[UCI]I804:24011-200000234250

2007年　2月

博士學位論文

표준화 후 유의한 유전자 선택 방법 조합을 이용한 마이크로어레이 분류 시스템 구현

朝鮮大學校 大學院

電算統計學科

朴　受　泳

표준화 후 유의한 유전자 선택 방법 조합을 이용한 마이크로어레이 분류 시스템 구현

A Implementation of System on Microarray Classification Using Combination of Significant Gene Selection Method after Normalization

2007년 2월 일

朝鮮大學校 大學院

電算統計學科

朴 受 泳

표준화 후 유의한 유전자 선택 방법 조합을 이용한 마이크로어레이 분류 시스템 구현

指 導 敎 授　丁 采 榮

이 論文을 理學博士學位申請 論文으로 提出함

2006년 10월 일

朝鮮大學校 大學院

電算統計學科

朴 受 泳

목 차

표 목 차 ····· iv
그림 목차 ····· v
ABSTRACT ····· vii

Ⅰ. 서 론 ····· 1

Ⅱ. 관련연구 ····· 4
A. 마이크로어레이 개요 ····· 4
B. 마이크로어레이의 종류 ····· 9
1. cDNA 마이크로어레이 ····· 10
2. 올리고뉴클레오티드 마이크로어레이 ····· 12
C. 마이크로어레이의 제작 방법 ····· 15
1. 애피메트릭스 올리고뉴클레오타이드 마이크로어레이 ····· 15
2. cDNA 마이크로어레이 ····· 19
D. 마이크로어레이 데이터의 특징 ····· 22
E. 마이크로어레이 기반의 기존 암 분류 연구 ····· 24
1. 현대 암 진단 방법 ····· 24
2. 분자 생물학을 이용한 암 분류 ····· 24
3. 마이크로어레이를 이용한 암 분류 연구 분야 ····· 25
4. 요구되는 최소 시스템의 개요 ····· 26
5. 마이크로어레이 기반 암 분류 연구 사례 ····· 27
a. Golub의 방법 ····· 27
b. 다중 분류기 시스템을 이용한 방법 ····· 30

Ⅲ. 유의한 유전자 선택 방법 ····· 34
A. 마이크로어레이의 표준화 ····· 34
1. 포괄적인 표준화 방법 ····· 38

2. 특성화한 표준화 방법 ····· 40
B. 유의한 유전자 선택 ····· 45
C. 조합 방법 ····· 50

Ⅳ. 신경망 ····· 52
A. 신경망 ····· 52
B. 지도 학습과 비지도 학습 ····· 56
1. 지도 학습 ····· 56
2. 비지도 학습 ····· 57
C. 신경망 학습 방법 ····· 59
1. 퍼셉트론 ····· 59
a. 단순 퍼셉트론 ····· 60
b. 다층 퍼셉트론 ····· 61

Ⅴ. 실 험 ····· 63
A. 제안하는 시스템 구조도 ····· 63
B. 실험 데이터 ····· 65
C. 실험 결과 및 고찰 ····· 66
1. 상위 200개 유전자에 대한 다양한 조합에 따른 분류 성능 ····· 68
2. 상위 90개 유전자에 대한 다양한 조합에 따른 분류 성능 ····· 71
3. 상위 50개 유전자에 대한 다양한 조합에 따른 분류 성능 ····· 75
4. 평균 실행 시간 ····· 79

Ⅵ. 결론 및 향후과제 ····· 80

각 슬라이드에서 잘못 분류된 유전자 정보 ····· 81
참고문헌 ····· 86

표 목 차

[표 2.1] 마이크로어레이를 이용한 암 분류 연구 분야 ····· 25
[표 5.1] 상위 200개 유전자에 대한 단일 척도 사용에 따른 분류 성능 ····· 68
[표 5.2] 상위 200개 유전자에 대한 2개 척도 조합에 따른 분류 성능 ····· 70
[표 5.3] 상위 200개 유전자에 대한 3개 척도 조합에 따른 분류 성능 ····· 71
[표 5.4] 상위 90개 유전자에 대한 단일 척도 사용에 따른 분류 성능 ····· 71
[표 5.5] 상위 90개 유전자에 대한 2개 척도 조합에 따른 분류 성능 ····· 73
[표 5.6] 상위 90개 유전자에 대한 3개 척도 조합에 따른 분류 성능 ····· 74
[표 5.4] 상위 50개 유전자에 대한 단일 척도 사용에 따른 분류 성능 ····· 75
[표 5.5] 상위 50개 유전자에 대한 2개 척도 조합에 따른 분류 성능 ····· 77
[표 5.6] 상위 50개 유전자에 대한 3개 척도 조합에 따른 분류 성능 ····· 78
[표 5.7] 평균 실행 시간 ····· 79

그림 목차

[그림 2.1] DNA 구조 ····· 4
[그림 2.2] 염색체의 구조 ····· 6
[그림 2.3] Central dogma ····· 10
[그림 2.4] cDNA 마이크로어레이 ····· 11
[그림 2.5] 애피메트릭스 올리고뉴클레오타이드 마이크로어레이 ····· 12
[그림 2.6] 1애피메트릭스 올리고뉴클레오타이드 마이크로어레이의 제작 ····· 16
[그림 2.7] 불완전 짝 검출 유전자의 적용 예 ····· 18
[그림 2.8] 마이크로어레이 크기 ····· 22
[그림 2.9] 행렬로 나타낸 유전자 발현 데이터 ····· 23
[그림 2.10] 요구되는 최소 시스템의 개요 ····· 26
[그림 2.11] 암 분류를 위한 다중 분류기 시스템의 구조 ····· 32
[그림 3.1] Cy3 & Cy5 plot(표준화하기 전의 자료) ····· 35
[그림 3.2] logG & logR plot ····· 37
[그림 3.3] M-A plot ····· 37
[그림 3.4] Global 표준화 후 plot ····· 39
[그림 3.5] Lowess 표준화 후 plot ····· 44
[그림 3.6] 유전자 발현 정도와 이상 유전자 모델 사이의 관계 ····· 46
[그림 3.7] 유의한 유전자 선택을 위한 기존 유사성 척도 ····· 48
[그림 3.8] 정보력 있는 유전자 선택을 위한 조합 방법 알고리즘 ····· 51
[그림 4.1] 생물학적인 뉴런의 구조 ····· 52
[그림 4.2] 인공뉴런의 구조 ····· 53
[그림 4.3] 단층 신경망 ····· 54
[그림 4.4] 다층 신경망 ····· 55
[그림 4.5] 퍼셉트론의 구조 ····· 59
[그림 4.6] 단순 퍼셉트론의 구조 ····· 60

[그림 4.7] 다층 퍼셉트론의 구조 ····· 61
[그림 5.1] 제안하는 분류 시스템 구조도 ····· 64
[그림 5.2] 마이크로어레이 분류 시스템 ····· 67
[그림 5.3] 단일 척도 사용에 따른 상위 200개 유전자 M-A plot ····· 69
[그림 5.4] 단일 척도 사용에 따른 상위 90개 유전자 M-A plot ····· 72
[그림 5.5] 단일 척도 사용에 따른 상위 50개 유전자 M-A plot ····· 76

Abstract

A Implementation of System on Microarray Classification Using Combination of Significant Gene Selection Method after Normalization

by Park, Su-Young
Advisor : Prof. Jung, Chai-Young
Dept. of Computer Science and Statistics
The Graduate School
Chosun University

Microarray technology provides large-scale gene expression profiles according to experimental conditions such as quantitative values. The transcription of genes are triggered or inhibited by complex biological interactions and associations. To understand not only the function and the role of a gene in a cell but also the mechanism of cellular phenomena, we need to capture snapshots of cellular process using microarray technology. Nowadays, microarray technology is expected to contribute to filling the ultimate purpose of the total study in bioinformatics. Microarray technology allows the monitoring of expression levels for thousands of genes simultaneously. This novel technology helps us to understand gene

regulation as well as a gene by gene interactions more systematically.

In the microarray experiment, however, many undesirable systematic variations are observed. Normalization is the process of removing some sources of variation which affect the measured gene expression levels. Normalization plays an important role in the earlier stage of microarray data analysis. The subsequent analysis results are highly dependent on normalization.

Microarray, based on its gene expression data information, has been applicable to the field of cancer diagnosis with the computer-aided classification and prediction technology. Current clinical practice involves an experienced hematopathologist's interpretation of the tumor's morphology. In this way, however, classification remains imperfect and errors do occur. So, it has been suggested that such microarrays could provide a tool for correct cancer classification. But from a datamining-based point of view, there is one difficulty in microarray data analysis that the number of samples is very small while the number of attributes(i.e., genes) is very large. Therefore, to classify the subtypes of cancer correctly using current microarray technology, we should select the informative genes whose expression pattern was strongly correlated with the class distinction to be predicted.

Independently separated informative genes can contribute to inspiring the study of medical cure after the correct classification by them. These informative genes list data suggest that genes useful for cancer class prediction may also provide insight into cancer pathogenesis and pharmacology. So for this application, we have to separate this informative genes list independently. And that list should be consistent, trusted, and strongly correlated with the class distinction to be predicted.

In this paper, the system that can create the separated informative genes list after Lowess normalization is proposed. The effectiveness of these system and method was evaluated through some experiments. In the experimental results using MLP(Multi-Layer Perceptron)-based classifier system, it was found that the proposed system and the suggested combination method create an independently separated informative genes list with consistency, trust, and strong correlation with the class distinction for microarray data.

Therefore, the proposed system and the suggested combination method in this paper are expected to contribute to providing insight into cancer pathogenesis and pharmacology as well as correct classification of cancer types.

Ⅰ. 서 론

세계의 관심이 집중되었던 인간유전체(*Human Genome*) 프로젝트의 목표는 인간이 보유한 22쌍의 상염색체와 *X, Y*로 표현되는 성염색체 등에 포함된 유전자의 염기서열 해독과 유전체지도의 작성이었다. 성공적인 연구의 결과로, 2002년 2월 인간유전체지도가 완성되었다. 이는 2000년 6월에 발표되었던 초안 위에 생물학적 표지를 심어 특정 유전자의 위치를 표시한 99% 수준의 상세 지도로, 당시 누락되었던 부분에 대해서도 대폭 보완된 것이다. 지도의 완성으로, 인간유전체로부터 각 유전자들의 생체기능을 밝히고 개인, 생물간 유전체(*Post-Genome*) 시대가 열리게 되었다.

인간유전체 프로젝트를 통해 여러 개체의 완전한 유전자 서열정보 등 방대한 양의 자료들을 밝혀내면서 이제는 자료 자체보다는 얻어지는 방대한 양의 자료를 어떻게 해석하고 이용하는 것이 더 어렵고 당면한 문제로 부각되었다. 기존의 전통적인 분자생물학에 의한 연구방법은 단일 유전자에 대한 실험에 근거하여 진행되기 때문에 그 결과물 또한 매우 제한적이고 유전자의 전체 움직임을 관찰하기에는 한계가 있다. 반면에, *DNA* 마이크로어레이(*microarray* 또는 *microchip*)는 하나의 칩(*chip*)상에서 전체 유전체(*genome*)의 발현양상을 탐색할 수 있고, 동시에 수천 개의 유전자들 간의 상호작용도 관찰할 수 있다. 또한 짧은 시간에 엄청난 분량의 자료를 만들어낼 수 있는 가능성을 지니고 있어, 최근 주목받고 있다.

DNA 마이크로어레이 실험과정을 거쳐 생성되는 유전자 발현자료들을 이용하여, 생물학, 의학 및 약학 분야의 연구자들은 유전자의 기능, 유전자의 발현 및 통제방식, 그리고 유전체구조의 비밀 등과 같은 문제를 해결하기 위해 노력하고 있다. 그 중에서도, 관심 있는 생물학적 현상과 관련 있는 정보를 제공해 주는 유전자들을 찾는 것은 중요한 과제라 할 수 있다. *DNA* 마이크로어레이 실험의 특성상 원하는 만큼의 반복실험을 할 수 없기 때문에 관측치의 개수보다 변수의 개수가 더 많은 특성을 갖고 있다. 따라서 기존의 통계학 방법들을 적용할 수 없는 새로운 유

형의 자료구조를 갖게 된다. 판별분석, 군집분석, 분류분석과 같은 통계방법들을 이용하기 위해서는 자료의 차원을 축소시켜주어야 하는데, 이 때 수천 개의 유전자들 중에서 각 집단 간의 발현양상에서 유의한 차이를 보이는 유전자들을 선택하여 자료로 이용하게 된다. 이 유의한 유전자들은 정상 조직과 서로 다른 질병의 차이, 질병의 소분류들의 차이 등을 밝히거나 새로운 표본조직의 생물학적 현상을 예측하는데 이용할 수 있어, 이들을 찾기 위해 다양한 유전자 선택(*gene selection*)방법들이 개발되었고 새로운 방법에 대한 연구들도 활발히 이루어지고 있다.

현재 이루어지고 있는 마이크로어레이 데이터 분석은 크게 새로운 생물학적 하위 클래스를 찾기 위한 클래스 발견(*class discovery*)과 이미 알려져 있는 기존 하위 클래스에 새로운 개체를 적절하게 분류하기 위한 클래스 예측(*class prediction*)으로 분류할 수 있다. 이와 같은 연구를 기반으로, 유전자 발현 분석 시스템을 구축하여 유전 질환에 대해 분류 및 예측을 통한 진단을 하고, 보다 구체적이고 종합적인 분석을 위해 염기 서열 분석 시스템을 구축하여 유전자의 기능 예측을 통한 유전자 망(*gene network*) 구축을 성공적으로 완수할 수 있을 것이라는 기대가 높아지고 있다. 특히, 마이크로어레이 데이터를 이용하여 유전 질환의 분류 및 예측을 하는 진단 분야는 기존 전산학의 발전을 바탕으로 가장 활발하게 연구가 진행되고 있는 부분이다[1][2].

수천 개에서 수 만여 개의 유전자들이 들어있는 마이크로어레이 데이터는 종양 샘플을 구하기가 쉽지 않을 뿐만 아니라 실험 비용도 매우 비싸 실제 표본의 개수에 비해 유전자의 개수가 훨씬 많다는 특성을 가지고 있다. 따라서, 수많은 유전자들로부터 실제 종양들의 세부 부류에 따라 확연하게 발현량이 변하는 표본 분류에 유용한 유전자들을 추출하기 위한 특징 추출(*feature selection*)방법과 이 유전자들을 이용하여 보다 정확한 종양 분류 모델(*tumor classification*)을 구축하는 것이 매우 중요하게 부각되고 있다[3][4].

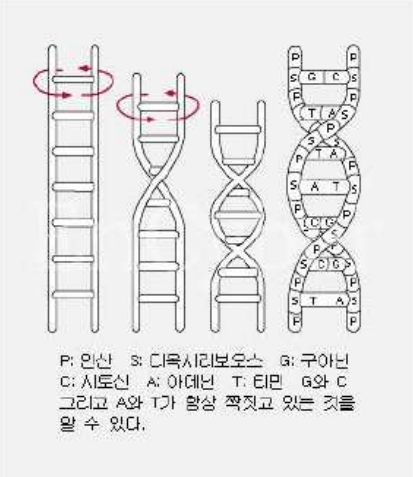
본 연구는 쥐 외피 줄기 세포의 신경 분화 동안 유전자 발현 프로파일의 변화를 찾는 실험에서 획득한 유전자에 대해 *Lowess* 표준화 후 정보력이 있는 유전자 목록을 별도로 추출할 수 있는 시스템을 고안하고 보다 정보력이 있는 유전자를 선택하기 위한 조합 방법을 제안한다. 제안한 방법의 실험적 검증을 위해서 마이크로어레이 데이터에 대해 *Lowess* 표준화 후 기존의 다양한 유전자 선택 방법을 적용하여 그에 따른 분류 성능을 비교 평가한다. 또한 기존의 선택 방법을 적절하게 조합함으로써 보다 신뢰성 있는 유용한 유전자 리스트를 확보할 수 있고 그에 따른 분류 성능도 향상시킬 수 있음을 보이고자 한다.

본 논문의 구성은 Ⅱ장에서 마이크로어레이에 대해 먼저 소개하고, 마이크로어레이 기술을 이용한 암 분류와 관련된 기존의 연구 성과들에 대해 소개한다. Ⅲ장에서 표준화 방법들을 소개하고, 기존의 특징 추출 방법을 이용한 유전자 선택 방법과 기존의 특징 추출 방법을 조합해서 보다 정보력이 있는 유전자들을 선택하기 위한 방법에 대해 기술한다. Ⅳ장에서 신경망의 기본적인 개념과 제안된 마이크로어레이 분류 시스템에서 사용하고자 하는 멀티퍼셉트론 학습 방법에 대하여 설명한다. Ⅴ장에서는 Ⅲ장에서 제안한 조합 방법과 본 장에서 제안한 분류 시스템이 어떠한 성능을 보이는지 실제 마이크로어레이 데이터 모의실험에 대한 결과를 기술하고, 이를 분석한다. 6장에서는 결론을 도출하고 향후 연구 과제에 대해 논하도록 한다.

Ⅱ. 관련연구

A. 마이크로어레이 개요

생명체의 생명 현상을 근원적으로 제어하는 것은 세포 핵 내부의 염색체에 존재하는 *DNA*(*deoxyribo nucleotide acid*)이다. 염색체는 히스톤 단백질(*histone protein*)과 *DNA*로 구성되어 있고, 간기에서는 염색사의 형태로 존재한다. 히스톤 단백질은 진핵 생물체에서 원핵 생물체의 *DNA*보다 더욱 긴 *DNA*가 한정된 공간에 문제없이 엉키지 않고 염색체의 형태로 존재할 수 있도록 돕는 역할을 한다. *DNA*는 뉴클레오타이드(*nucleotide*)가 많이 연결된 고분자 유기물로서, 디옥시리보오스, 인산, 염기 한 분자씩의 연결체이다. 염기에는 *A*(*adenine*), *T*(*thymine*), *G*(*guanine*), *C*(*cytosine*) 네 가지의 종류가 있다. *A*와 *T*는 2중 수소 결합을 이루고 있으며, *G*와 *C*는 3중 수소 결합을 이루고 있다. 이에 따라 DNA는 이중 나선 구조를 형성하게 된다. [그림 2.1]은 이러한 *DNA*의 구조를 나타내고 있다.

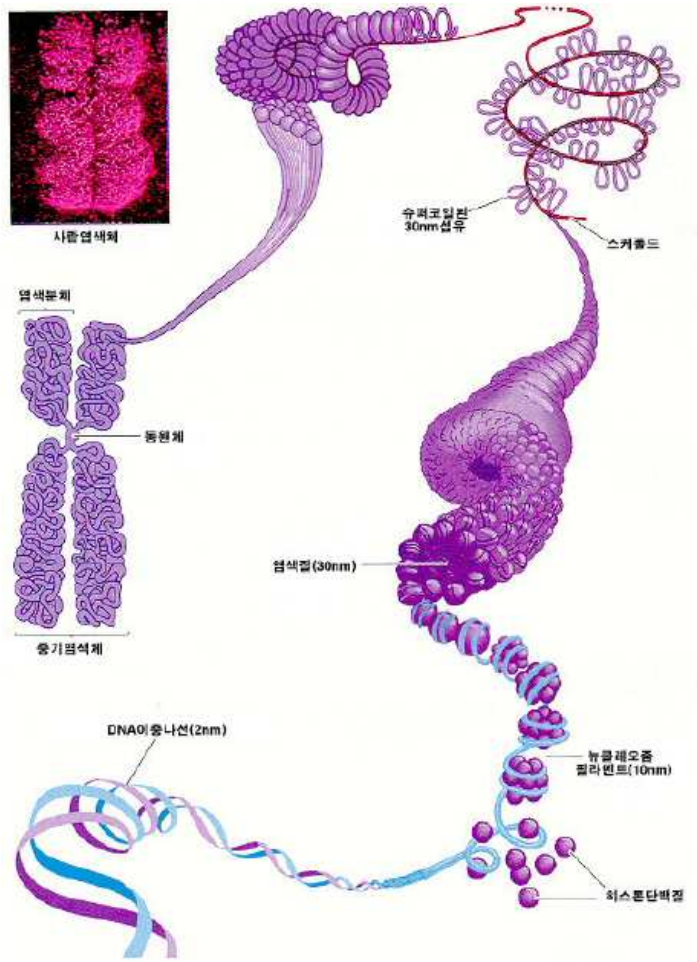


[그림 2.1] DNA 구조

*DNA*를 그 기능적 역할에 따라 엑슨(*exon*), 인트론(*intron*), *UTR* 부분으로 구분하기도 하는데, 엑슨은 생성하고자 하는 단백질 정보를 전송하는 역할을 한다. 유전자는 *DNA* 중에서 단백질 생성을 위한 정보를 전달하는 특정 부분만을 의미하는 것으로 정의된다. 유전자의 분포는 이산적이다. 유전체(*genome*)는 특정 개체의 전체 *DNA*를 의미한다. *DNA*의 길이와 염기 서열 구성은 생물체에 따라 다르며, 인간의 *DNA*는 약 30억 *bp*(*base pair*)로 이루어져 있다.

