(X59871, 5q31)등이 ALL환자의 염색체 변이와 관련된 것으로 확인되었다. 그림 4은 선택된 유전자로 주성분 분석과 퍼지 분류기를 수행한 결과이다. 분류결과 47명의 ALL환자 중 하나의 샘플(#17)을 제외한 모든 환자의 예측결과가 실제로 ALL 유형과 일치함을 알 수 있었다. 이러한 결과는 선별된 유전자들이 신호 대 노이즈 비 (signal-to-noise ratio)가 높다고 알려진 유전자 칩의 데이터들 중에서 각 유형의 분류를 위한 주요 유전자(super genes)로 사용될 수 있음을 보여 주고 있다.

3.3 AML 환자의 유형 분류

일반적으로 AML의 경우 6가지 유형 (M1, M2, M3, M4, M5, M6)으로 다시 분류될 수 있다. M5 유형은 기존의 French-American-British (FAB)의 방법으로는 분류하기 어렵고, 특히 M3 유형은 retinoids성분의 약에 일반적으로 반응하지 않기 때문에 이 M3 유형의 분류는 특히 중요하다. 퍼지 분류기를 AML subclass인 M1, M2, M4, M5의 다유형 분류에 적용해 보았다. 20개 샘플을 학습 샘플로 사용하여 DPLS모델링을 수행하였고, 25개의 유전자를 선택하였다. 남아있는 5명의 샘플(#62-66)은 기존의 Golub *et al.* [6]방법으로 분류가 되지 않아 테스트 샘플로 사용하였다. 그림 5(a)는 20개 AML 학습 샘플 분석 (M1, M2, M4, or M5)과 5개 테스트 샘플(#62-66)들에 대한 주성분 분석 결과이다. AML 환자는 4개 유형으로 구성되어 있고 20개의 학습 샘플에 대하여 퍼지 분류기는 어떤 에러도 없이 분류가 가능하였다. 그림 5(b)은 5명의 subclass가 알려지지 않은 AML환자들(#62-66)에 퍼지 분류기를 적용한 결과이다. 그 결과 3명의 AML환자들(#63, 64, 65)은 유형 M1이고 두 명의 AML환자들(#62, 66)은 유형 M2로 예측되었다. 또한 선택된 유전자들은 생물학적 병리학적으로 AML 유형과 관련성을 찾을 수 있었다. 이러한 다유형 암의 경우에 퍼지 분류결과에 따라 암의 아종을 예측할 수 있고 예측 결과는 각 환자에 적합한 치료방법을 결정하는데 좋은 정보로 사용 될 수 있다.

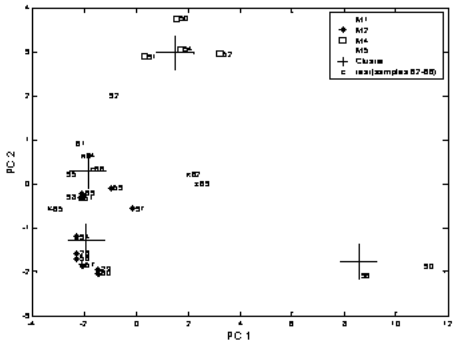


(a) Cluster 1 (T-cell lineage)

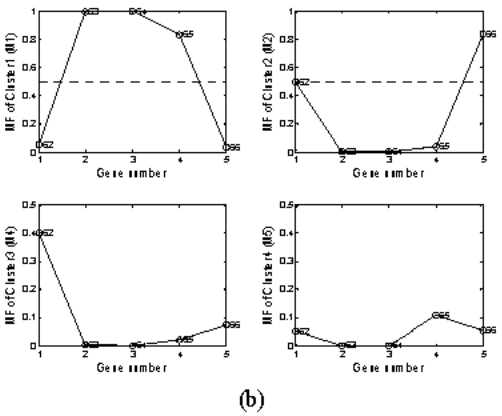


(b) Cluster 2 (B-cell lineage)

그림 4. 퍼지 분류기를 이용한 47개 ALL 샘플들의 예측 결과.
Fig. 4. FCM clustering results of ALL samples of 47 patients.