유전자로부터 발현 전송된 단백질 구성을 위한 데이터들은 전사(*transcription*)의 과정을 거쳐 *mRNA(messenger RNA)*를 형성한다. 유전자 데이터가 *mRNA*의 형태로 나타나는 것을 유전자 발현(*gene expression*)이라 정의한다. 여기에서 *mRNA*에 포함되어 있는 실제 단백질 구현 관련 정보가 번역(*translation*) 과정에 이용되면서 최종 산물인 단백질이 생성된다. 이렇게 만들어진 단백질은 생물체의 생명 현상을 유지하고 제어하기 위한 각종 역할을 하는 데에 사용된다. 즉, *DNA*의 부분집합인 유전자는 단백질 생성에 필요한 핵심 정보를 포함하고 있다고 할 수 있다.

[그림 2.2]는 앞서 언급한 염색체의 구조를 도식화한 것이다.


사람염색체
염색분체
동원체
중기염색체
슈퍼코일된
30nm섬유
스캐폴드
염색질(30nm)
DNA이중나선(2nm)
뉴클레오좀
필라멘트(10nm)
히스톤단백질

[그림 2.2] 염색체의 구조

유전자의 발현은 특정 상황에서 필요한 목적에 따라 단백질을 생성하기 위해 발생한다. 이 때 발현은 하나의 유전자에 의한 단독 발현일 수도 있고, 두 개 이상의 유전자에 의한 상호 발현일 수도 있다. 또한, 어떤 유전자의 발현으로 인하여 다른 유전자의 발현이 촉진되거나 억제될 수도 있다. 유전자의 발현 메커니즘은 매우 복잡한 구조로 이루어져 있으며, 기존의 분자 생물학적 접근으로는 이러한 메커니즘에 대해 분석하고 이해하는 것이 결코 쉽지 않다. 오늘날 인류는 인간 게놈 프로젝트에 의해 길이가 30억(*bp*)이나 되는 게놈 서열과 8만 여개로 추정되는 유전자들의 지도 및 이들에 관련된 분류 정보를 획득했다. 또한 30여종이 넘는 유기 생명체에 대한 *DNA* 염기 서열 정보를 완벽하게 밝혀냈다.

그러나 이러한 유전자 지도는 그 자체만으로는 실질적으로 무의미하다. 유전자 지도 자체만으로는 어떠한 유전자가 어떠한 단백질 생성에 관여하며, 이렇게 생성된 단백질이 어떤 세포 형성에 영향을 주고, 유기체 전체에 어떤 효과를 줄 수 있는지, 그리고 전체적인 의미에서 이와 관련된 유전자로는 어떤 것들이 있는가와 같은 물음에 대한 대답은 알 수가 없기 때문이다. 기존의 분자 생물학적 방법과 같이 한 번에 유전자 하나에 대한 접근을 시도하고 이로부터 획득한 데이터들을 수집해서 전체적인 의미를 도출하는 형태의 방법은 대량의 데이터를 포함하는 종합적인 분석에 적합하지 않다. 분자 생물학과 공학 기술을 결합하여 고안된 마이크로어레이 기술은 동시에 대량의 유전자에 대해 특정 조건에 따른 발현 정보를 관찰할 수 있게 해 줌으로써 이러한 분야에서의 효과적인 분석을 가능하게 하고 있다.

마이크로어레이란 형태상으로는 현미경 슬라이드 정도 크기의 유리판과 같이 투명하고 딱딱한 판 위에 수천 혹은 수만 개의 *DNA* 조각을 격자 유리 모양으로 가지런히 배열해 놓은 분자 생물학의 도구이다.

마이크로어레이는 그 형태상 특징에서도 나타나듯이 특정 조건에서 대량의 염기 서열이 보이는 경향을 관찰하기 위한 도구이다. 또한 대량의 유전자 발현 데이터를 획득하고 처리할 수 있도록 해 준다. 이는 기존 개별적인 한두 개의 유전자 경향 관찰에서 벗어나 개체의 대량 유전자가 보이는 발현 현상을 동시에 볼 수 있게 되었다는 점에서 매우 중요한 의의를 지닌다.

마이크로어레이를 개별 슬라이드 수준으로 이해해서는 안되고, 마이크로어레이의 제조, 마이크로어레이 실험, 마이크로어레이 실험 결과 추출, 그리고 이를 기반으로 한 데이터 분석까지를 마이크로어레이와 관련된 하나의 시스템으로 이해하여야 한다.

B. 마이크로어레이의 종류

실제로 존재하는 마이크로어레이 종류는 매우 다양하다. 그러나 특정 목적을 위해 마이크로어레이를 칭할 때, 그것이 의미하는 마이크로어레이의 종류는 크게 다음과 같이 분류될 수 있다.

(1) *cDNA* 마이크로어레이(*cDNA microarray*)

(2) 애피메트릭스 올리고뉴클레오타이드 마이크로어레이
(*Affymetrix oligonucleotide microarray*)

(3) 스팟티드 올리고뉴클레오타이드 마이크로어레이
(*Spotted oligonucleotide microarray*)

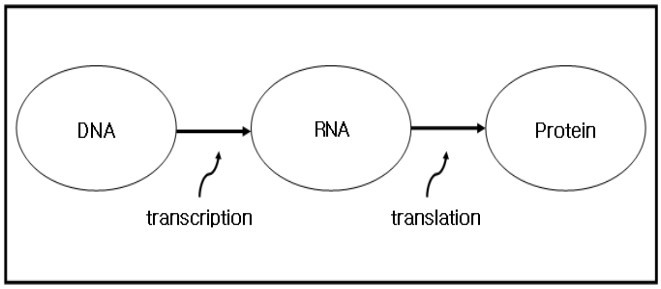
(4) 신세틱 올리고뉴클레오타이드 마이크어레이
(*Synthetic oligonucleotide microarray*)

이러한 분류는 마이크로어레이의 슬라이드에 부착하는 검출 *DNA*(*probe DNA*)가 어떻게 만들어지는가에 기반한 것이다. 검출 *DNA*의 종류에 따라서 마이크로어레이를 크게 *cDNA* 칩과 올리고 칩으로도 분류할 수 있다.

1. *cDNA* 마이크로어레이

cDNA 칩은 마이크로어레이의 슬라이드에 *cDNA*를 붙인 것으로, *cDNA*는 결과적으로 전체 염기서열 내의 유전자 가운데 단백질로 전사되는 부분에 해당한다.

다음의 [그림 2.3]은 *DNA*, *RNA* 그리고 단백질 사이의 관계를 나타낸 *central dogma*를 도식화한 것이다.



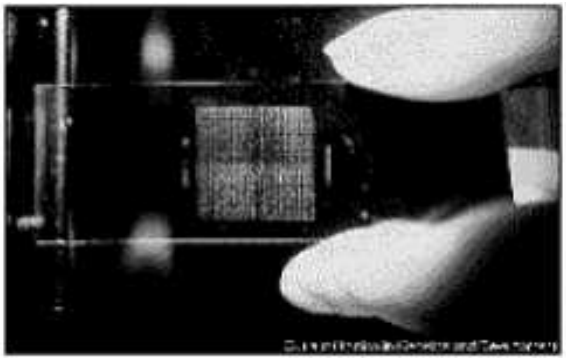
[그림 2.3] Central dogma

간단하게 표현된 [그림 2.3]의 도식은 단백질 합성에 매우 중요한 개념이다. 초기 *central dogma*의 개념이 알려지고 난 후, 이 이론은 거의 절대적인 것으로 받아들여졌다. 그러나 일부 바이러스들이 자신의 유전 정보를 *RNA* 상태로 보존하고 있다는 사실이 발견되면서 분자 생물학은 새로운 전기를 맞이하게 되었다. 이러한 종류의 바이러스들은 역전사 효소(*reverse transcriptase*)를 갖고 있으며 숙주의 내부에서 이 효소를 이용하여 자신의 *DNA*를 만들어 내고 복제를 하여 새로운 바이러스를 만들어 증식한다. 분자생물학자들은 이 역전사 효소를 분리하여 *mRNA*로부터 *cDNA*들을 만들어 내는 방법을 고안하였다. 이 방법을 이용하여 무작위로 분리된 *mRNA*로부터 *cDNA*들을 얻을 수 있게 되었고 이렇게 확보한 데이터들을 통합하여 대규모의 라이브러리를 구축할 수 있게 되었다. 이러한 *cDNA* 라이브러리를 이용하여 새로운 유전자를 발견할 수도 있게 되었는데 그 대표적인 방법으로는 *yeast-to-hybrid system*이 있다. 단백질의 발현을 연구하는 데 있어서 *RNA* 수준

에서의 조절을 시도할 때에도 역전사 효소를 이용하여 *cDNA*를 얻은 뒤, 이를 *PCR*(*polymerase chain reaction*)을 통하여 관찰하는 방법을 이용한다.

즉, 환자의 종양 샘플로부터 얻어진 세포에서 해당 마이크로어레이를 제작하기 위해 필요한 대량의 *DNA*를 추출하는 것은 어렵지 않으나, 이 *DNA* 중 단백질 생성을 위한 실질적 데이터들을 포함하고 있는 유전자 부분만을 선택하여 추출하는 것은 현재의 기술로 불가능하다. 따라서 획득한 *DNA* 전부를 이용하여 *mRNA*를 확보하고 이 *mRNA*를 역전사 효소를 이용하여 역전사시키면 정 방향 전사 때와는 반대로 그에 대응하는 *cDNA*를 얻을 수 있다. 이 *cDNA*는 *mRNA*가 생성될 때 정보를 전달해 준 정보력이 있는 유전자들과 똑같은 데이터만으로 이루어지게 된다. 따라서 *cDNA* 마이크로어레이의 검출 *DNA*도 이와 같이 역전사 효소를 사용하여 얻는다.

실제 마이크로어레이를 사용하여 측정하고자 하는 것이 조건에 따른 유전자의 발현량(*gene expression level*)이므로 마이크로어레이의 슬라이드에 부착한 것은 유전자를 검출할 수 있는 분자이어야 하는데, 이러한 맥락에서 *cDNA*는 매우 자연스러운 선택이다. [그림 2.4]는 실제 *cDNA* 마이크로어레이 형태이다.

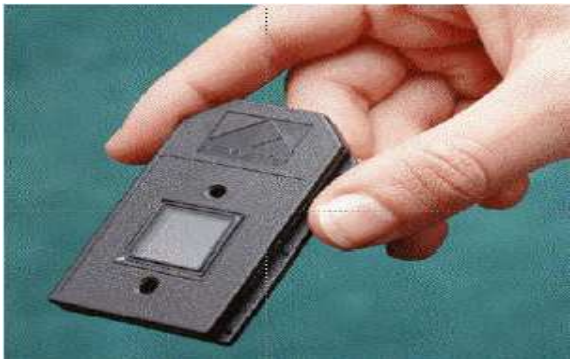


[그림 2.4] cDNA 마이크로어레이

2. 올리고뉴클레오타이드 마이크로어레이

올리고 칩은 이들 *cDNA* 가운데에서 각 유전자마다 고유한 서열을 가진 일정한 길이의 영역만을 선택해서 슬라이드에 붙인 것이다. 따라서 올리고 칩은 바이오인포메틱스의 결정체라고 할 수 있다. 왜냐하면 인간의 유전자 전체를 측정하기 위한 올리고 칩의 경우에는, 인간이 가진 전체 유전자들에 대해서 각 유전자가 갖는 고유한 일정 길이의 서열을 모두 찾아내야 하기 때문이다. 이론적으로 생각했을 때, 모든 생물체는 각 유전자를 특정한 상황에 맞추어 발현시키므로 각 유전자 고유의 영역이 있을 것이라고 생각하는 것이 전혀 근거 없는 일은 아니다. 그러나 모든 유전자에 대해서 다른 유전자와는 확연히 구분되는 각 유전자의 고유한 영역을 찾는다는 것은 결코 쉽지 않은 문제이다.

[그림 2.5]는 대표적인 올리고 칩인 애피메트릭스 올리고뉴클레오타이드 마이크로어레이의 형태이다.



[그림 2.5] 애피메트릭스 올리고뉴클레오타이드 마이크로어레이

올리고 칩도 여러 가지 종류로 분류할 수 있는데, 애피메트릭스 올리고뉴클레오타이드 마이크로어레이, 스팟티드 올리고뉴클레오타이드 마이크로어레이, 그리고 신세틱 올리고뉴클레오타이드 마이크로어레이가 그 대표적인 것들이다. 이와 같은 분류는 올리고 검출 *DNA*를 제작해서 슬라이드에 붙이는 방법에 근거한 것이다.

애피메트릭스 올리고뉴클레오타이드 마이크로어레이는 미국의 애피메트릭스사가 특허권을 보유하고 있는 *photo-lithography* 기술을 이용하여 각 위치에 염기를 한 층씩 쌓아 만든 올리고 칩이다. 그 제작 방법에 대해서는 이후에 좀 더 자세히 언급하도록 한다.

스팟티드 올리고뉴클레오타이드 마이크로어레이는 이미 제작되어진 올리고 *DNA*를 슬라이드에 붙이는 방식으로 제작된다. 즉, *cDNA* 마이크로어레이에서 *cDNA* 대신 올리고뉴클레오타이드를 검출 유전자로 사용한 것이다. 많은 수의 생물체에 대해 전체 *DNA*의 서열이 밝혀진 현재에는 특정한 영역에 해당하는 염기 서열을 만드는 것이 매우 쉬워졌기 때문에, 특히 오늘날에는 스팟티드 올리고뉴클레오타이드 마이크로어레이에 대한 선호도가 더욱 증가하고 있다.

이에 비해 신세틱 올리고뉴클레오타이드 마이크로어레이는 애피메트릭스 올리고뉴클레오타이드 마이크로어레이와 스팟티드 올리고뉴클레오타이드 마이크로어레이 그리고 *cDNA* 마이크로어레이를 혼합해 놓은 올리고 칩이다. 신세틱 올리고뉴클레오타이드 마이크로어레이는 슬라이드 위에서 검출 유전자를 쌓아 올린다는 점에서는 애피메트릭스 올리고뉴클레오타이드 마이크로어레이와 같다. 그러나 쌓아올리는 방법으로 염기 하나 하나를 특정한 위치에 붙인다는 점에서는 *cDNA* 마이크로어레이와 같고 그 검출 유전자의 길이는 스팟티드 올리고뉴클레오타이드 마이크로어레이와 비슷하다. 이 칩은 에이질런트(*Agilent*)사에서 독점적으로 제작되고 있는데 그 과정에서 잉크젯 프린터에 적용되는 기법이 사용된다. 슬라이드 위에서 필요한 염기를 특정한 위치에 한 층씩 뿌려 원하는 서열이 되도록 키우는 방식을 사용한다.

앞서 언급한 네 종류의 마이크로어레이를 다른 기준으로 분류해본다면, 슬라이드 위에서 검출 유전자를 합성하는 형태의 마이크로어레이인 애피메트릭스 올리고뉴클레오타이드 마이크로어레이와 신세틱 올리고뉴클레오타이드 마이크로어레이, 그리고 이미 외부에서 만들어진 검출 유전자를 특정한 위치에 고정시키는 형태의 마이크로어레이인 스팟티드 올리고뉴클레오타이드 마이크로어레이와 *cDNA* 마이크로어레이로 분류할 수도 있다. 현재에도 충분히 다양한 형태의 마이크로어레이가 존재하고 있으나, 시간이 흐를수록 더욱 많은 형태의 발전된 마이크로어레이가 제작될 것이다.

C. 마이크로어레이의 제작 방법

1. 애피메트릭스 올리고뉴클레오타이드 마이크로어레이

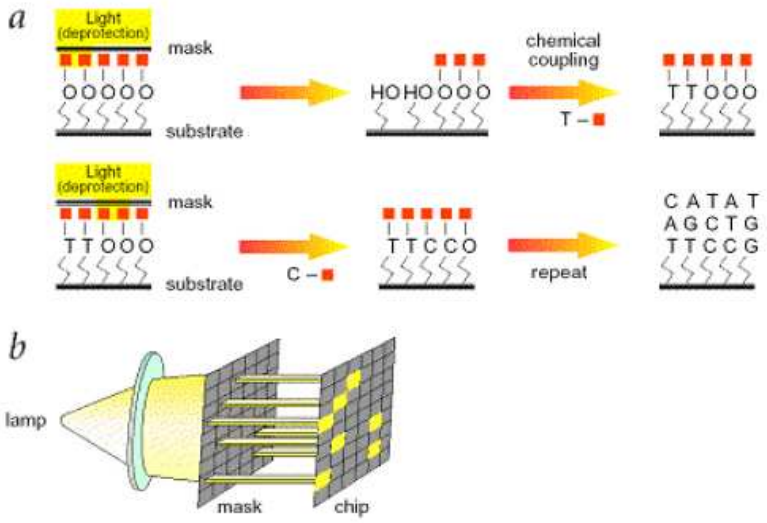
앞서 언급한 것과 같이 애피메트릭스 올리고뉴클레오타이드 마이크로어레이는 미국의 애피메트릭스 사가 독점적으로 제조, 공급하고 있는 특정한 형태의 올리고 칩을 의미한다. [그림 2.6]에서 검은색 플라스틱 중앙에 위치한 정사각형의 창(*window*)과 유사한 모양을 하고 있는 부분이 검출 유전자가 찍혀 있는 부분이다. 이 카트리지는 이 상태로 애피메트릭스 사의 결합반응기(*hybridization chamber*)에 들어가게 된다. 창의 위와 아래 부분에 있는 홈으로 형광물질이 착색된 *mRNA* 용액이 흘러들어가서 흘러나오는 방식으로 검출 유전자와 *mRNA* 사이의 반응이 정확하게 발생하도록 되어 있다. 애피메트릭스 올리고뉴클레오타이드 마이크로어레이는 제조 방법이 매우 독특하고 정교하기 때문에 제작하기가 쉽지 않다. 그러나 이 칩의 성능이 매우 뛰어나기 때문에 많은 연구자들이 이 마이크로어레이를 이용하고 있다.

애피메트릭스 올리고 칩을 제조하는 방법은 다음과 같다. 우선 슬라이드의 특정한 위치에 어떤 유전자를 둘 것이지를 결정했다고 가정하자. 슬라이드에 *DNA*가 쉽게 붙어있지 않기 때문에 슬라이드에 *DNA*가 붙어있을 수 있도록 처리를 한다. 이 과정은 다른 마이크로어레이도 동일하게 적용된다. 이 층을 가-층이라고 하자. 다음으로 이 위에 접착 물질을 한 번 더 처리하여 가-층의 물질과 나-층을 만든다. 중요한 사실은 나-층의 물질은 가-층의 물질과 화학적으로 결합하고 있으나 이 화학 결합은 빛을 쪼이면 끊어져서 나-층의 물질은 분리될 수 있다는 것이다. 즉 하나의 층이 벗겨지면서 묻혀있던 가-층이 드러나는 것이다. 이렇게 드러난 가-층에 원하는 염기를 결합시킨다. 염기는 기본적으로 뼈대(*backbone*)를 중심으로 안쪽과 바깥쪽으로 나눌 수 있는데, 이 때 가-층과 결합하는 바깥쪽에는 앞에서 사용한 빛에 반응하는 물질이 붙어있기 때문에 염기가 가-층의 모든 영역에 결합하면 그 위

는 다시 빛과 반응하는 나-층의 물질로 덮인 상태가 된다. 이 과정을 반복하면서 염기를 한 층씩 쌓아가는 방법으로 애피메트릭스 올리고뉴클레오타이드 마이크로어레이를 제작한다.