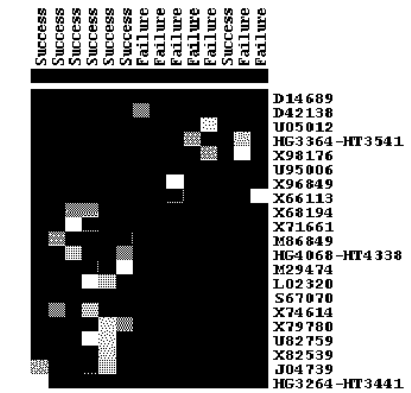
(a)



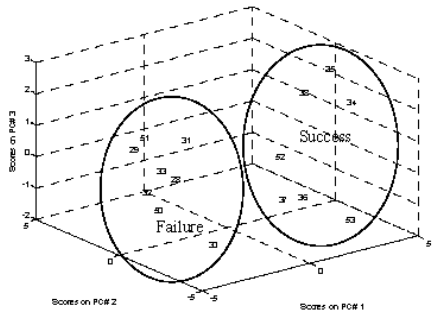
(b)

그림 5. (a) 주성분 분석을 이용한 20개 AML 학습 샘플 분석 (M1, M2, M4, or M5)과 5개 테스트 샘플 분석(62-66). (b) 퍼지 분류기를 이용한 5개 테스트 샘플의 예측 결과.

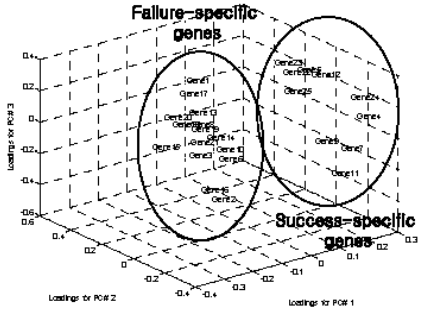
Fig. 5. (a) PCA score plot of 20 AML patients with M1, M2, M4, or M5 and 5 test samples (62-66). (b) Prediction result of membership values of FCM for 5 test AML samples (62-55) with subclass M1, M2, M4, or M5.



(a)



(b)



(c)

그림 6. (a) 백혈병유전자중 AML 환자의 치료결과에 따른 유전자 heat map, (b) 15명의 AML환자의 스코어 플롯, (c) 로딩 플롯 (8명: 치료 실패, 7명: 치료 성공).

Fig. 6. (a) Gene expression maps of a leukemia data set based on the 25 selected genes most relevant for discrimination between success and failure of AML leukemia treatment, (b) Score plot using PCA for 15 AML patients (8 failure and 7 success samples), (c) Loading plot using PCA for 15 AML patients (8 failure and 7 success samples).

3.4 AML 환자의 치료결과 예측

현재까지 백혈병의 메커니즘과 발병에 관한 다양한 변수가 연구되어 일부 환자의 경우 치료결과를 예상한 것과 같은 결과를 보였다. 하지만 일부 환자는 예상과는 정 반대의 결과를 보여 이 외에도 다양한 요소가 연관되어 있음을 알 수 있다[24]. 암 환자의 유전자 발현 패턴은 치료의 성공여부에 따라 큰 차이를 보인다. 따라서 유전자 발현 패턴의 이용은 기존의 치료방법이 실패할 확률이 높은 환자를 분류하는데 더 높은 성능을 보일 것으로 기대되고 있다.

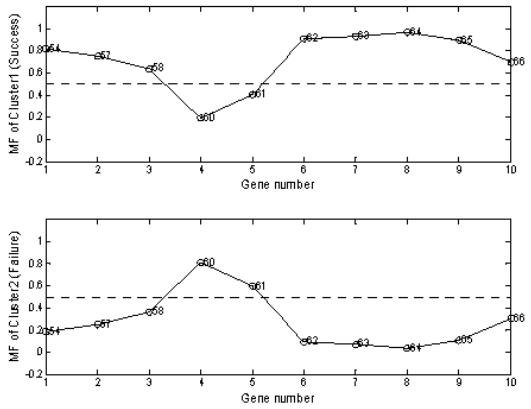


그림 7. 퍼지 분류기를 이용한 10개 AML 테스트 샘플의 치료 성공/실패의 예측 결과 (#54, 57, 58, 60-66).