슬라이드의 특정한 위치에만 빛을 쪼이도록 하는 데에는 반도체의 기판 등을 만들 때 사용하는 *photo-lithography*라는 기술을 사용한다. 이 기술은 스탠실을 축소시킨 것과 유사하다. 즉, 광원이 되는 레이저의 앞에 스크린을 두고 이 스크린의 특정한 위치에만 구멍을 뚫어 놓으면 그곳으로만 빛이 투과하게 되고 레이저의 고휘도성을 이용하면 아주 조그만 위치에만 빛을 조준하는 것이 가능하다.

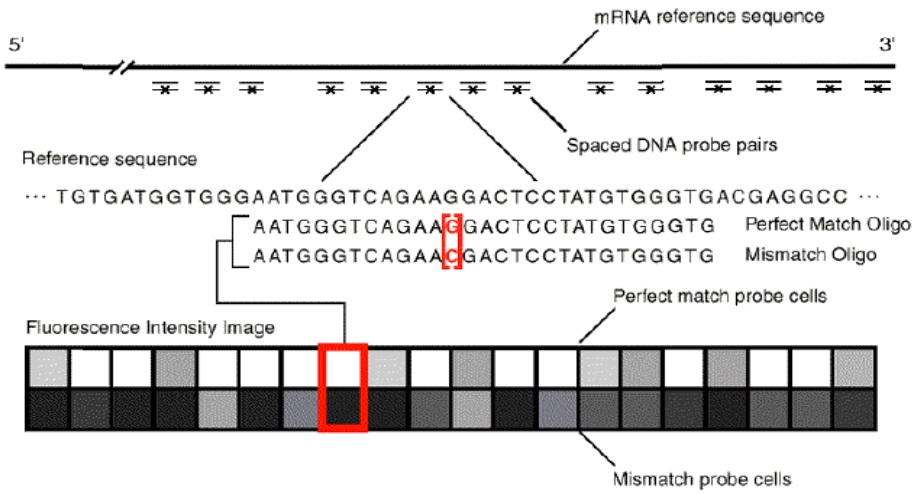
[그리 2.6]은 애피메트릭스 올리고뉴클레오타이드 마이크로어레이의 제작 방법을 나타내고 있다.



[그림 2.6] 애피메트릭스 올리고뉴클레오타이드 마이크로어레이의 제작

그런데 이렇게 많은 수의 미세한 영역에 직접 염기를 하나씩 쌓아서 검출 *DNA*를 만든다고 한다면 한 층을 쌓는 데에도 상당히 많은 조작이 필요하다. 실제로 판매가 될 정도의 가격으로 칩을 제작하기 위해서는 하나의 단위 영역에 쌓는 층의 개수가 많지 않아야 한다. 층의 수가 늘어날수록 조작의 횟수도 늘어나고 비용도 증가한다. 따라서 애피메트릭스에서는 하나의 단위영역에 쌓는 층의 개수를 25개로 제한했다. 그러나 염기의 서열은 결코 완전히 무작위로 나타나지 않고 대부분의 많은 *DNA* 서열은 서로 비슷하다. 그러므로 25개의 염기는 충분하지 못하다. 하나의 유전자를 하나의 단위 영역으로 나타내지 않고 다수의 단위 영역으로 분할하여 나타내는 방법을 사용함으로써 이 문제는 해결된다. 마지막으로, 어떤 염기서열을 쌓아올려야 하는가에 대해 결정해야 하는 문제가 남는다. 올리고 칩의 특성 상 쌓아올려야 하는 염기 서열은 하나의 유전자에 대해 고유한 것이어야 한다. 애피메트릭스는 불완전한 짝(*mis-match*) 검출 유전자를 사용하여 이 문제를 해결하였다. 슬라이드에 검출 유전자를 쌓아올릴 때, 한 단계 아래 라인에 원래 쌓아올리려 했던 염기 서열(*perfect match*)의 가운데 염기 하나를 다른 염기로 변경해서 쌓아올린다. 이것은 이종 결합(*cross-hybridization*)의 수준을 측정해서 반영하기 위한 것이다. 즉 염기 서열은 비록 서로 다른 유전자에서 나왔다고 하더라도 서로 매우 비슷하기 때문에, 근본적으로 의도했던 짝이 아닌 다른 짝에서 나온 염기서열과 검출 유전자가 결합을 하는 경우가 없지 않다. 따라서 이러한 경우가 어느 정도인지를 조사하여 이를 전체 데이터에 반영하기 위해 애피매트릭스는 불완전 짝의 아이디어를 칩에 반영하였다. 이는 이것이 데이터의 분석 과정에서 적지 않은 혼동을 야기하고 있다는 점도 결코 간과할 수 없음을 언급하기 위함이다.

[그림 2.7]은 불완전 짝 검출 유전자의 적용 사례를 보여주고 있다.



[그림 2.7] 불완전 짝 검출 유전자의 적용 예

2. *cDNA(complementary DNA)* 마이크로어레이

cDNA 마이크로어레이는 *cDNA*가 직접 슬라이드에 붙어 있는 형태의 마이크로어레이를 의미한다. 슬라이드의 가운데에 바둑판 모양으로 배열된 점이 각 *cDNA* 스팟(*spot*)이다. 애피메트릭스 올리고뉴클레오타이드 마이크로어레이에서와 같이 *DNA*가 슬라이드에 부착될 수 있도록 처리를 하고 그 위에 *cDNA*를 얹은 뒤 *UV* 크로스 링크(*UV cross link*) 등의 처리를 하면 슬라이드에 *cDNA*가 붙는다. 그런데 이와 같이 *cDNA*를 슬라이드에 붙이면 한 쌍의 *cDNA*가 이미 결합이 되어있는 상태이기 때문에 이 상태에서는 샘플에서 *mRNA*를 추출하여 이와 결합 반응(*hybridization*)을 수행한다고 하여도 *mRNA*와 검출 유전자가 결합할 여지가 없게 된다. 따라서 *cDNA*를 슬라이드에 붙인 뒤에는 이 슬라이드를 담고 있는 용액의 온도를 높여서 이미 결합 되어있는 두 개의 *cDNA* 가닥이 서로 분리되도록 한 다음 한 가닥을 완전히 제거해야 한다.

현재 *DNA*로부터 *cDNA*를 직접 구할 수 있는 방법은 기술적으로 존재하지 않는다. 불행하게도 전체 *DNA*에서 유전자 부분을 직접적으로 구분하거나 그 내부 구조를 파악하는 작업 역시 완벽하지 못하다. 그러나 세포는 그 작업을 완벽하게 수행한다. 그러므로 인위적으로 *cDNA*를 얻고자 할 때에는 세포가 가지고 있는 작업 방식들을 이용한다. 그것은 역전사(*reverse transcription*)라는 과정인데, 유전자가 세포핵에서 발현된 후 핵막을 통과하여 세포질로 나오는 과정에서 엑슨으로부터 전사된 부분만 남겨두고 인트론으로부터 전사된 부분을 제거한 다음 *mRNA*를 만든다. *mRNA*를 역으로 *DNA*로 만든 것을 *cDNA*라고 하고 이 과정을 역전사라고 하는데 이 과정은 현재 충분히 알려져 있다. 이 과정을 여기서 기술하는 이유는 이 모든 과정이 전체 유전자에 대해서 적용된다는 사실을 언급하기 위해서이다. 인위적으로 하나의 특정한 *cDNA*를 선택하여 처리할 수 있는 방법은 없다. 즉, 전체 유전자에 대해서 *mRNA*를 만들고 역전사를 거쳐서 *cDNA*[1)]를 통째로 만든 후에 서열 조사를 거쳐야 원하는 부분을 확보할 수 있다. 주목할 점은 *DNA*의 전체 서열이나 특정한 유전자의 서열 등에 관한 사전 지식이 필요 없다는 사실이다. 이러한

1) 이를 *cDNA* 클론이라 한다.

특성은 올리고 칩 계열의 다른 마이크로어레이 속성과는 전혀 다른 것이다. 뿐만 아니라 원칙적으로는 한 번의 조작을 거쳐서 모든 세포의 *cDNA* 조합을 완전히 확보할 수 있기 때문에 이를 이용하면 한 종의 *cDNA*를 거의 모두 포함하고 있는 마이크로어레이를 제작할 수도 있다.

역전사를 이용하여 *cDNA* 라이브러리를 만드는 과정은 일반적으로 다음과 같이 진행된다. 먼저 세포막 내부에 들어있는 내용물을 모두 쏟아낸다[2]. 다음으로 *mRNA*를 정제한다. 세포핵에서 *DNA* 전사가 일어나지만 이때 생성되는 *RNA*는 단순히 초기 *RNA*(*primary RNA*)라 한다. 이 초기 *RNA*가 세포핵막을 거쳐 세포질로 나오면서 몇 가지 과정을 거치게 된다. 초기 *RNA*에서 *DNA*의 인트론으로부터 전사된 부분을 들어내는 과정[3]이 진행된다. 이렇게 해서 엑슨 부분에 의한 *RNA*만을 추출한다. 그리고 초기 *RNA*의 5′ 부분에 모자를 씌우고[4] 3′ 부분에는 염기 A를 많이 붙인다[5]. 이와 같은 과정을 통해 최종적으로 생성된 *RNA*를 *mRNA*라 한다. 세포내에서 역전사가 진행될 때 작용하는 효소가 역전사 효소인데 이 효소가 정상적으로 작용하기 위해서는 프라이머라는 조그만 *RNA*조각이 *mRNA*에 붙어 있어야 한다. 이 경우 프라이머는 단순히 *mRNA*의 꼬리쪽에 달린 여러 *A* 염기에 붙어있기만 하면 되기 때문에 프라이머는 염기 *T*로만 구성되어 있으면 된다. 이렇게 프라이머가 붙은 *mRNA*에 역전사 효소가 작용을 하면 *mRNA*를 기준으로 *cDNA*가 *mRNA*를 따라서 생성된다. 이 때에는 *mRNA*가 *cDNA*에 달라붙어 있는 상태인데 여기에 *RNase H*라는 효소(*enzyme*)를 가하면, *mRNA-cDNA* 쌍에서 *mRNA* 부분만 녹는다. 다음으로 *DNA* 전사 효소가 여기에 작용하여 *cDNA* 쌍을 만든다. *DNA* 전사 효소가 정상적으로 동작하려 하더라도 프라이머가 필요한데 이 때에는 아직 완전히 녹지 않고 부분적으로 남아 있는 *mRNA*의 조각들이 프라이머의 역할을 한다. 이렇게 생성된 *cDNA*쌍을 플라즈미드나 바이러스 벡터에 끼워놓은 다음 대량으로 증가시키는데, 하나의 세포에서 만들어진 모든 *cDNA*를 개별적으로 플라즈마나 벡터에 끼워놓은 전체를 가리켜 *cDNA* 라이브러리라고 한다.

2) 이 과정을 세포 용해(*lysis*)라 한다.

3) 이 과정을 *RNA* 중첩(*RNA splicing*)이라 한다.

4) 이 과정을 뚜껑 씌우기(*capping*)라 한다.

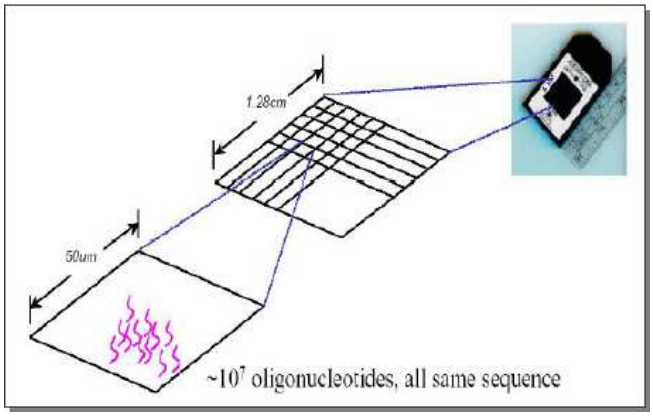
5) 이 과정을 *poly-A tail*라 한다.

마지막으로 라이브러리에 들어있는 각 *cDNA*를 벡터로부터 분리하여 그 양을 *PCR*(*Polymerase Chain Reaction*)을 이용해서 증가시킨다. 이는 *cDNA* 마이크로어레이를 만드는데 필요한 모든 *cDNA*를 확보하기 위함이다. 이렇게 대량의 *cDNA*가 생성되었다면 이것들을 바둑판 모양으로 슬라이드에 붙여야 하는데 이 과정은 마이크로어레이어(*microarrayer*)라는 기계가 한다. 기본적인 과정은 만년필촉과 같은 핀이 증폭된 *cDNA* 용액이 담긴 피펫에서 일정량의 용액을 집은 후 컴퓨터 연산을 통해 정해진 위치에 해당 용액을 찍는 방식으로 진행된다. 많은 양의 검출 유전자를 슬라이드에 붙여야 하기 때문에 정확하고 미세한 조작이 필요해서 컴퓨터가 제어하도록 한다. 물론 이 과정에서 하나의 *cDNA*를 정해진 위치에 찍은 뒤에는 잔여물이 남지 않도록 핀을 씻고 말리는 과정이 몇 번에 걸쳐 반복된다. 그렇지 않으면 기존의 기대했던 *DNA*가 아닌 다른 *DNA*가 해당 위치에 남아 있어서 데이터의 오류를 증가시키기 때문이다. 이러한 현상을 일반적으로 이종 결합(*cross-hybridization*)이라고 한다.

D. 마이크로어레이 데이터의 특징

위에서 살펴본 바와 같이 마이크로어레이 슬라이드의 크기는 실제로 매우 작다. 따라서 미세하고 정교한 작업을 하는 제작 과정과 실험 과정에서 물리적 잡음(*noise*)이 발생할 가능성을 결코 배제할 수 없다. [그림 2.8]은 마이크로어레이의 실제 크기를 나타내고 있다. 그리고 실험 과정에서 이종 결합으로 인한 잡음 발생 가능성 역시 간과할 수 없다.

마이크로어레이를 통한 실험은 가상 시뮬레이션이 아닌 실제 상황이다. 마이크로어레이는 실제 생물학적 샘플 데이터를 이용하여 실질적인 반응을 통해 결과 데이터를 확보하는 도구이다. 따라서 마이크로어레이를 이용하여 특정한 조건에서 관찰된 특정 유전자의 발현은 실제 상황에서도 그대로 적용될 수 있다. 예를 들어 특정 세포에서 추출한 유전자 정보를 이용하여 마이크로어레이 실험을 하였을 때, 특정 유전자들이 해당 조건에 대해 현저한 발현을 보였다면 이는 그 유전자들이 이 세포를 구성하는 단백질 발현에 실제로도 직간접적인 연관을 갖는다는 것을 의미한다.



[그림 2.8] 마이크로어레이 크기

올리고 칩은 정상적인 염기 서열로 이루어진 *perfect match probe*와 염기 서열 하나가 다른 *mismatch probe*가 *probe pair*를 이루고 있는 형태로 되어있다. 따라서 올리고 칩을 이용하면 하나의 염기 변화에 의한 다양성(*SNP*, *single nucleotide polymorphism*)을 연구할 수 있다. 즉 한 염기에 생긴 돌연변이(*point mutation*)를 탐색할 수 있음을 의미하는데, 많은 종류의 암이나 유전 질환 발달이 특정 유전자에 발생한 작은 돌연 변이에 의해서 유발되므로 이러한 특징을 이용하여 지금까지 밝혀진 암 관련 유전자를 만든다면 한 번의 실험으로 쉽게 돌연변이를 찾을 수 있다. 결국, *cDNA* 칩은 유전자 발현 양상의 연구에 주로 쓰이고 환자의 질병과 연관된 검사는 어려운 반면, 올리고 칩은 유전적 다형성을 감지하여 유전병을 진단하고 예측할 수 있게 해 준다.

마이크로어레이를 이용하여 얻게 되는 유전자 발현 데이터는 스캐너에 의해 읽혀지고 통계학적 계산에 의해 [그림 2.9]과 같이 수치로 변환된 후 일반적으로 행렬의 형태로 표현된다. 이 행렬에서 각 열(*column*)은 하나의 샘플을 나타내고 각 행(*raw*)은 각 샘플에 대한 속성으로서의 유전자를 나타낸다. 마이크로어레이의 특성상 각 열의 총 수는 수십에서 수백으로 상대적으로 매우 적고, 각 행의 총 수는 수천에서 수만 이상으로 상대적으로 매우 많다[6][7].

[그림 2.9] 행렬로 나타낸 유전자 발현 데이터

m samples

n genes

GeneId	Condition 1	Condition 2	...	Condition *m*
Gene1	103.02	58.79	...	101.54
Gene2	40.55	1246.87	...	1432.12
...	...	...	...	...
Gene *n*	78.13	66.25	...	823.09

E. 마이크로어레이 기반의 기존 암 분류 연구

1. 현대 암 진단 방법

현대적인 암 진단은 종양의 해부학적 발생 위치를 정확히 결정하는 데에서부터 출발한다. 이를 기반으로 발생 부위에 적합한 진단을 하고 처방을 하게 된다.
따라서 암에 대한 정확한 진단은 현대 의학에 있어 암의 치료를 위해 매우 중요한 역할을 한다. 그러나 이러한 방법들은 여러 다른 종양 사이의 형태학적 유사성으로 인해 그 정확성이 떨어지고, 후행적인 진단이 될 가능성이 높다는 단점이 있다. 형태적으로 동일해 보이는 암이라 할지라도, 어떤 경우에는 그 진행 속도가 상당히 빨라서 매우 공격적인 치료를 해야 하는 반면, 어떤 경우에는 그 진행 속도가 너무 느려서 오히려 치료를 하지 않는 것이 더 나을 수도 있다.

2. 분자 생물학을 이용한 암 분류

분자 생물학을 이용한 암 분류는 앞서 언급한 문제점들을 해결할 수 있는 가능성을 지니고 있다. *DNA* 칩을 활용한 유전자 발현 분석 기술은 암 분류 응용에 유망한 기술로, 동일한 종류의 암은 그 종류마다 고유한 유전자 발현 패턴을 지닌다는 사실을 전제로 한다. 경우에 따라서는 이제까지 알려지지 않았던 암의 종류를 예측할 수도 있다. 오늘날 가장 이슈가 되고 있는 부분 중 하나는 특정 암을 식별할 수 있는 유전적 표지(genetic markers) 즉, 정보력이 있는 유전자를 선택하고 이것을 이용하여 암을 정확히 분류 할 수 있는 시스템을 구축하는 것이다. 이 과정에서 선택된 정보력이 있는 유전자 리스트는 그 염기 서열 분석과 더불어 정확한 분류 이후 신약 제조와 같은 치료를 위한 과정에 능동적인 동기를 부여할 수 있으므로 이 역시 간과할 수 없는 부분이다[8].

3. 마이크로어레이를 이용한 암 분류 연구 분야

마이크로어레이 데이터를 이용한 암 분류 연구는 다음과 같이 크게 클래스 발견(class discovery: cluster analysis)과 클래스 예측(class prediction: discriminant analysis) 두 부분으로 분류될 수 있다.