Fig. 7. Prediction result of membership values of FCM for 10 test samples (54, 57, 58, 60-66) with AML patients who lived and died after treatment.

백혈병 환자의 치료결과를 예측하는데 유용한 유전자군을 찾기 위한 문제의 일부로, AML 환자의 치료결과를 예측하는 유전자를 찾았다. 25명의 AML 환자 중 15명을 학습 집단으로, 10명을 테스트 집단으로 나누고 DPLS를 이용하여 AML 환자의 치료결과를 성공과 실패로 분류하는데 기여도가 큰 상위 25개의 유전자를 골랐다. 백혈병 환자는 염색체 상의 유전자 위치가 변경된 경우가 많고 이는 종종 백혈병 진단에 중요하기 때문에 25개의 선택된 유전자의 위치를 NCBI LocusLink로 찾아보았다.

25개의 선택된 유전자 대부분은 이전에 AML이나 다른 백혈병 타입과 관련성이 있다는 것이 입증된 것이었다. Lyons-Weiler *et al.* [4]이 밝힌 유전자의 대부분이 본 연구에서 선택되었다 (HoxA9, PIG-B, MACH-alpha-2 protein, BPI Bactericidal/permeability increasing protein, Autoantigen PM-SCL, ERGIC-53 Protein, and so on). 선택된 유전자 중 하나인 HoxA9의 경우, 발현양이 증가하면 leukocytes와 lymphocytes의 생산을 증가시키는 것으로 추정된다. 실제로, HoxA9와 Meis1의 발현양이 증가되도록 조작된 primary bone marrow cells을 실험용 쥐에 주입하면 석 달 안에 AML이 발병된다[6]. 그림 6(a)는 선택된 유전자중 치료결과와 연관성이 높은 상위 25개의 유전자의 발현 패턴을 보여준다. 그림에서 보듯이 선택된 유전자의 발현 수준은 AML 치료결과에 따라 현저한 차이를 보인다.

25개의 유전자를 선택한 뒤 주성분 분석을 이용하여 데이터의 차원을 더욱 줄여 보았다. 4개의 주성분으로 데이터를 투영 한 후 그림 6(b)의 3차원 그림을 이용하여 15명의 AML 환자의 치료결과와 선택된 유전자간의 상관관계를 조사하였다. 이 중 그림 6(c)의 loading plot은 25개의 유전자간의 상호연관성을 조사하는데 사용될 수 있다. Loading plot에서 환자의 치료결과가 좋을 때 발현양이 높은 유전자는 오른쪽에, 치료결과가 나쁠 때 발현양이 높은 유전자는 왼쪽에 나타난다. 즉 이는 그룹 특이적인 조절 패턴인 coregulation 패턴을 나타낸다고 볼 수 있다[20,25].

그림 7은 10명의 AML 환자의 퍼지 분류기의 치료 예측 결과이다. 이 중 8명은 치료가 성공적일 것으로, 2명은 실패

할 것으로 예측되었다. 따라서 제안된 방법은 AML 환자의 치료결과를 예측하는데 사용될 수 있으며 또한 이는 백혈병의 재발이나 이와 연관된 특정 유전자를 찾아낼 수 있기 때문에 이를 이용한 신약개발에 도움이 될 수 있다. 비록 치료결과는 환자의 나이, 진단 시점 등 다양한 변인에 의해 영향받을 수 있지만, 본 연구결과를 바탕으로, 제안된 방법이 환자의 유전적 특이성을 바탕으로 한 백혈병의 치료결과를 예측하는 새로운 기준을 밝힐 수 있을 것으로 기대된다.