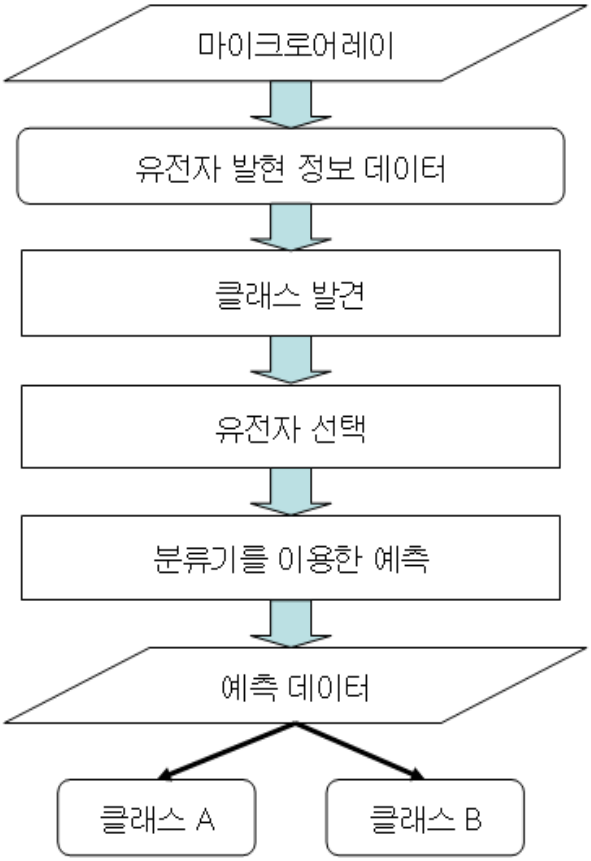
[표 2.1]은 마이크로어레이를 이용한 두 가지의 대표적인 암 분류 연구 분야에 대한 몇 가지 특징을 나타내고 있다.

[표 2.1] 마이크로어레이를 이용한 암 분류 연구 분야

Class Discovery	*Class Prediction*
현재 알려져 있지 않은 새로운 종류의 암을 발견하는 것	새로운 샘플을 기존의 알려져 있는 클래스로 정확히 지정하는 것
Hierarchical Clustering Self Organizing Map K-means Clustering Delta Clustering	Artificial Neural Network Support Vector Machine Discriminant Analysis K-Nearest Neighbor
사전 지식이 이용되지 않음	정의되어 있는 클래스 정보에 기초하여 새로운 샘플을 적절한 클래스로 할당함

4. 요구되는 최소 시스템의 개요

이렇게 마이크로어레이 데이터를 이용하여 암 분류를 하기 위해 요구되는 최소한의 시스템 개요는 [그림 2.10]과 같다.



[그림 2.10] 요구되는 최소 시스템의 요구

우선 마이크로어레이를 이용하여 해당 종양 샘플에 대해 특정 조건에 따른 유전자 발현 데이터를 확보한다. 이 마이크로어레이 데이터를 기반으로 클래스 발견 작업을 수행한다. 기존에 알려져 있는 부 클래스가 있다면 해당 정보를 사용할 수 있다. 클래스 발견 작업이 완료되면 각 클래스로 분류하는 데에 있어 극단적으로 큰 연관을 갖는 가상의 유전자를 이상 유전자 모델(ideal set)로 결정한다. 이 이상 유전자 모델의 정보를 기준으로 분류하는 데에 큰 관련을 갖는 유전자를 선택하는 작업을 한다. 이렇게 선택된 유전자 리스트를 기반으로 분류기 혹은 예측기를 구현하다. 마지막으로 이 분류기 혹은 예측기를 이용하여 새로운 샘플 데이터가 어떤 클래스에 해당하는지 그 여부를 결정하는 분류와 예측 작업을 한다.

5. 마이크로어레이 기반 암 분류 연구 사례

마이크로어레이 데이터를 이용한 기존의 암 분류 연구 중 분자 생물학적 관점을 충분히 반영한 연구 사례와 전산학적 관점을 충분히 반영한 대조적인 연구 사례 두 가지를 제시하고 두 연구의 장점과 단점을 언급한다.

*a. Golub*의 방법

*Golub*은 지난 30년간 진행되어 온 기존의 암 분류화 작업을 전체적이고 시스템적이며 일반적인 접근을 시도하려 하였다. 당시에는 새로운 클래스를 정의하고 알려진 클래스에 종양(*tumor*)을 할당하기 위한 일반적인 접근이 없었다. *Golub*은 *DNA* 마이크로어레이에 의한 유전자 발현 정보를 관찰하는 것을 기반으로 암 분류에 대한 일반적인 접근을 시도하는 방법을 급성 백혈병에 적용하여 그 가능성을 타진하였다. 기존에는 암의 치료를 위해서 특성화된 목적 치료(*target therapies*)를 시도했다. 따라서 정확한 암 종류를 판별하는 것이 매우 중요한 과제였다. 대상 암이 어떤 종류의 암인지를 정확히 파악해야 해당 암에 대한 특성화된 치료를 할 수 있기 때문이다. 과거에는 이렇게 특정 암을 분류하는 데에 일반적으로 종양의 외적

특징(*morphological appearance of the tumor*)을 활용하였다. 그러나 이러한 방법은 제약이 따를 수밖에 없다. 형태학적인 외적 양상은 매우 비슷하나 치료에 서로 다르게 반응하는 종양들이 발견됨으로써 이러한 방법의 한계가 입증되었다. 따라서 분자 수준의 외적 표지에 의한 기존의 방법으로는 정의되지 않은 중요한 부 클래스(subclass)를 놓치기 쉽다. 실례로 신장 암(*prostate cancer*) 종양에서 동일한 외적 양상이 발견되는 경우 이것이 상당히 다른 의료적 접근을 요구하는 경우가 많다. 기존에는 이러한 분류화 작업의 보완을 위해 효소를 이용한 분석 기법이 소개되었고, 이것이 발전하여 분류화의 기틀을 마련하였다. 최근에는 급성 백혈병의 특정 부 클래스가 특정 염색체와 연관되어 있다는 사실이 밝혀지면서, 유전자 정보를 이용한 분류화가 시도되기에 이르렀다.

*Golub*이 제안한 암 분류화 과정은 다음과 같다.

우선, 부 클래스들로 분류를 시도하기 위해 대상 클래스의 조건에 따른 종양 샘플에 대한 마이크로어레이 데이터를 확보한다. 이 과정에서 마이크로어레이로부터 얻어진 조건에 따른 각각의 유전자 발현 정보를 해당 차원의 공간에 사상한 다음 우선적으로 많은 유전자가 모여 있는 곳을 조건에 따라 유사한 반응 패턴을 지니고 있는 클러스터로 볼 수 있다. 여기서 여러 조건에서 나타난 발현 정보가 몇 개의 일정한 레벨로 나타나는 유전자들을 찾는다. 이러한 유전자는 몇 가지 레벨의 형태로 일정하게 분산이 적은 상태로 반응을 보여야하고, 여기서 새로운 클래스의 정의 즉, 클래스 발견을 시도할 수 있다. 이미 대상 클래스의 부 클래스가 알려져 있는 경우, 어느 샘플이 어떤 부 클래스에 해당되는지 병리적 분석(clinical diagnosis)을 통해 데이터 확보 순간부터 그 여부를 알 수 있기 때문에 이 때에는 적절한 이상 유전자 모델을 가정하고 좀 더 이상 유전자 모델에 가까운 정보력이 있는 유전자를 선택하도록 한다. 해당 이상 유전자 모델이 위치하는 곳에 클러스터가 형성되어 있을 경우 이상 유전자 모델로부터 가장 가까운 유전자들을 순차적으로 선택하면 그것들이 곧 정보력이 있는 유전자들이 된다. *Golub*은 이 방법을 사용하였고, 이상 유전자 모델의 속성을 무작위로 변경하면서 그 위치를 이동시켰다.

다음으로는 선택된 정보력이 있는 유전자들을 긍정적으로 평가받은 순서대로 정렬시킨다. 선택된 유전자들을 행렬 형태로 정렬시키고, 각 샘플에 대해 교차 입증 작업(*cross validation*)을 한다.

교차 입증 작업 과정은 다음과 같다.

*Golub*의 실험에서와 같이 38개의 샘플을 이용하여 예측기를 구현한다면, 38개의 샘플 중 하나의 샘플을 택하여, 나머지 37개의 샘플에 대해 각 유전자 발현 정보의 차이를 계산하여 가중치 합을 구한다. 가중치는 서열화 된 유전자 순서에 따라 높은 값을 할당하고 차이는 절대값을 사용하여 양수화한 뒤 계산한다. 클래스 *A*(*AML*)방향과 클래스 *B*(*ALL*)방향을 따로 구분하여 계산을 수행하며, 이 둘을 비교하여 그 합이 더 큰 쪽으로 부 클래스 할당 여부를 판정한다. 옳은 예측인 경우에 V_{win}을, 옳지 못한 예측인 경우에 V_{lose}를 부여하고 다음의 정의에 의거하여 *PS* 값을 구한다. 예측이 옳거나 옳지 못하다는 판단은 샘플 데이터 확보 시부터 이미 주어져 있는 병리적 분석 결과에 따른다.

$$PS(Prediction\ Strength) = (V_{win} - V_{lose})/(V_{win} + V_{lose}) \qquad (2.1)$$

모든 각각의 샘플에 대해 이를 수행하여, *PS*의 값이 특정 값 이하이면 스래시홀드(*threshold*)를 적용하여 해당 샘플을 제거한다. 이러한 과정을 통해 양질의 샘플과 정보력이 있는 유전자들로 새로운 행렬을 구축하게 된다. 이것은 예측기로 사용된다.

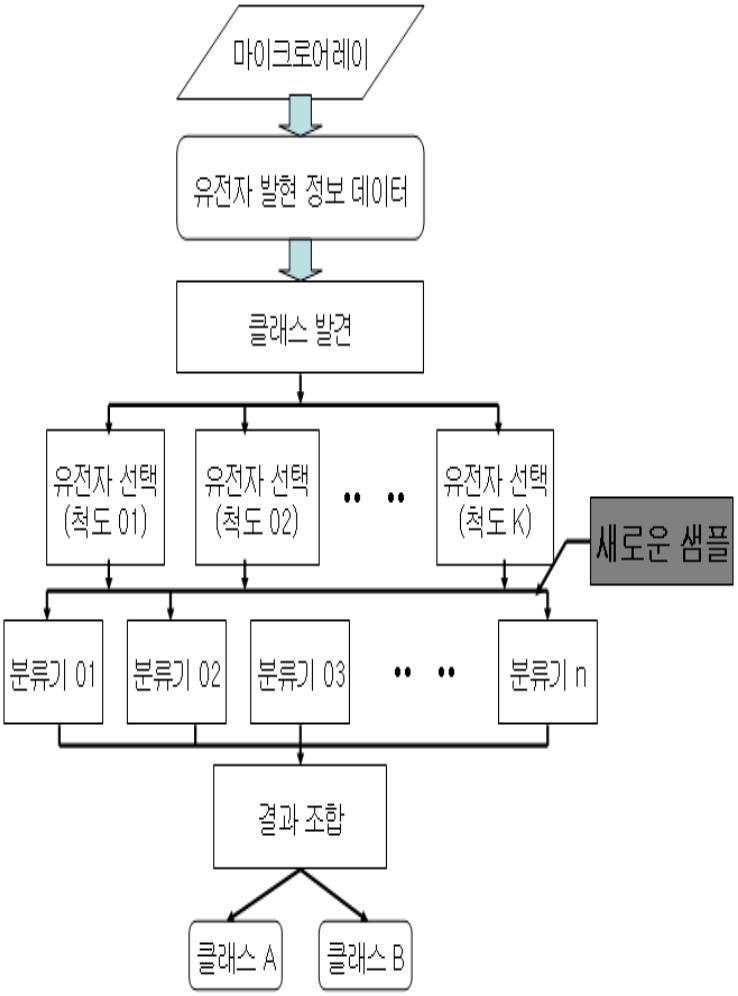
여기까지의 작업으로 생성된 것이 예측기이다. 이제 새로운 샘플 데이터를 입력하여 그 새로운 샘플이 어떤 부 클래스에 속할 가능성이 높은지를 추측할 수 있도록 한다. *Golub*은 새로운 데이터의 입력으로 이 예측기의 수행 성능을 측정하기 위해 34개의 샘플 데이터를 사용하였다. 입력된 샘플 데이터는 36번의 유전자 대조를 통해 각 반복에서의 유전자와의 차를 구하여 가중치 합을 계산하는 데에 사용되고, 이렇게 계산된 값을 기반으로 더 높은 값을 갖는 클래스로 해당 샘플을 할당한다. 이 경우에도 *PS*를 구하고 *PS*가 낮은 경우는 판정을 보류하도록 한다.

예측기에 사용된 정보력이 있는 유전자 리스트는 특정 유전자가 급성 백혈병에 관여하고 있음을 나타내고, 또한 두 가지 부 클래스에서 매우 상이하게 발현하고 있으므로 유전 질환 치료를 위한 병인론(*pathogenesis*) 연구에 직접적인 동기를 부여할 수 있음을 암시한다. 따라서 이러한 클래스 예측 방법론은 종양들 사이에서 측정 가능한 구분에 적용될 수도 있는데, 그러한 구분은 해당 전립선암이 무통성 종양인지 여부 혹은 유방암이 기존 화학 요법에 효과적으로 반응할지의 여부와 같은 미래의 임상 성과에 긍정적으로 관여할 수 있음을 암시한다. 따라서 이 정보력이 있는 유전자들의 리스트가 보다 정확해야 한다. 이러한 맥락에서 *Golub*의 방법론은 정보력이 있는 유용한 유전자를 별도로 추출해 낼 수 있다는 점이 매우 긍정적이다. 그러나 이렇게 별도로 분리할 수 있는 유용한 유전자들의 리스트가 실질적으로 병인론 등의 치료 관련 연구에 동기를 부여하기 위해서는 상당한 수준의 신뢰성이 바탕이 되어야 한다. 그럼에도 불구하고 *Golub*의 방법론으로는 정보력이 있는 유전자를 추출하기 위해 단일 유사성 척도를 사용했기 때문에 척도를 변경할 때마다 정보력이 있는 유전자의 순위가 변경된다는 단점을 갖고 있다. 실질적으로 신뢰성 있고 일관된 정보력이 있는 유전자를 추출하지 못할 가능성이 높다. 또한 클래스 발견 부분에서의 데이터 랜덤 처리에 적지 않은 비용이 요구된다. 마지막으로 상대적으로 분류 성능이 낮다는 단점이 있다.

b. 다중 분류기 시스템을 이용한 방법

마이크로어레이를 이용한 암 분류에서 최근의 전산학적 접근으로는 다중 분류기 시스템의 활용이 대표적이다. 암 분류 문제에 있어서 좀 더 높은 분류 성능을 확보하고자 기계 학습 기반 다중 분류기 시스템을 암 분류에 이용하는 사례가 많다. [그림 2.11]은 암 분류를 위한 다중 분류기 시스템의 구조도를 나타내고 있다.

서로 다른 분류기들의 결과를 조합하거나 분류기를 선택하여 전체 성능을 향상시키는 다중 분류기 시스템은 그룹의 의견 종합이라는 측면 또는 지역적인 전문가를 양성한다는 측면에서 단일 분류기의 사용보다 일반적으로 좋은 성능을 보인다 [8].



[그림 2.11] 암 분류를 위한 다중 분류기 시스템의 구조도

다중 분류기 시스템이 단일 분류기 시스템보다 더 좋은 성능을 갖는 이유는 다음과 같이 설명될 수 있다. 예를 들어 만약, 이진 분류 문제에서 오류 확률이 0.5보다 적은 가설 공간 L 에서, 만약 기반 분류기들이 오류가 서로 상관관계가 없을 경우, 다중 분류기 시스템의 다수결 방법은 단일 분류기보다 더 적은 오류 확률을 갖는다. 실제로 21개의 기반 분류기들이 모두 똑같이 0.3의 오류 확률을 갖고, 각각의 오류들이 서로 독립일 경우, 다수결 방식을 이용하는 다중 분류기 시스템 전체의 오류는 아래와 같이 이항 분포를 따르게 된다.

$$P_{error} = \sum_{i=[L/2]}^{l} \binom{L}{i} p^i (1-p)^{L-i} \rightarrow P_{error} = 0.026 \ll p = 0.3 \tag{2.2}$$

따라서, 단일 분류기 시스템보다 다중 분류기 시스템이 더 낮은 오류 확률을 갖게 되어 상대적으로 높은 성능을 나타내게 된다.

위와 같은 형태의 다중 분류기 시스템을 이용하여 암 분류를 시도하게 되면 상대적으로 높은 분류 성능을 얻을 수 있다. 그러나 여러 분류기를 운용하는 데에 따르는 고비용과 느린 속도가 단점으로 지적된다. 또한 정보력이 있는 유전자 리스트를 별도로 추출한다는 것 자체가 거의 불가능하며 분리할 수 있다고 하더라도 유사성 척도를 단일하게 사용하고 있기 때문에 척도를 변경하게 되면 역시 정보력이 있는 유전자 리스트의 일관성을 유지하기 어렵다는 단점이 있다.

Ⅲ. 유의한 유전자 선택 방법

A. 마이크로어레이 표준화

실험에 의해 얻어진 마이크로어레이 자료는 자료의 특성상 많은 잡음을 포함하고 있다. 잡음에는 크게 두 가지가 있는데 하나는 자료자체에 내재하는 잡음과 실험 수행에서 발생하는 잡음이다. 전자의 경우 해당 자료의 산포를 증대시켜 통계적 검정력을 약화시키나 개개의 자료에 대해 랜덤하게 발생하므로 제거가 어려운 반면 후자의 실험에 의한 잡음인 경우 통계적 분석의 결과에 유의한 영향을 미칠 수 있는 위험성을 지니고 있으므로 표준화를 통하여 다양한 형태의 잡음을 찾아내어 제거해야 한다. 즉 표준화란 마이크로어레이 실험에서 슬라이드 내 자료의 유전자 발현 수준을 정확하게 측정할 수 있도록 유전자 발현 수준에 영향을 미치는 여러 잡음을 찾아 제거함과 동시에 각 슬라이드 혹은 실험 간의 결과를 비교할 수 있도록 보정하는 것이다.

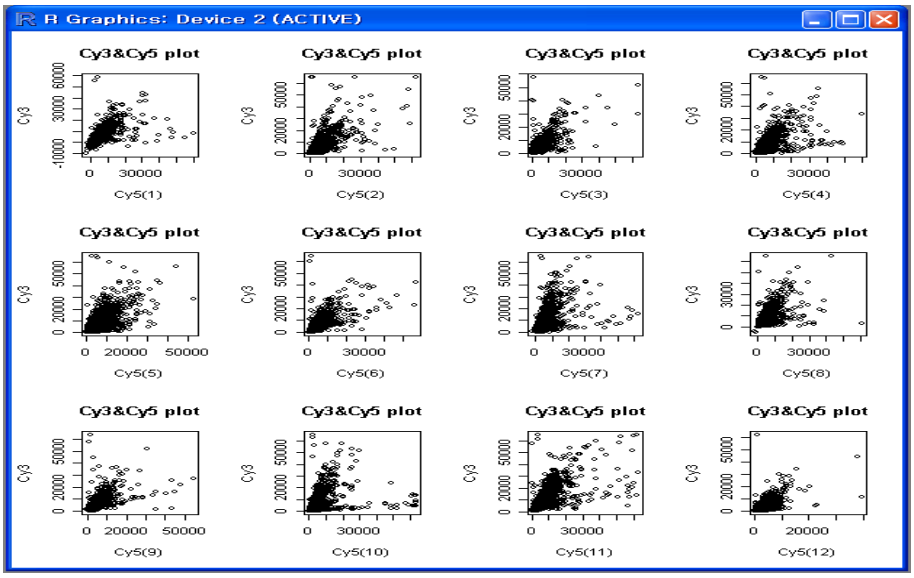
실험에 의해 생성된 마이크로어레이 자료에서 적색 강도를 R, 녹색 강도를 G 라 하면 이의 로그 변환한 값 log_2R과 log_2G를 좌표축으로 하여 2차원상의 산점도를 통해 $log_2R = log_2G$의 직선상에서 벗어난 유전자의 양을 관찰함으로써 자료를 분석한다. 또한 로그 변환된 인텐시티의 비율인 M과 평균인 A를 이용한 산점도(M-A plot)를 그려 보기도 한다. 따라서 표준화 하는데 대상이 되는 척도는 M과 A이고 식 (3.1)과 식 (3.2)와 같이 나타낼 수 있다.

$$M = \log_2 \frac{R}{G} = \log_2 R - \log_2 G \tag{3.1}$$

$$A = \log_2(\sqrt{R \times G}) = \frac{1}{2}(\log_2 R + \log_2 G) \tag{3.2}$$

이 두 값 M과 A를 이용하는 경우에는 log_2R과 log_2G를 이용한 산점도를 45°회전시킨 것으로 0값을 기준으로 자료를 관찰하여 분석한다.

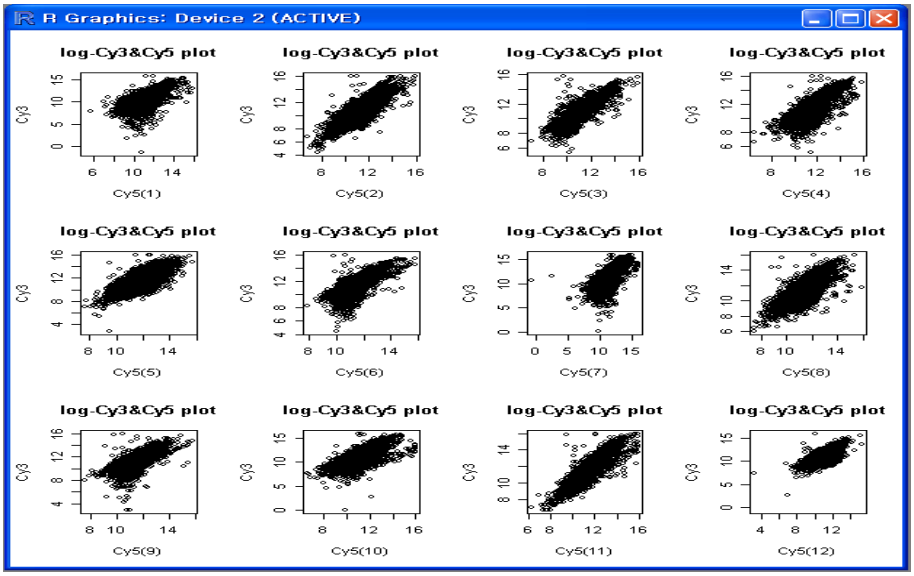
슬라이드 내에 발생하는 잡음의 원인으로는 형광물질의 특성상 열이나 빛에 대한 반응의 차이로 녹색이 적색보다 강도가 높게 측정되는 경향과 형광물질 혼합의 효율성 문제, 자료를 모으고 스캐닝하는 과정에서의 잡음, 프린트 팁 간의 차이 등이 있다. 슬라이드 내에서의 표준화는 보통 로그 변환한 두 값 log_2R과 log_2G의 균형을 맞추기 위해 행해지는데, 위치 보정의 경우는 일반적으로 로그비율 M 의 평균이나 중위수를 사용하여 보정하며, 스케일 보정일 경우 분산 값을 보정하여 표준화시켜 준다[8]. [그림 3.1]은 본 논문에서 사용한 마이크로어레이 데이터의 *Cy3 & Cy5 plot*의 일부분이다[8].



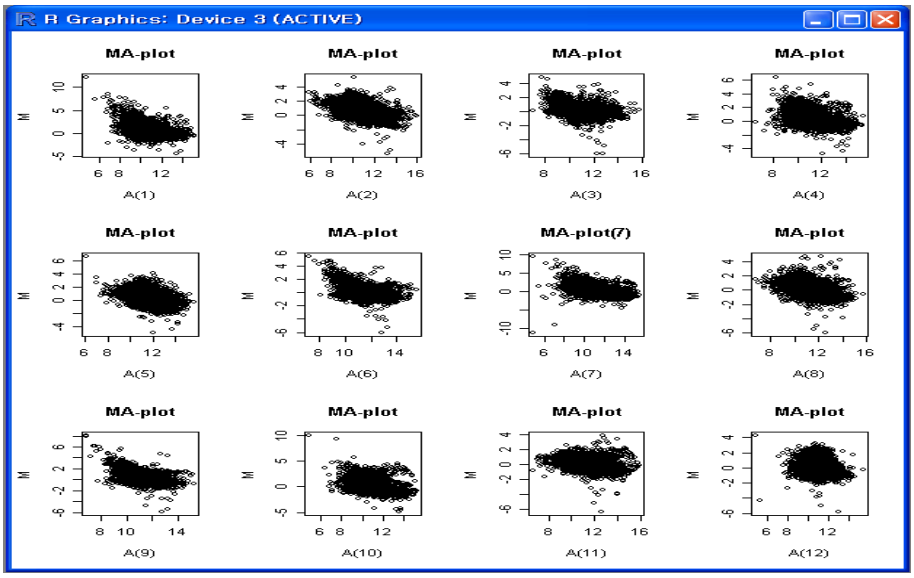
[그림 3.1] *Cy3 & Cy5* plot

대다수(95%이상)의 유전자의 발현수준이 서로 비슷하며 극소수의 유전자만이 유의하게 다른 발현강도를 보일 것이므로, *R-G* plot이 대각선(45^0) 중심이어야 하는데, 실제로 이러한 경향을 보이는 경우는 거의 없다. 일반적으로 곡선 등의 비선형 추세가 나타나거나 점들이 한 쪽으로 치우친 모습을 하고 있다.

[그림 3.2]와 [그림 3.3]은 각각 로그로 변환된 log_2G & log_2R *plot* 과 *M-A plot* 이다.



[그림 3.2] *log_2G & log_2R plot*



[그림 3.3] *M-A plot*

1. 포괄적인 표준화 방법

포괄적인 표준화 방법(*Global normalization*)은 고전적인 표준화로써 전통적인 통계적 실험에서의 표준화 방법에 근거하여 로그 변환한 값을 표준화하는 것이다.

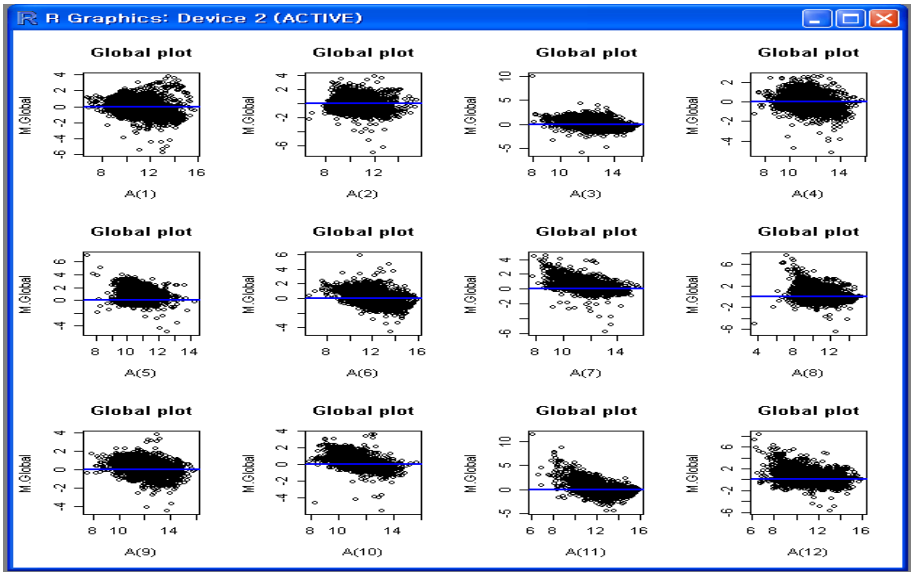
먼저, Chen et al.[9](1997)에 의해 제안된 방법은 G와 R값이 한 슬라이드 내에서 일정한 비를 이루고 있다($R = k \cdot G$)는 가정을 한 것으로 그 비율 분포의 중심을 상수(c)의 가감에 의해 0에 맞추어 가는 것이다. 이는 투입되는 *Cy3*, *Cy5* 형광물질의 특성상 이러한 잡음이 첨가되리라는 것에 기초한 것이다.

$$M = \log_2 \frac{R}{G} \Rightarrow \log_2 \frac{R}{G} - c = \log_2 \frac{R}{k \cdot G} \tag{3.3}$$

따라서, 각 유전자(j)마다 표준화된 값을 M_j^{Global} 이라 하면 그 값은 식 (3.4)와 같이 정리할 수 있다.

$$M_g^{Global} = M_j - \hat{c} \tag{3.4}$$

$c = \log k$의 추정방법으로는 최대우도추정법이 제안되었으나 중위수나 평균을 사용하기도 한다. 보통 이상치의 영향을 받지 않기 위해 중위수를 많이 사용하며 이 방법을 M의 중위수를 이용한 표준화 방법(*Global median of log ratio*)이라고 한다[9].



[그림 3.4] *Global* 표준화 *plot*

Global normalzation M-A 산점도를 보았을 때 수평축(0)이 중심이나 수평축(0) 중심에서 벗어나는 정도가 유전자 발현강도에 따라 분산이 다른 경우 편차가 큰 모습을 보이고 있다.

2. 특성화한 표준화 방법

포괄적인 표준화는 G나 R값 중 하나의 값을 고정한 후에 표준화하는 것으로 비율의 특성상 자료의 형태는 기본적으로 $R = G$에 대칭적으로 분포하게 되는데 한쪽을 고정하여 고려하는 경우 대칭성이 깨질 우려가 있다. 이러한 고려 하에 Yang et al.[10](2001)은 M과 직교하는 척도 인텐시티 A를 제안하여 이를 기준으로 표준화하는 방법을 제안하였다. 인텐시티는 각 형광 이미지 파일에서 측정된 강도 R과 G의 로그 변환한 값의 평균으로 구한다.

가장 먼저 간단한 가정으로 선형모형에 대한 가정을 할 수 있으며 식 (3.5)와 같다.

$$M = \beta_0^{MA} + \beta_1^{MA} A \tag{3.5}$$

식 (3.5)에 의해 각 유전자(j)에 대한 표준화된 값을 M_j^{MA}라 두고, 식 (3.6)에 의해 구한다.

$$M_j^{MA} = M_j - \widehat{M}_j \tag{3.6}$$

여기서, $\widehat{M}_j$은 회귀분석의 추정치를 이용하고, M_j^{MA}는 잔차값을 이용하게 된다. 이 방법을 A를 고려한 선형 표준화 방법(*Intensity Dependent Linear Regression normalization*)이라고 한다.

선형모형에서 더 나아가 비선형적인 잡음이 첨가되는 경우에 대해 *robust smoother lowess* 산점도를 사용하여 A에 의존한 비선형모형으로 확장할 수 있다. 즉, $M = k(A)$와 같은 A에 의존하는 일반적인 함수형태를 가정하고, 이상점에 대해 상대적으로 덜 민감한 *Lowess* 함수 추정법으로 추정하는 방법이다. 각 유전자

(j)별로 *Lowess* 함수를 추정하며 각 유전자(j)에 대해 표준화된 M_j^{LOWESS}는 식 (3.7)과 같이 로그비율에서 함수에 의해 추정된 값을 제거함으로써 구해진다.

$$M_j^{LOWESS} = M_j - k(\widehat{A_j}) \tag{3.7}$$

여기서 $k(A)$는 알고리즘을 적용한 소프트웨어에 의해 자동으로 추정되며 추정된 식에 의해 표준화된 값 M_j^{LOWESS}를 얻게 된다. 이 방법을 A를 고려한 비선형 표준화 방법(*Intensity dependent Lowess Normalization*)이라 한다[10].

위의 비선형 표준화 방법에 추가로 프린트 팁에 의한 변이를 고려할 필요가 있다. 서로 다른 k개의 다른 프린트 팁을 사용하여 형광물질의 스팟을 출력하므로 각 팁(*tip*)의 오프닝이나 길이 또는 많은 시간이 흐른 후 출력 이미지의 변형 같은 잡음이 발생할 우려가 있다. 프린트 팁에 의한 잡음의 첨가가 두드러질 때 함수의 추정을 프린트 팁(k)별로 구별하여 각각 추정할 수 있다. 이 방법은 간단하게 [프린트 팁 + A]에 의한 변이를 고려한 표준화 방법으로 표현할 수 있으며 위치를 보정하는 것과 스케일을 보정하는 두 과정으로 나눌 수 있다. 먼저 위치를 보정하는 과정은 식 (3.8)에 의해 수행된다.

$$\begin{aligned} M &= \log_2 \frac{R}{G} \Rightarrow \log_2 R - c(A_k) \\ &= \log_2 \frac{R}{(k(A_k) \cdot G)} \\ & \qquad\qquad k = 1, \ldots, k \end{aligned} \tag{3.8}$$

따라서, 프린트 팁(k)별로 구별하여 각 유전자(j)에 대해 표준화된 값 M_j^{LOWESS}은 식 (3.9)와 같이 된다.

$$\begin{aligned} M_j^{LOWESS} &= M_j - c(\widehat{A_{jk}}) \\ & \qquad\qquad k = 1, \ldots, k \end{aligned} \tag{3.9}$$

식에서 $c_k(A)$는 k 번째 프린트 팁 그룹에서 M-A 산점도 상의 *Lowess* 적합 값이다. 즉, 이 방법은 A척도를 이용한 표준화 방법과는 다르게 전체적으로 *Lowess* 적합을 하지 않고, 각 프린트 팁 그룹마다 *Lowess* 적합을 한 후 표준화시켜준다. 이 방법을 프린트 팁 간의 위치 보정을 통한 표준화(*Within Print tip Lowess*)라 한다.

그러나, 프린트 팁 별로 위치 보정한 후 자료는 서로 다른 프린트 팁 그룹별로 각 프린트 팁에 대한 분포의 중심이 0으로 표준화되지만, 분포의 스케일은 서로 같지 않을 경우가 종종 발생한다. 이로 인해 스케일 조정이 필요한데, 이를 위해서는 먼저 i 번째 프린트 팁 그룹으로부터의 모든 M은 다음과 같이 평균이 0이고, 분산은 $a_i \bullet \sigma^2$인 정규분포 (식 (3.10))을 따른다는 것을 가정해야 한다.

$$M_i \sim N(0,\ a_i^2 \bullet \sigma^2) \tag{3.10}$$

여기서 a_k^2는 k 번째 프린트 팁의 스케일 요인이며 σ^2는 실제 로그 비율의 분산이다. 위와 같이 가정한 후 프린트 팁 그룹 간에 다른 분산을 갖게 하는 요인 a_i를 식 (3.11)의 제약조건 하에서 최대우도추정법에 의해 추정한다.

$$\sum_{i=1}^{I} \log a_i^2 = 0 \tag{3.11}$$

이는 프린트 팁(k) 그룹별 스케일 요인의 합을 0으로 만드는 것이다. 최대우도추정법을 이용하여 추정된 $\hat{a}_k^2$는 식 (3.12)와 같다.

$$\hat{a}_k^2 = \frac{\sum_{j=1}^{n_i} M_{ij}^2}{\sqrt{\prod_{k=1}^{I} \sum_{j=1}^{n_k} M_{kj}^2}} \qquad j = 1, \ldots, n_k \tag{3.12}$$

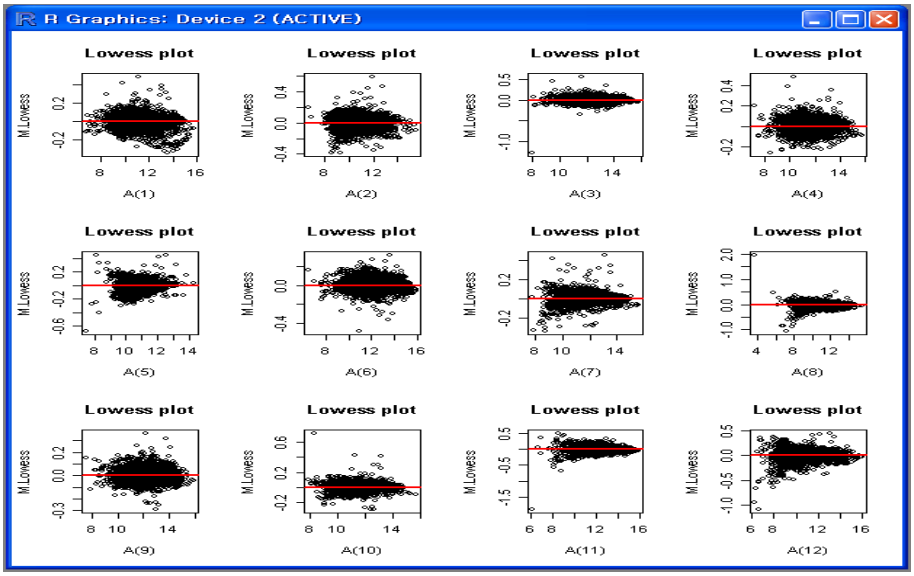
여기서 M_{ij}는 i 번째 프리트 그룹의 j 번째 로그비율을 의미한다.

최대우도추정량을 사용하는 것보다 식 (3.14)와 같이 중위수 (*MAD* : *median absolute deviation*)를 사용하여 추정하는 것이 좀 더 로버스트하다. 통계량 *MAD*는 *robust smoother lowess* 산점도를 이용한 것과 비슷하게 *M-A* 산점도 상에서 이상치의 영향을 받지 않는다. 이 경우 추정된 a_i는 식 (3.13)과 같다.

$$\hat{a}_i = \frac{MAD_i}{\sqrt{\prod_{i=1}^{I} MAD_i}} \tag{3.13}$$

$$MAD_i = median_j\left(|M_{ij} - median_j(M_{ij})|\right) \tag{3.14}$$

이렇게 추정된 $\hat{a}_i$를 각 프린트 팁 그룹별로 제거한 후 표준화를 수행한다. 이를 프린트 팁 간의 스케일 표준화(*scale normalization* : 스케일 보정을 통한 표준화)라 한다. 또한 슬라이드 혹은 실험 간의 비교를 목적으로 하는 경우 각 슬라이드 간의 스케일 보정을 통한 표준화를 수행해야 하는데 이는 슬라이드 내에서의 크기 조정의 방법과 같다. [그림 3.5]는 *Lowess normalization plot* 이다.



[그림 3.5] *Lowess* 표준화 *plot*

*Global normalization*이 선형적인 패턴을 보정했을 뿐 패턴 자체를 제거하지 못한 것에 반해 *Lowess normalization*은 선형적인 패턴뿐만 아니라 비선형적인 패턴 자체를 제거하였다.