III. 결론

인간 및 다양한 생물체를 대상으로 한 게놈 프로젝트가 완결되어 유전자 시퀀스와 수 천개의 유전자 발현 패턴을 동시에 탐색하는 마이크로어레이 데이터가 날로 증가하고 있다. 특히 암분류에 사용되는 마이크로어레이 데이터분석의 목표는 각 암에 관련된 유전자들을 규명함으로써 일반적인 질병 유발 유전자들을 대상으로 하는 약의 개발과 환자마다 특이한 유전자들을 검출하여 환자중심의 치료제를 개발하데 고급 정보를 제공하는 것을 최종 목적으로 한다. 실제 환자들의 유전자 발현 데이터 분석은 분석 결과가 직접 임상에 사용될 수 있다는 측면에서 그 효용성과 적용 가능성이 크다고 할 수 있다. 이미 많은 대학 병원을 중심으로 방대한 양의 암환자 마이크로어레이 데이터 베이스가 구축되어 있으며, 데이터 분석 과정에서도 의사들이 직접 참여함으로써 분석 결과를 임상에 적용하려 노력하고 있다.

본 연구에서는 마이크로어레이 데이터분석에 관한 기본적인 개념을 소개하였으며 전산생물학을 이용한 암환자의 데이터분류를 위한 유전자 선택방법과 특히 두 개 이상의 카테고리에 적용할 수 있는 다유형 암분류 방법을 구성하는 것을 목표로 하였다. 제시된 방법이 본 데이터 분석의 최종 목표인 질병 치유를 위한 메커니즘의 규명과 신약의 개발에 통계적이고 논리적인 정보를 제공할 수 있기를 기대한다.

참고문헌

[1] G M. Hampton and H. F. Frierson, "Classifying human cancers by gene expression analysis," *Trends Mol. Med.*, vol. 9, 5-10, 2003.

[2] J. Quackenbush, "Computational analysis of microarray data," *Nat. Rev. Genet.*, vol. 2, 418, 2001.

[3] Y. Lu and J. Han, "Cancer classification using gene expression data," *Information Systems*, *28*, 243-268, 2003.

[4] J. Lyons-Weiler, Patel, S. and S. Bhattacharya, "A classification-based machine learning approach for the analysis of genome-wide expression data," *Genome Res.*, vol. ***13***, 503-512, 2003.

[5] S. Ramaswamy and T. R. Golub, "DNA microarrays in clinical oncology," *J. Clin. Onc.*, vol. 20, 1932-1945, 2002.

[6] T. R. Golub, D. K. Slonim, P. Tamayo and E. S. Lander, "Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring," *Science*, vol. 286, 531-537, 1999.

[7] P. Kebriaei, J. Anastasi, & R. A. Larson, Acute lymphoblastic leukaemia: diagnosis and classification. *Best Pract Res Clin Haematol.*, vol. 15, 597-621, 2002.

[8] M. F. Ochs and A. K. Godwin, "Microarrays in cancer: research and applications," *BioTechniques*, vol. 34, S4-S15, 2003.

[9] S. Dudoit, J. Fridlyand, and T. Speed, "Comparison of discrimination methods for the classification of tumor using gene expression data," *J. Am. Stat. Assoc.*, vol. 97, 77-87, 2002.

[10] D. V. Nguyen, and D. M. Rocke, "Tumor classification by partial least squares using microarray gene expression.," *Bioinformatics*, vol. 18(1), 39-50, 2002.

[11] A. Alizadeh, M. B. Eisen, and L. M. Staudt, "Different types of diffuse large b-cell lymphoma identified by gene expression profiling," *Nature*, vol. 403, 503-511, 2000.

[12] U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine, "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays," *Proc. Natl. Acad., Sci. USA*, vol. 96, 6745-6750, 1999.

[13] S. Bicciato, M. Pandin, G Didone and C. Di Bello, "Pattern identification and classification in gene expression data using an autoassociative neural network model," *Biotechnol. Bioeng.*, vol. 81, 594-606, 2002.