B. 유의한 유전자 선택

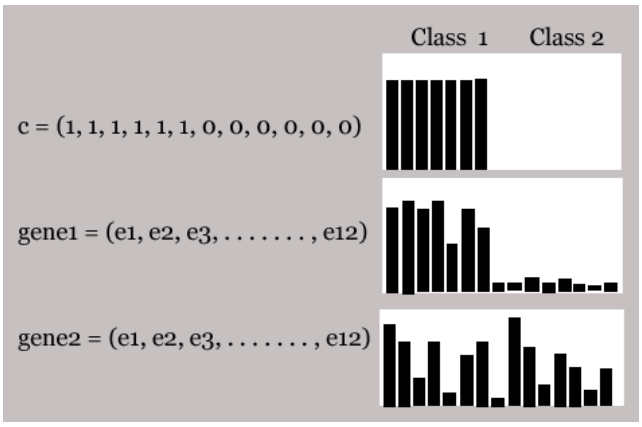
마이크로어레이를 통하여 얻게 되는 유전자 발현 정보 데이터는 기계 학습 기반 분류기를 사용하여 분류를 시도하던 기존 전통적 분야의 데이터 집합과는 매우 다른 특성을 지닌다. 직접 종양 샘플에서부터 마이크로어레이 기술에 의해 데이터가 생성되기 때문에 샘플의 수는 적고 각 샘플에 대한 속성 즉, 유전자의 수는 수천 개에서 수만 개 내외가 되는 것이 일반적이다. 이러한 특징은 기계 학습 기반 분류기를 이용하여 학습을 하기에 적절하지 않다. 마이크로어레이로부터 얻어지는 유전자 데이터 중에서 실제로 각 샘플의 특정 클래스, 즉 암의 특정 클래스와 연관이 큰 유전자는 그 수가 매우 적으며, 실험 과정 중에 발생한 잡음을 포함하고 있는 유전자 발현 데이터 또한 많다. 따라서 기계 학습 기반 분류기를 이용하여 현실적으로 효과적인 학습을 하기 위해서는 해당 클래스와 연관성이 높은 유전자들을 시스템의 전단부인 전처리 과정에서 선택해야만 한다. 이것을 유전자 선택과정(*gene selection*)이라고 한다. 주로 통계적 유사성 척도나 클러스터링 기법 등을 이용하여 해당 클래스 분류에 상대적으로 관련이 높은 유전자들을 선택한다. 마이크로어레이로부터 얻은 데이터 집합이 샘플 M개와 유전자 N개로 이루어져 있다고 가정한다. M개의 샘플은 클래스 A와 클래스 B 두 가지 종류로 구분된다고 가정하고, K 개의 샘플을 정상 세포 혹은 암 세포라고 하고 M-K개의 샘플은 다른 종류의 암 세포의 샘플이라고 가정하면, 각 유전자 발현 데이터는 식 (3.15)와 같은 형태의 벡터로 표현할 수 있다[28].

$$G_i = (e_1, e_2, e_3, \quad , e_k, e_{k+1}, \quad , e_M) \tag{3.15}$$

각 클래스에 대한 특징을 극단적으로 뚜렷하게 나타내면서 이상적으로 발현하는 유전자를 G_{ideal}이라고 하면, 처음의 K개의 클래스 A의 특징을 1로 정의하고 나머지 M-K개의 클래스 B혹은 처음의 K 와는 다른 암세포의 특징을 0으로 정의하여 식 (3.16)과 같은 벡터로 표현할 수 있다. G_{ideal}은 앞장에서 언급한 이상 유전자 모델과 같은 의미이다.

$$G_{ideal} = (1, 1, 1, ..., 1, 0, 0, 0, ..., 0) \tag{3.16}$$

[그림 3.6]은 유전자 발현 정도와 이상 유전자 모델 사이의 관계를 나타내고 있다.



[그림 3.6] 유전자 발현 정도와 이상 유전자 모델 사이의 관계

[그림 3.6]에서 *gene2*의 경우에는 *Class 1*과 *Class 2* 모두에서 발현 정도가 거의 비슷하고 또 일정하지도 않다. 따라서 *gene2*는 *Class 1*과 *Class 2*를 구별한 데에 직접적인 연관을 갖는 유용한 유전자라 하기 어렵다. 즉 이 샘플에서 *gene2*는 정보력이 있는 유전자가 아니다. 그러나 *gene1*의 경우 *Class 1*과 *Class 2* 두 클래스에서 평균적인 발현 정도의 차이를 보이고 있다. 또한 각 클래스에서의 발현 정도가 크고 들락 없이 고르게 유지되고 있음을 볼 수 있다. 즉, *gene1*은 *Class 1*과 *Class 2*를 구별하는 데에 유용한 정보를 가지고 있는 유전자라 가정할 수 있다.

*Golub*은 식 (3.16)을 'signal to noise'라 정의하고 각 클래스에서 나타는 유전자 발현 데이터들의 평균값의 차이는 크면서 동시에 표준 편차의 합은 작을수록 해당하는 유전자가 유용한 유전자일 가능성이 높다고 하였다[28]. 이러한 양상이 극단적으로 나타나는 매우 이상적인 발현을 보이는 유전자가 [그림 3.6]의 *c*와 같은 경우로 이를 이상 유전자 모델이라 정의한다.

이제 여러 개의 유사성 척도를 각각 사용하여 식 (3.16)과 각 유전자 사이의 유사성 여부를 측정한다. 각 유사성 척도별로 이상 유전자 모델과 유사도가 높은 유전자들을 순차 정렬하고 상위의 유전자 일부를 선택하여 기계 학습 기반 분류기의 학습 데이터로 사용한다. 이 때 선택해야 하는 상위 유전자의 수는 20에서 200개가 안정적인 분류 결과를 나타내는 것으로 알려져 있다[28].

본 논문에서는 표준화 후 상위 유의한 유전자 50개, 90개, 200개를 선택하여 학습 데이터로 사용하였다.

유전자 선택을 위해 개별적으로 사용되던 기존의 유사성 척도는 [그림 3.7]과 같다.

• Pearson correlation coefficient(PC)

$$PC(G_i, G_{ideal}) = \frac{\sum G_i G_{ideal} - \frac{\sum G_i \sum G_{ideal}}{N}}{\sqrt{\left(\sum G_i^2 - \frac{(\sum G_i)^2}{N}\right)\left(\sum G_{ideal}^2 - \frac{(\sum G_{ideal})^2}{N}\right)}}$$

• Euclidean distance(ED)

$$ED(G_i, G_{ideal}) = \sqrt{\sum (G_i - G_{ideal})^2}$$

• Spearman correlation coefficient(SC)

$$SC(G_i, G_{ideal}) = 1 - \frac{6\sum (D_G - D_{ideal})^2}{N(N^2 - 1)}$$

(D_G and D_{ideal} are the rank matrices of G_i and G_{ideal})

[그림 3.7] 유의한 유전자 선택을 위한 기존 유사성 척도

두 변수의 평균으로부터의 차이, 즉 편차를 서로 곱한 값들의 평균인 공분산은 변수를 측정한 척도의 단위에 따라 그 값이 달라지기 때문에, 공분산만으로 두 변수 간의 선형관계를 정확하게 파악하기 어렵다. 따라서 측정 단위에 따라 그 크기가 바뀌지 않는 공분산을 구하면 그 값이 곧 상관계수가 된다. 변수의 측정 단위에 따라 그 크기가 변하지 않는 공분산은 공분산 계산에 사용되는 두 변수의 편차를 해당 변수의 표준편차로 나누어 줌으로써, 즉 각 변수의 편차를 표준화한 편차 값들을 사용하여 구할 수 있다. 다시 말해서, 공분산을 구할 때 사용하는 각 변수의 편차를 그 변수의 표준편차로 나누어 편차를 표준편차 단위로 바꾼 다음 이들을 서로 곱한 값들의 평균을 공분산으로 한다. 이렇게 함으로써 측정 단위에 따라 변하지 않는 두 변수 간의 선형관계를 파악할 수 있는 표준화된 공분산을 도출할 수 있는데 이 값이 상관계수이다. 이처럼 두 변수의 표준화된 편차 곱의 평균으로 구한 상관계수를 피어슨 적률 상관계수라 한다. 상관계수의 크기는 -1에서 +1사이의 값을 갖게 된다. 기본적으로 두 변수 간의 선형관계의 정도를 나타내는 상관계수는 두 변수 간의 선형관계 정도와 방향을 수학적인 수치로 정량화하여 표시하는 지수로서, 상관계수가 +1 혹은 -1일 경우에 두 변수는 완전한 양과 음의 상관관계를 갖는 반면, 상관계수가 0에 가까울 때 두 변수는 선형이 아닌 다른 관계이거나 서로 독립적이어서 무관한 관계에 있다고 볼 수 있다[29][30].

일반적으로 상관분석은 등간척도나 비율척도를 이용하여 측정된 변수들 간의 선형관계를 파악하는 데에 사용된다. 그러나 비모수 통계분석을 이용하면 서열척도로 측정된 변수들 간의 연관성도 파악할 수 있다. 서열척도로 측정된 변수 간의 상관관계는 스피어만의 서열 상관계수나 켄달의 타우 기법으로 산출하는 서열 상관계수를 이용하여 분석할 수 있다. 이들 두 가지 방법은 큰 차이가 없으므로 비교적 계산이 간편한 스피어만의 서열 상관계수가 많이 사용된다. 스피어만 서열 상관계수는 분석하고자 하는 각각의 대상이 가지고 있는 속성에 대하여 서열 척도로 측정한 2개 변수 간의 연관성, 즉 상관관계를 나타내는 계수이다.

이 밖에도 유클리디안 거리척도가 사용된다.

C. 조합 방법(Combination Method)

기존 방법과 같이 [그림 3.7]의 각 유사성 척도를 개별적으로 사용하여 유용한 유전자 목록을 만들게 되면, 정확한 암 분류에 있어서 중요한 정보를 내포하고 있다고 판단된 유전자 목록이 각 유사성 척도를 달리할 때마다 상이하게 나타난다. 이것은 각 유사성 척도마다 정보력이 있는 유전자 데이터를 판단하는 기준이 다르기 때문이다. 따라서 일관적이고 신뢰성이 높으면서 동시에 정보력이 있는 유전자 목록을 확보할 수 없다. 이러한 문제를 해결하기 위해 본 논문에서 새로운 시스템을 제안하고자 한다. 제안된 시스템에서는 유사성 척도 한 가지를 사용해서 얻게 되는 유전자 목록의 일관성과 신뢰성의 결여를 보안하기 위해, 여러 개의 유사성 척도를 함께 활용하여 정보력이 있는 유전자 목록을 만든다. 구체적으로는 유전자 발현 데이터들에 대해 각 유사성 척도를 사용하여 두 가지 이상의 유사성 척도에서 정보력이 있는 유용한 유전자로 높게 평가 받은 유전자들을 정량화된 유용성 정도에 따라 서열화하고 이들의 상위 부분을 선발하여 정보력이 있는 유전자 목록으로 확정하기 위한 방법을 고안해야 한다. 즉, 다수의 유사성 척도로부터 정보력이 있다고 평가된 유전자 데이터를 선택하기 위해, 각 유사성 척도에 의해 측정된 값으로 이루어진 배열을 적절하게 활용하도록 한다.

이와 같은 조합 방법을 구현하기 위해 사용할 수 있는 정보는 각 유사성 척도에 의해 계산된 유전자 각각의 이상 유전자 모델(G_{ideal})에 대한 유사도 수치이다. 유사도 수치를 이용하여 다수의 척도에서 정보력있는 유전자로 평가받은 의미 있는 유전자들을 선택하는 방법을 제안하고자 한다. 이를 위한 알고리즘은 [그림 3.8]에 정리하였다.

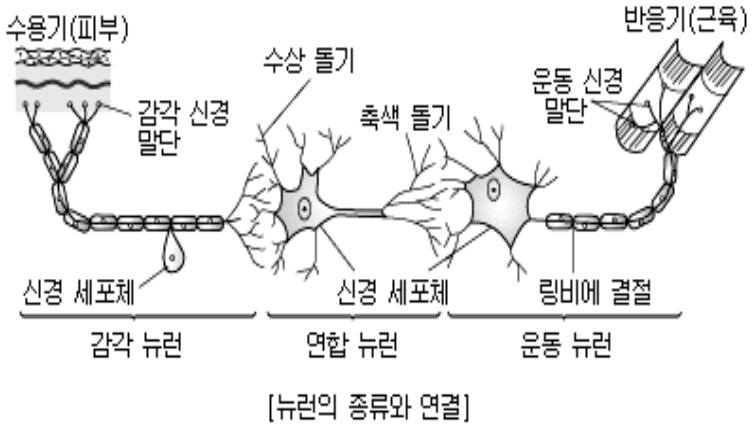
Step 1) N choice of similarity scales	조합하고자 하는 유사성 척도 N개 선택
Step 2) Compute $S_i = \sum_{i=1}^{n} E_i$ n : number of sample	유사성 척도에서 i 번째 유전자 G_i 가 평가 받은 유사도 수치의 총계
Step 3) Compute $V_i = \frac{\sum_{i=1}^{N} S_i}{N}$	합산한 결과 값의 평균
Step 4) combination value = V_i	step 3)의 결과를 유사성 척도 N개의 조합 값
Step 5) Sort of an ascending series	유사도 값이 높은 G_i 을 선택하여 유전자 목록 완성

[그림 3.8] 정보력이 있는 유전자 선택을 위한 조합 알고리즘

Ⅳ. 신경망

A. 신경망

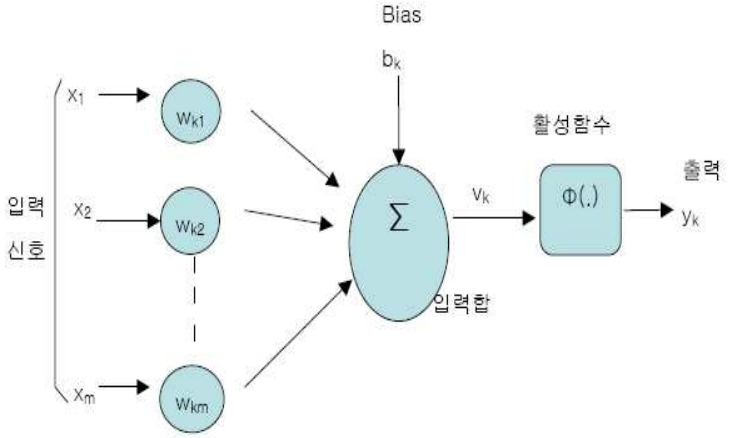
신경망이란 인간의 두뇌 작용을 신경세포들 간의 연결 관계로 모델링한 것으로 간단히 정의할 수 있다. 인간의 두뇌는 무수히 많은 신경세포(뇌의 기본 소자: *Neuron*)들과 이들을 상호 연결하는 연결 체들로 구성되어 있다. 이러한 신경세포들은 그들 사이의 정보 교환을 병렬적으로 수행하면서 이러한 정보 교환 기능을 학습이라는 과정을 통해 향상시킨다. 신경세포(*Neuron*)는 세포체(*Cell Body*), 신경절(*Synapse*), 축색돌기(*Axon*), 수상돌기(*Dendrite*)로 크게 구성되어지며, 신경 과학자들의 연구에 의하면 인간의 뇌에는 이러한 신경세포가 1,011개 정도가 존재하며, 이들은 1015개의 상호연결 관계를 가지고 이루어져 있는 것으로 추정되고 있다. [그림 4.1]은 인간의 뇌에 존재하는 뉴런의 생물학적인 구조를 보이고 있다.



[그림 4.1] 생물학적인 뉴런의 구조

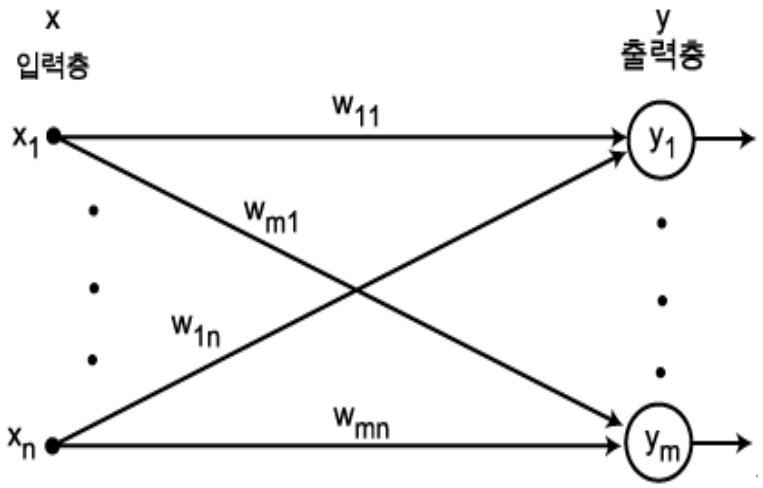
신경망에 관한 연구는 생물학적인 신경망에 관한 연구와 인공신경망의 연구로 구분이 된다. 전자의 경우는 인간의 생체기능을 근원적으로 연구하는 분야이며, 후자의 경우 즉, 인공신경망은 전기, 전자, 컴퓨터, 기계 공학자들을 중심으로 생물학적 신경망을 인공적으로 설계하고 이를 하드웨어화하는 연구를 포함한다.

신경과학자들의 연구에 의하면 복잡한 인지 작업을 수행하는 각 뉴런은 약 천개에서 10만개의 뉴런과 신경절을 통해 서로 연결이 되어 정보를 전달한다. 이러한 인간 뇌의 기능을 모방하고자 인공신경망은 노드(*Node*) 혹은 처리요소라고 하는 많은 수의 뉴런들이 복잡하게 연결된 망으로 구성되며, 이러한 인공 신경망은 인간의 뇌가 수행하는 예제에 의한 학습, 일반화, 연상기억, 결함 극복 등의 특징을 보이고 있다. 이러한 특징을 가지는 인공신경망은 아래와 같이 모델링 할 수 있다[그림 4.2].



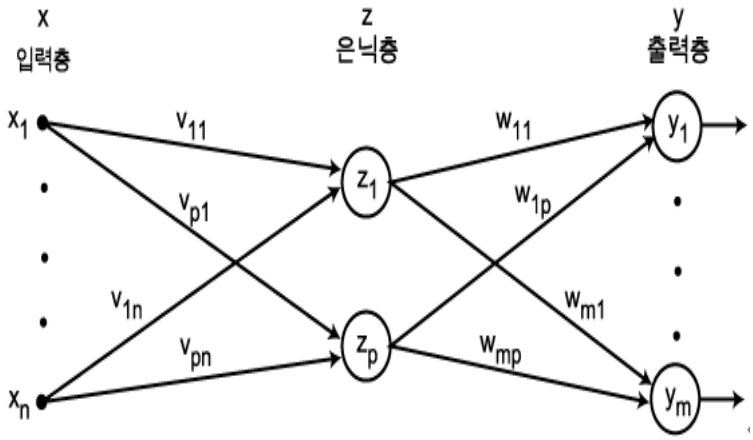
[그림 4.2] 인공 뉴런의 구조

뉴런 한 개는 단순한 기능을 가지지만, 이들이 서로 연결된 신경망은 강력한 기능을 수행할 수 있다. 가장 간단한 신경망은 입력 뉴런들로 구성된 입력층과 출력 뉴런들로 구성된 출력층으로 구성된 단층 신경망(*Single Layer Neural Network*)으로서 [그림 4.3]은 단층 신경망의 구조를 보여준다. 일반적인 경우 입력층은 단순히 입력 벡터들을 전송하는 역할만을 수행한다.



[그림 4.3] 단층 신경망

대체로 신경망이 더 커지고 복잡해지면 보다 나은 계산 기능을 수행할 수 있다. 신경망의 입력 층과 출력 층 사이에 새로운 층들을 추가하는 다층 신경망(*Multi Layer Neural Network*)은 보다 복잡한 기능을 수행할 수 있게 되며, 이 경우 입력층과 출력층 사이의 층들은 은닉층(*Hidden Layer*) 혹은 중간층(*Internal Layer*)이라 부른다. [그림 4.4]는 한 개의 은닉층을 가진 다층 신경망의 구조를 보인다.



[그림 4.4] 다층 신경망

이와 같은 구조를 가지는 인공신경망은 다양한 학습규칙에 의해 학습될 수 있다. 신경망의 학습이란 정보처리 시스템의 목적에 맞도록 각 뉴런들이 가지는 가중치 벡터들을 조절해 가는 것을 의미한다. 이러한 학습규칙은 크게 교사 신호를 인가해 주는 지도 학습과 교사 없이 입력신호의 성질에만 기초를 둔 비지도 학습으로 나눌 수 있다.

B. 지도 학습과 비지도 학습

1. 지도 학습

지도 학습(*Supervised Learning*)이란 실험데이터의 입력벡터와 그에 대응하는 출력 값을 함께 신경망에 입력시킨 후 학습시키는 방식이다. 대표적인 학습 알고리즘으로는 델타규칙(*delta Rule*), 오류 역 전파(*Error Backpropagation*) 학습규칙 등이 있다.

델타규칙은 선형 활성화 함수를 가지는 단층 신경망의 학습규칙 중의 대표적인 알고리즘이다. 이 알고리즘은 전체 출력 뉴런들에 대해 계산된 출력값과 목적값 간의 오차를 최소화시키는 알고리즘이다. 두 값 사이의 오차가 존재하지 않는다면 가중치 벡터의 조절은 수행되지 않으며, 오차가 있다면 오차를 줄이는 방향으로 가중치 벡터를 조절한다.

$$W_{ij}(t+1) = W_{ij}(t) + \alpha[T_i(t) - \sum_i W_{ij}(t)X_i(t)] \qquad (5.1)$$

식 (5.1)에서 $W_{ij}(t)$는 입력뉴런 i와 출력뉴런 j 간의 가중치 벡터를 나타내고, α는 학습률 상수, $T_i(t)$는 목표값, $X_i(t)$는 입력값을 각각 의미한다. 일명, 일반화된 델타규칙으로도 불리는 오류 역전파 학습 규칙은 다층 신경망을 학습시키는 가장 일반적인 방법이다. 이 알고리즘은 원하는 목표값 T_i와 실제값인 O_i 사이의 오차 자승의 합으로 정의된 비용함수 E의 값을 최급강하법(*Gradient Descent Method*)에 의해 최소화하는 방향으로 학습을 진행하며 각 뉴런은 비선형의 활성화 함수를 사용한다.

$$E = \sum_p E_p = \frac{1}{2}\sum_p \sum_i (T_i^p - O_i^p)^2 \tag{5.2}$$

이 학습 규칙은 은닉층의 학습을 위해 출력층에서 발생한 오차를 사용하여 은닉층의 오차를 계산하고, 다시 이 값을 입력층으로 역전파시켜서 출력층의 오차가 원하는 수준이 될 때까지 반복 수행하는 알고리즘이다[22,23].

지도 학습은 실험 데이터의 입력벡터와 그에 대응하는 출력값을 함께 신경망에 입력시킨 후 학습시키는 방식이기 때문에 학습 전에 이미 엄선된 각 단계별 데이터와 입력된 데이터의 출력결과가 의도한 결과와 일치하는지를 알려주는 교사를 필요로 한다. 학습 시에 교사 신호와 잘 분류된 학습 데이터를 필요로 하므로, 데이터의 선정시간이 지나치게 길거나 교사가 일일이 학습에 관여해야하는 등의 단점을 지닌다.

2. 비지도 학습

비지도 학습(*Unsupervised Learning*)은 말 그대로 신경망의 학습과정에서 입력패턴에 상응되는 어떤 목적값의 인가가 없는 학습을 뜻한다. 하나의 입력벡터가 신경망에 주어지면 각 뉴런의 가중치 벡터들은 인가된 입력벡터와 유사해지도록 조절된다. 이러한 과정의 반복은 신경망이 가지는 뉴런이 각기 특정한 입력에 대해서만 반응을 보이게 되고 최종적으로는 입력패턴들을 몇 개의 클러스터로 분류하게 된다. 이러한 비지도 학습은 주어진 입력에 대해 어떤 뉴런을 반응하게 할 것인가에 대한 선정의 문제가 뒤따른다.

대표적인 비지도 학습의 알고리즘으로는 승자독점(*Winner Take All*)의 학습 형식을 가지는 *SCL*(*Simple Competitive Learning*), 승자 뉴런의 결정시에 뉴런에게 양심을 부여하는 *FSCL*(*Frequency Sensitive Competitive Learning*), 승자를 포함한 이웃 뉴런에 대해서 학습을 수행하는 *Kohonen*의 *SOFM*(*Self Organizing*

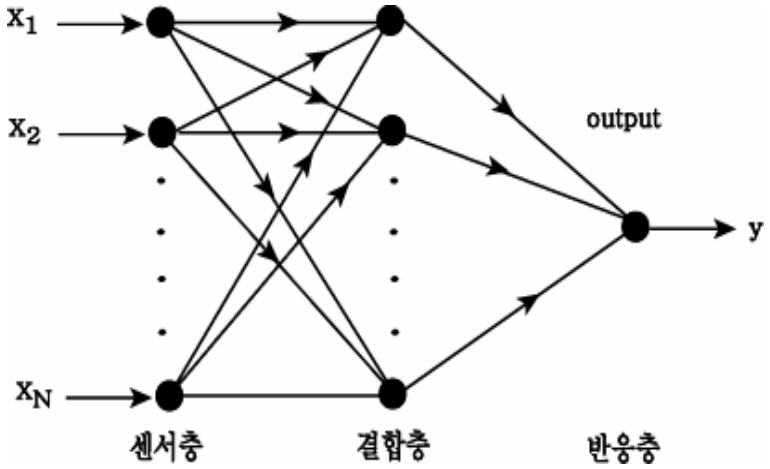
Feature Map) 등의 경쟁학습과 학습 과정 중에 출력 뉴런을 생성하는 *Carpenter*와 *Grossberg*의 *ART*(*Adaptive Resonance Theory*) 모델이 있다.

이와 같은 비지도 학습은 인간의 자율적인 지능 획득 원리에 보다 접근해 보자는 시도이고, 인간의 감각기관이 보여주는 다양한 특징을 보여준다. 비지도 학습은 매우 빠른 학습속도를 그 특징으로 하는 반면, 지도 학습에 비해서 응용범위가 넓지 못하다는 단점을 가진다.

C. 신경망 학습방법

1. 퍼셉트론

퍼셉트론(*Perceptron*)은 *Rosenblatt*에 의해 제안된 패턴 인식의 모델이다. 이것은 학습에 따라 인식 능력을 향상시키는 능력을 갖는 최초의 신경망 모델이다. 퍼셉트론은 [그림 4.5]에 나타나듯이 세 개의 층으로 되어 있다. 제 일층은 입력층(*Sensory Layer*, *S*층, 센서층)이라 불리우고 외부의 입력을 받아들인다. 입력층의 신호는 다음의 중간층(*Association Layer*, *A*층, 결합층)에 보내져, 마지막으로 출력층(*Response Layer*, *R*층, 반응층)에서 인식결과가 출력된다.

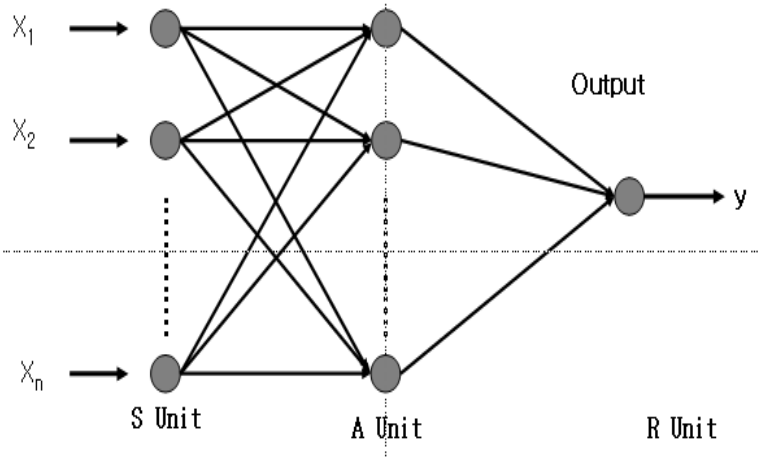


[그림 4.5] 퍼셉트론의 구조

입력층에는 입력이 있으면 1, 없으면 0을 출력하는 뉴런이 배열되어 있다. 입력층의 출력 신호를 받아들이는 것이 중간층이고, 여기에는 시냅스의 하중(시냅스 하중, 결합무게, 무게)에 따라서 무게가 더해진 입력의 총화가 수치를 넘으면 1을 출력하는 형식 뉴런(유니트)이 늘어선다.

a. 단순 퍼셉트론

단순한 퍼셉트론에서의 뉴런은 층상으로 나열되어 있어 각 층은 S 유니트(*Sensory Unit*), A 유니트(*Associate Unit*) 및 R 유니트(*Response Unit*)로 불려지고 있다.



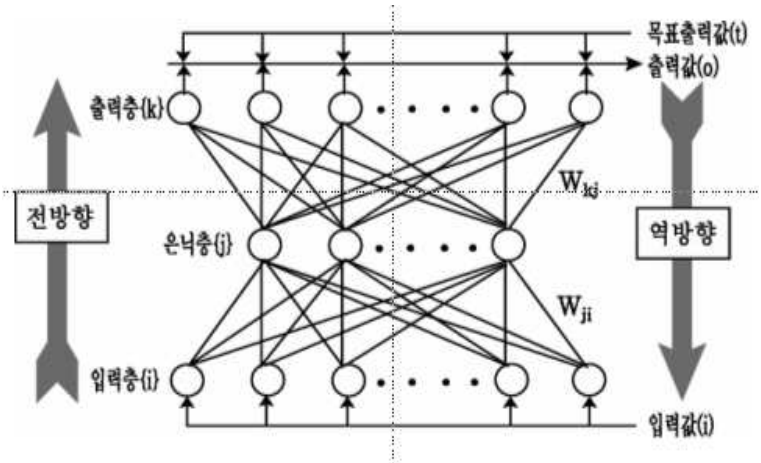
[그림 4.6] 단순 퍼셉트론의 구조

분류대상 패턴을 입력하면 S 유니트가 반응한다. 다음에 A 유니트는 S 유니트로부터 입력을 받아서 신호를 출력한다. 그리고, R 유니트는 A 유니트로부터의 입력에 의해 응답하고 입력패턴의 분류에 대응하는 신호를 출력한다[그림 4.6].

A 유니트와 R 유니트에는 비선형 입출력함수의 일종인 역치함수를 사용한다. S 유니트로부터 A 유니트로의 결합의 강도는 최초의 부여된 값으로 정해지는 것에 비해 A 유니트에서 R 유니트로는 학습에 의해 변화한다.

b. 다층 퍼셉트론

다층 퍼셉트론은 입력층과 출력층 사이에 하나 이상의 중간층이 존재하는 신경망으로 [그림 4.7]에 나타낸 것과 같은 계층구조를 가진다. 이 때 입력층과 출력층 사이에 중간층을 은닉층이라 부른다. 네트워크는 입력층, 은닉층, 출력층 방향으로 연결되어 있으며, 각 층내의 연결과 출력층에서 입력층으로의 직접적인 연결은 존재하지 않는 전방향의 네트워크이다.



[그림 4.7] 다층 퍼셉트론의 구조

다층 퍼셉트론은 단층 퍼셉트론과 유사한 구조를 가지고 있지만, 중간층과 각 유니트의 입출력 특성을 비선형화 함으로써 네트워크의 능력을 향상시켜 단층 퍼셉트론의 여러 가지 단점들을 극복했다. 다층 퍼셉트론은 층의 개수가 증가할수록 퍼셉트론이 형성하는 결정 구역의 특성은 더욱 고급화된다. 즉, 단층일 경우 패턴공간을 두 구역으로 나누어주고, 2층인 경우 오목한 개구역 또는 오목한 폐구역을 형성하며, 3층인 경우 어떠한 형태의 구역도 형성할 수 있다.

본 논문에서 제안하는 시스템의 분류기로 기계 학습 기반 분류기를 사용했으며 대표적인 인공 신경망인 다층퍼셉트론을 사용하여 이를 구현하였다.

인공 신경망의 대표적인 기계 학습 알고리즘인 다층퍼셉트론은 대부분의 패턴 인식 문제에 대해 안정적인 성능을 보이며, 일단 학습이 끝나면 응용 단계에서는 매우 빠르게 결과를 출력한다. 대부분의 다층퍼셉트론은 백프로퍼게이션(back propagation) 학습 알고리즘을 사용하여 학습시킬 수 있다[31].

V. 실 험

A. 제안하는 시스템 구조도

제안하고자 하는 효과적인 유전자 선택 방법의 현실적 구현을 위해서는 기존의 암 분류를 위한 유전자 발현 분석 시스템의 구조를 변경해야 한다. 변경된 시스템의 주요 단계는 유의한 유전자 선택 단계, 유전자 선택 방법 조합 단계, 클래스 분류 단계, 성능 평가 네 단계로 이루어진다.

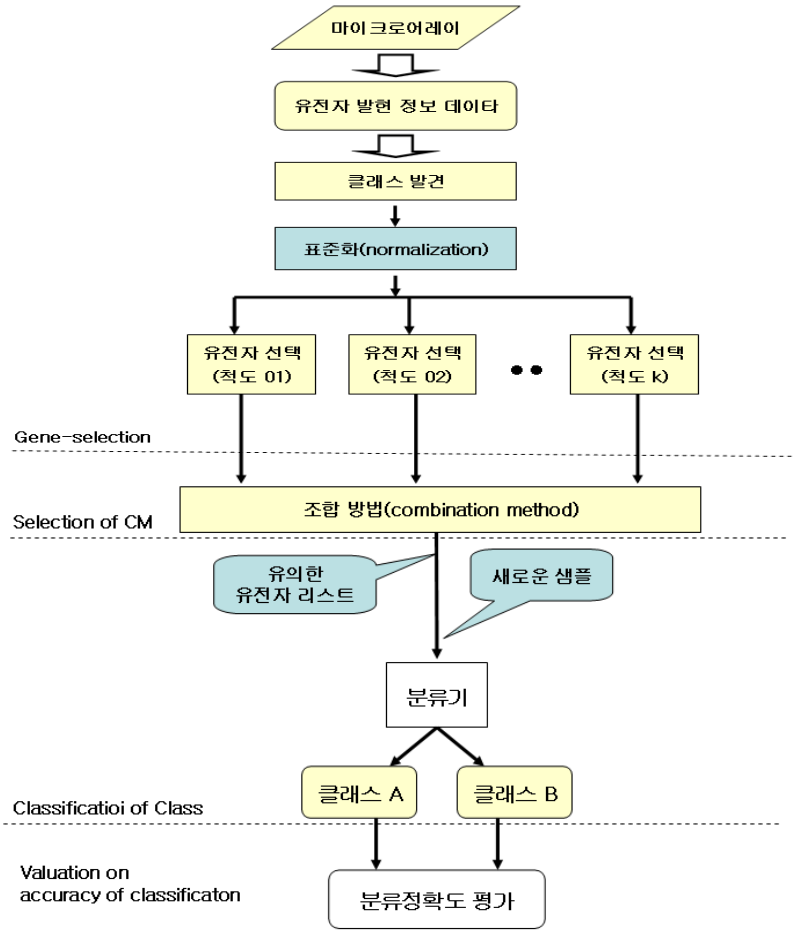
먼저 마이크로어레이로부터 유전자 발현 데이터를 획득한다. 클래스 발견 단계를 거쳐 이상 유전자 모델이 확정되고 나면, 각각의 유전자 발현 데이터를 표준화 후 각 유사성 척도를 사용하여 이상 유전자 모델과의 유사성 정도를 정량적으로 평가한다. 이들 중 여러 개의 척도(최소 2개 이상의 척도)에서 정보력이 있는 유용한 유전자로 평가 받은 유전자들을 정량화된 유용성 정도에 따라 상위 부분을 모아 정보력이 있는 유전자 목록을 확정한다.

마이크로어레이 데이터 분류에 있어서, 정보력이 있는 유전자들의 목록은 그 자체로서도 생물학적으로 중요한 의미를 지닌다. 따라서 이 목록을 별도로 추출한 정보력이 있는 유전자들은 암의 분류와 예측을 통한 진단 분야뿐만 아니라 정확한 분류 이후의 치료 분야에도 긍정적으로 활용될 수 있다. 이러한 정보력이 있는 유전자들은 분류와의 관련 정도가 커서 정확한 분류를 위해 효과적으로 이용될 수 있어야 하고 그 신뢰도 또한 높아야 하며 목록의 일관성을 유지할 수 있어야 한다.

따라서 이를 위해 본 논문 3장의 실험에서 얻어진 결과에 따라 선형 패턴뿐만 아니라 비선형 패턴도 제거해 주는 *Lowess* 표준화를 수행 한 후 단일 유사성 척도를 이용하여 유전자를 선택하던 기존 방식의 문제점을 보완하고 상대적으로 낮은 비용으로 특정 클래스를 분류하는 데에 연관이 크고 신뢰도가 높은 정보력이 있는 유전자 선택을 위한 조합 방법 알고리즘을 제안하였다. 제안된 알고리즘을 사용하여 일관성 있는 양질의 목록을 구성하고 다층퍼셉트론을 사용하여 분류 정확도를

측정하는 마이크로어레이 분류 시스템을 설계하고 구현하였다.

[그림 5.1]은 본 논문에서 제안된 시스템의 구조도를 나타낸 것이다.



[그림 5.1] 제안하는 분류 시스템 구조도

B. 실험 데이타

본 논문에서는 암에 걸린 흰쥐와 암에 걸리지 않은 흰쥐의 각 뇌신경조직 부위에서 획득한 유전체의 조절 인자를 각각 *Cy5, Cy3*로 염색한 다음, 2400개 이상의 알려진 유전체와 1700여개의 새로운 유전체가 찍힌 유리칩을 이용한 *cDNA* 마이크로어레이 실험에서 획득한 마이크로어레이 데이터를 사용하였다.

제작된 마이크로어레이에 대해 *Lowess* 표준화 후 기존의 유전자 선택 방법과 3장에서 제안된 조합 방법을 적용하여 정보력이 있는 유전자를 선택하고 그 목록을 만들어 그에 따른 분류 성능을 다층퍼셉트론 기반 분류기를 통해 정량적으로 비교 평가한다. 실험에서 기존의 유전자 선택 방법을 본 논문에서 제안한 조합 방법을 이용하여 조합함으로써, 이러한 방법이 보다 정보력이 있는 유전자 목록을 별도로 분리 생성할 수 있고 이 목록이 분류 성능 역시 향상 시킬 수 있음을 보이고자 한다.

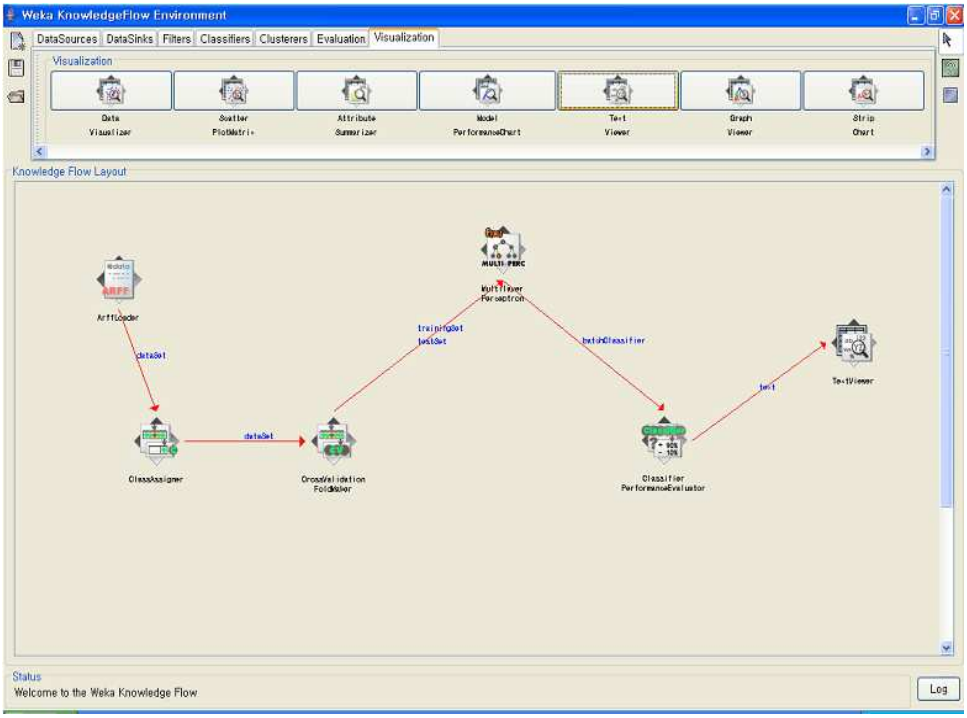
C. 실험 결과 및 고찰

암에 걸린 쥐 배아와 암에 걸리지 않는 쥐 배아의 각 뇌신경조직 부위에서 분리된 신경기간세포의 *cDNA* 마이크로어레이 실험에서 획득한 마이크로어레이 데이터 대하여 *Lowess* 표준화 후 기존의 단일 유사성 척도 3가지를 사용한 유전자 선택 방법과 이들을 조합한 유전자 선택 방법 4가지를 이용하여 정보력이 있는 유전자를 선택하고 목록을 만들어 다층퍼셉트론을 기반으로 하는 분류기를 통해 학습과 테스트를 한 분류 결과를 10-*fold cross-validation*을 사용하여 정확도를 측정하고 서로 비교분석하였다. 또한 조합 방법에 따른 성능대비 시간 분석을 위해 평균 수행 시간을 계산하였다.