[14] J.-H. Cho, D. K. Lee, J. H. Park, K. W. Kim and I.-B Lee, "Optimal approaches for classification of acute leukemia subtypes based on gene expression data," *Biotech. Prog.*, vol. 18(4), 847-854, 2002.

[15] C. K. Yoo, I. Lee, and P. A. Vanrolleghem; "Interpreting patterns and analysis of acute leukemia gene expression data by multivariate fuzzy statistical analysis," *Comp. & Chem. Eng.*, vol. 29, 1345-1356, 2005.

[16] J. Dorrie, H. Gerauer, Y. Wachter and S J. Zunino, "Resveratrol induces extensive apoptosis by depolarizing mitochondrial membranes and activating caspase-9 in acute lymphoblastic leukemia cells," *Cancer Res.*, vol. 61, 4731-4739, 2001.

[17] P. J. Park, L. Tian and I. S. Kohane, "Linking gene expression data with patient survival times using partial least squares," *Bioinformatics,* vol. 18(1), S120-S127, 2002.

[18] T. *et al.,* Ross, "Systematic variation in gene expression patterns in human cancer cell lines," *Nature Genetics*, vol. 24, 227-234, 2000.

[19] U. *et al.,* Scherf, "A gene expression database for the molecular pharmacology of cancer," *Nat. Genet.*, vol. 24, 236-244, 2000.

[20] G Stephanopoulos, D. H. Hwang, W. A. Schmit, J. Misra and G Stephanopoulos, "Mapping physiological states from microarray expression measurements," *Bioinformatics*, vol. 18(8), 1054-1063, 2002.

[21] J. Stephenson, "Human genome studies expected to revolutionize cancer classification," *J. Am. Med. Assoc.*, vol. 282, 927-92, 1999.

[22] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed., John Wiley & Sons. New York, 2001.

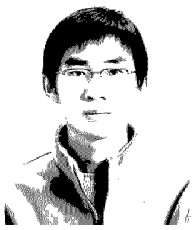
[23] Y. H. Yang, S. Dudoit, P. Lu, D. M. Lin, V. Peng, J. Ngai, and T. P. Speed, "Normalization for cDNA microarray data: A robust composite method addressing single and multiple slide systematic variation," *Nucleic Acids Res.*, vol. 30, 15-21, 2002.

[24] J. G. Thomas, J. M. Olson, S. J. Tapscott and L. P. Zhao, "An efficient and robust statistical modeling approach to discover differentially expressed genes using genomic expression profiles," *Genome Res.*, vol. 11, 1227-1236, 2001.

[25] E. *et al.,* Yeoh, "Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. *Cancer*," *Cell.*, vol. 1, 133-143, 2002.

유 창 규

1993년 연세대학교 화학공학과 학사. 1995년 포항공대 화학공학과 석사. 1998년 두산기술원 전임연구원. 2002년 포항공대 화학공학과 박사. 2003년~2004년 벨기에 겐트대학교 BIOMATH학과 Post-doc. 현재 포항공대 환경공학부 연구 조교수. 관심분야는 시스템즈 생물학(생명정보학), 데이타 마이닝, 공정제어 및 최적화, 환경시스템공학.



김 영 황

2005년 경북대 화학공학과 학사. 현재 포항공대 화학공학과 석사과정 재학중. 관심분야는 제어시스템, 반복제어.

이 민 영

2003년 포항공대 화학공학과 학사. 2005년 포항공대 화학공학과 석사. 현재 지능자동화연구센터 연구원. 관심분야는 시스템즈 생물학, 생물정보학.

이 인 범

1977년 연세대학교 화학공학과 학사. 1979년 KAIST 화학공학과 석사. 1982년 한국과학기술연구원 연구원. 1987년 Purdue univ. 화학공학과 박사. 1988년-현재 포항공대 화학공학과 교수. 1998년-현재 지능자동화연구센터 소장. 2004년-현재 한국공학한림원 정회원.