여기서 말하는 *cross-validation*은 교차검증을 뜻하는 것이다. *k-fold cross-validation*은 *k* 개의 부분샘플로 데이터집합을 나눈다. 그런 다음 *k-1*개의 부분 샘플은 *train*(훈련)데이터로 나머지 한 개의 부분샘플은 *test*(테스트)데이터로 사용한다. 그러므로 *10-fold cross-validation*은 주어진 *dataset*을 10개의 서브 데이터 셋으로 나누어 2-9번 서브 데이터 셋은 *train* 데이터로 1번 서브 데이터 셋은 *test* 데이터로 사용하여 모델을 만들고, 다시 1,3-9번 서브 데이터 셋은 *train* 데이터로 2번 서브 데이터 셋은 *test* 데이터로 사용하여 모델을 만들었다. 이런 식으로 하나의 서브 데이터 셋을 돌아가며 테스트 데이터 셋으로 사용하여 여러 모델을 만들어 평균을 낸다. 여러 모델이 있으면 새로운 데이터가 들어왔을 때 각 모델 별로 적용시켜 나온 결과들 중 많은 것을 택한다.

각 해당하는 절에 정량화된 결과 값을 표로 기술하였으며, 표에서 사용된 기호 '-'는 본 논문에서 제안된 조합 방법에 의해 기존의 유사성 척도를 이용한 유전자 선택 방법을 '조합'한 것을 의미한다. '*PC*'는 피어슨 적률 상관 계수(*Pearson correlation coefficient*)를 뜻하고, *SC*는 스피어만 상관 계수(*Spearman correlation coefficient*)를 나타내며, *ED*는 유클리디안 거리 계수(*Euclidean distance*)를 의미한다. 예를 들어, '*PC-ED*'로 표기된 것은 기존의 유사성 척도인 피어슨 적률 상관

계수와 유클리디안 거리 계수를 이용하여 선택된 유전자들을 본 논문에서 제안한 조합 방법에 의해 서로 조합하여 유전자를 새롭게 선택하고 목록을 재구성하였음을 의미한다. 표에 나타나 있는 분류 성능 수치는 해당 조합 조건에서 분류기의 분류 성공률을 의미하며 그 단위는 퍼센트(%)이다. *MSE*(*Mean Square Error*)는 평균제곱 오차를 나타내며, 실제 클래스와 예측한 클래스 차이를 제곱한 결과를 나타내며 이 값이 작을수록 좋은 분류를 나타낸다. [그림 5.2]는 데이터 마이닝 툴 *WEKA*를 이용한 마이크로어레이 분류 시스템을 설계한 그림이다.



[그림 5.2] 마이크로어레이 분류 시스템

1. 선택된 *200*개 유전자에 대한 다양한 조합에 따른 분류 성능

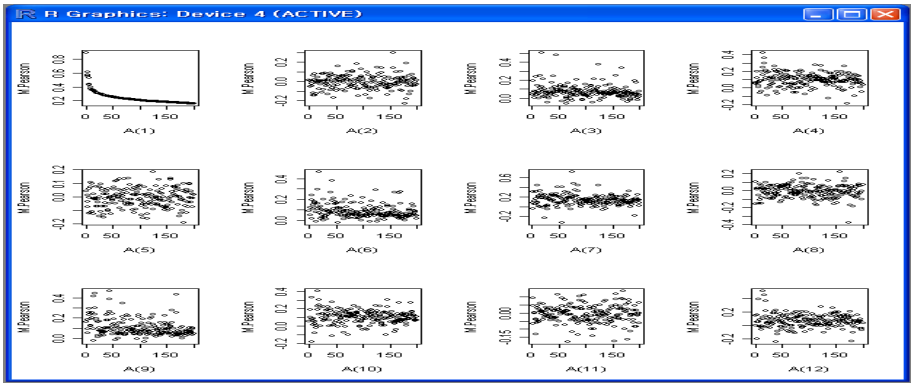
우선 기존의 단일 유사성 척도를 이용하여 상위 200개 유전자를 선택하고 목록을 만들어 분류 성능을 실험한 결과는 [표 5.1]과 같다. 이를 두 개 이상의 기존 척도 조합에 따른 분류 성능을 비교 평가하기 위한 실험의 대조군으로 한다.

[표 5.1] 상위 200개 유전자에 대한 단일 척도 사용에 따른 분류 성능

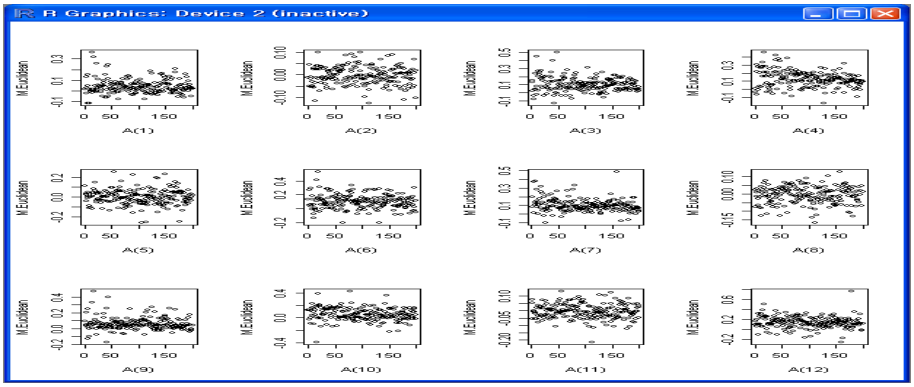
조 합	PC	ED	SC
분류 성능	68.14	67.35	58.24
MSE	0.46	0.48	0.52

기존의 유사성 척도인 피어슨 적률 상관관계 척도를 단일하게 사용하여 유전자를 선택하고 목록을 생성하여 활용하는 경우 다층퍼셉트론 기반의 분류기를 통해 분류 성능을 확인한 결과 68.14%의 분류 성공률을 보였다. 같은 방법으로 유클리디언 거리 척도를 사용하였을 경우 67.35%의 분류 성능을 보였으며, 스피어만 상관관계 척도를 사용하였을 때에는 58.24%의 가장 낮은 분류 성능을 나타내었다. 기존의 단일 유사성 척도를 사용하여 유전자 목록을 생성한 뒤 실험 한 경우 대부분 낮은 분류성능을 나타내었다.

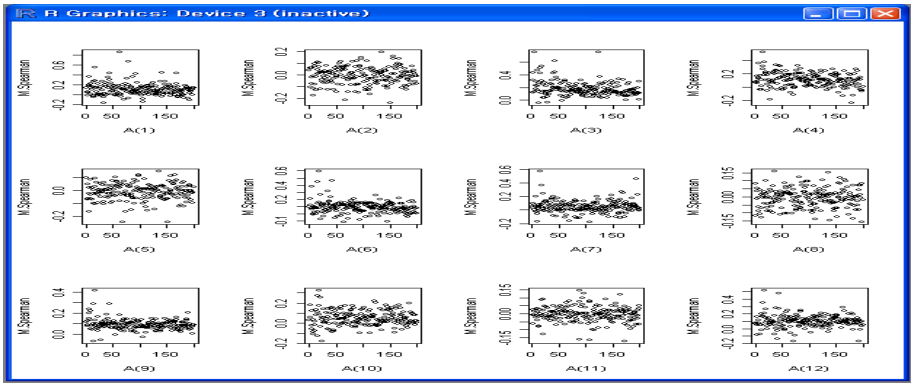
[그림 5.3]은 *Lowess* 표준화 후 기존의 단일 유사성 척도를 이용하여 선택한 상위 200개 유전자의 *M-A plot*의 일부분이다.



(a) 피어슨 적률 상관관계 계수



(b) 유클리디언 거리 계수



(c) 스피어만 상관관계 계수

[그림 5.3] 단일 척도 사용에 따른 상위 200개 유전자 *M-A plot*

그러나 이러한 기존의 단일 유사성 척도를 조합 방법에 의해 조합하여 보다 정보력이 있는 유전자 목록을 생성한 뒤 실험한 경우 분류기에서 향상된 분류 성능을 나타내었다. 마이크로어레이 데이터에 대해 기존의 유사성 척도를 단일하게 사용한 유전자 선택 방법 중 두 가지를 조합하여 실험한 결과, 관찰된 분류 성능은 [표 5.2]와 같다.

[표 5.2] 상위 200개 유전자에 대한 2개 척도 조합에 따른 분류 성능

조 합	PC-ED	PC-SC	ED-SC
분류 성능	82.35	72.29	78.14
MSE	0.24	0.38	0.32

기존의 유사성 척도인 피어슨 적률 상관관계 척도를 단일하게 사용하여 실험 하였을 때에는 그 분류성능이 68.14%로 나타났다. 유클리디안 거리 척도를 사용하여 같은 방법으로 실험한 결과 67.35%의 분류 성능을 보였다. 그러나 이 두 가지 기존 유사성 척도를 사용한 유전자 선택 방법을 조합 방법에 의해 조합하여 실험한 결과 *PC-ED*의 경우 82.35%의 분류 성능을 보였다. 피어슨 적률 상관관계 척도와 스피어만 상관관계 척도를 조함하여 실험한 *PC-SC*의 경우 역시 72.29%의 분류 성능을 내면서, 단일 유사성 척도를 사용하여 실험한 결과보다 높은 분류 성능을 나타냈다.

[표 5.3] 상위 200개 유전자에 대한 3개 척도 조합에 따른 분류 성능

조 합	PC-ED-SC
분류 성능	82.35
MSE	0.26

기존의 유사성 척도를 사용하여 유전자를 선택하는 방법 세 가지를 모두 조합하는 경우에 [표 5.3]과 같은 분류 성능 향상을 보였다. 이러한 분류 성능 향상은 두 가지 척도 즉, 피어슨 적률 상관 계수와 유클리디안 거리 계수를 조합하는 경우에서 가장 현저하게 나타나고 세 가지 척도를 조합하는 경우에는 다소 소극적으로 나타났다. 또한 0.26%의 *MSE*를 보여 두 가지 척도를 조합하는 경우보다 낮은 신뢰성을 보였다.

2. 선택된 *90*개 유전자에 대한 다양한 조합에 따른 분류성능

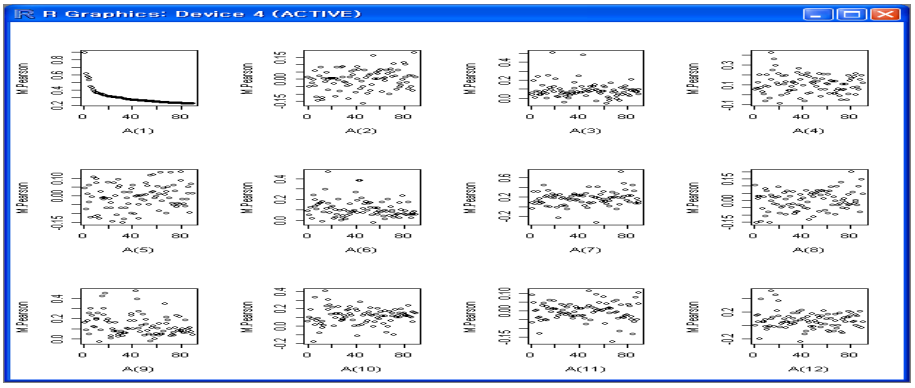
기존의 단일 유사성 척도를 이용하여 상위 90개 유전자를 선택하고 목록을 만들어 분류 성능을 실험한 결과는 [표 5.4]과 같다. 이를 두 개 이상의 기존 척도 조합에 따른 분류 성능을 비교 평가하기 위한 실험의 대조군으로 한다.

[표 5.4] 상위 90개 유전자에 대한 단일 척도 사용에 따른 분류 성능

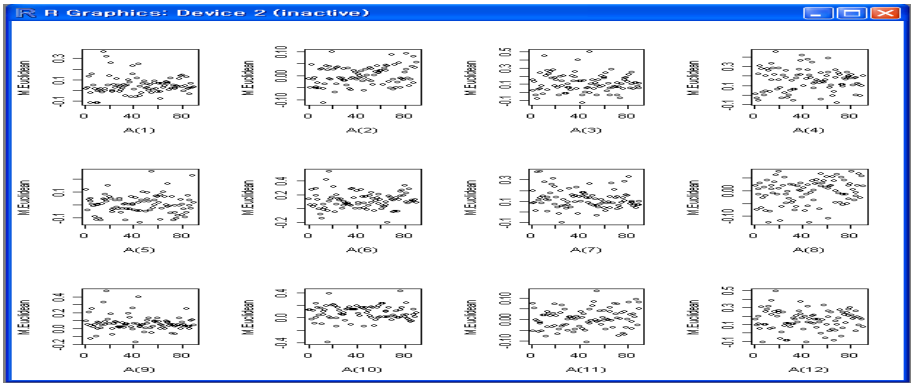
조 합	PC	ED	SC
분류 성능	80.65	79.23	76.09
MSE	0.25	0.28	0.33

기존의 유사성 척도인 피어슨 적률 상관관계 척도를 단일하게 사용하여 상위 유전자 90개를 선택하고 목록을 생성하여 활용하는 경우 다층퍼셉트론 기반의 분류기를 통해 분류 성능을 확인한 결과 80.65%의 분류 성공률을 보였다. 같은 방법으로 유클리디안 거리 척도를 사용하였을 경우 79.23%의 분류 성능을 보였으며, 스피어만 상관관계 척도를 사용하였을 때에는 76.09%의 분류 성능을 나타내었다.

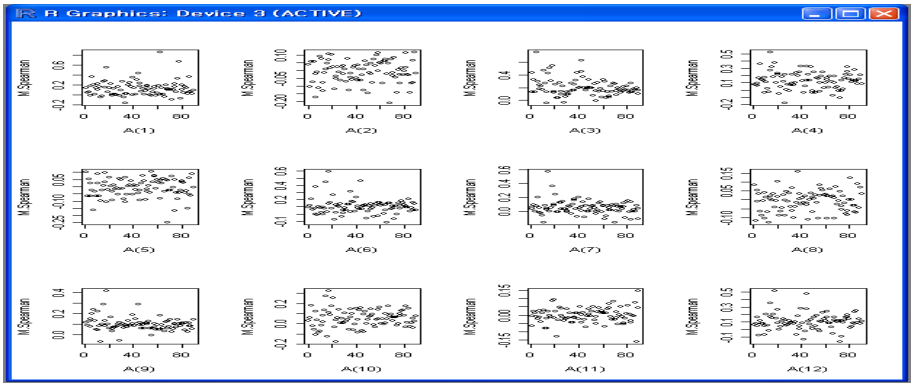
[그림 5.4]은 *Lowess* 표준화 후 기존의 단일 유사성 척도를 이용하여 선택한 상위 90개 유전자의 *M-A plot*의 일부분이다.



(a) 피어슨 적률 상관관계 계수



(b) 유클리디안 거리 계수



(c) 스피어만 상관관계 계수

[그림 5.4] 단일 척도 사용에 따른 상위 90개 유전자 *M-A plot*

그러나 이러한 기존의 단일 유사성 척도를 조합 방법에 의해 조합하여 보다 정보력이 있는 유전자 목록을 생성한 뒤 실험한 경우 분류기에서 향상된 분류 성능을 나타내었다. 또한 상위 200개 유전자의 경우 보다 전반적인 분류 성능이 향상되는 경향을 보였다. 이것은 높은 발현현상을 보이는 유의한 유전자를 사용한 것에 기인한 것으로 추정된다. 마이크로어레이 데이터에 대해 기존의 유사성 척도를 단일하게 사용한 유전자 선택 방법 중 두 가지를 조합하여 실험한 결과 관찰된 분류 성능은 [표 5.5]와 같다.

[표 5.5] 상위 90개 유전자에 대한 2개 척도 조합에 따른 분류 성능

조 합	PC-ED	PC-SC	ED-SC
분류 성능	88.78	86.14	81.12
MSE	0.11	0.14	0.22

기존의 유사성 척도인 피어슨 적률 상관관계 척도를 사용하여 실험하였을 때에는 그 분류 성능이 80.65%로 나타났다. 유클리디안 거리 척도를 사용하여 같은 방법으로 실험한 결과 79.23%의 분류 성능을 보였다. 그러나 이 두 가지 기존 유사성 척도를 사용한 유전자 선택 방법을 조합 방법에 의해 조합하여 실험한 결과 *PC-ED*의 경우 88.78%의 향상된 분류 성능을 보였다. 유클리디안 거리 척도와 스피어만 상관관계 척도를 조합하여 실험한 *ED-SC*의 경우 역시 81.12%의 분류 성능을 나타내면서, 유클리디안 거리 척도와 스피어만 상관관계 척도를 단일하게 사용하여 실험한 결과보다 높은 분류 성능을 나타냈다. 기존 유사성 척도를 사용하여 유전자를 선택하는 방법 두 가지를 조합하는 경우 한 가지 척도를 사용하여 실험을 했을 때보다 전반적으로 만족할 만한 분류 성능 향상을 보였다.

[표 5.6] 상위 90개 유전자에 대한 3개 척도 조합에 따른 분류 성능

조 합	PC-ED-SC
분류 성능	80.65
MSE	0.20

기존의 유사성 척도를 사용하여 유전자를 선택하는 방법 세 가지를 모두 조합하는 경우에도 [표 5.6]과 같은 분류 성능 향상을 보였다. 이러한 분류 성능 향상은 두 가지 척도를 모두 조합하는 경우 *PC-ED* 조합에서 가장 현저하게 나타나고 세 가지 척도를 조합하는 경우에는 두 가지 척도를 조합하는 경우보다 낮은 분류 성능을 보였다. 그러나 상위 90개 유전자를 사용한 경우에도 역시 중요한 사실은 기존의 유사성 척도를 단일하게 사용하여 유전자를 선택하고 목록을 생성해서 실험을 하는 경우보다 기존 유전자 선택 방법 중 두 개 이상을 조합 방법에 의해 조합하면 보다 정보력이 있는 유전자 목록이 생성되고 실험을 수행한 경우에 그 분류 성능이 향상되었다는 것이다.

3. 선택된 *50*개 유전자에 대한 다양한 조합에 따른 분류성능

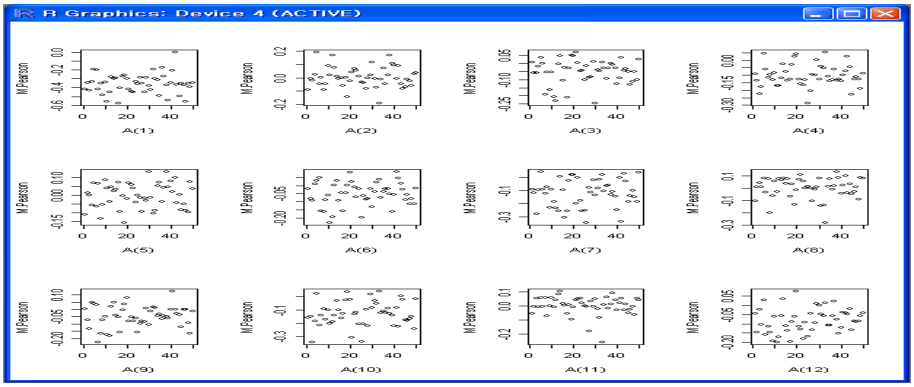
기존의 단일 유사성 척도를 이용하여 상위 50개 유전자를 선택하고 목록을 만들어 분류 성능을 실험한 결과는 [표 5.7]과 같다. 이를 두 개 이상의 기존 척도 조합에 따른 분류 성능을 비교 평가하기 위한 실험의 대조군으로 한다.

[표 5.7] 상위 50개 유전자에 대한 단일 척도 사용에 따른 분류 성능

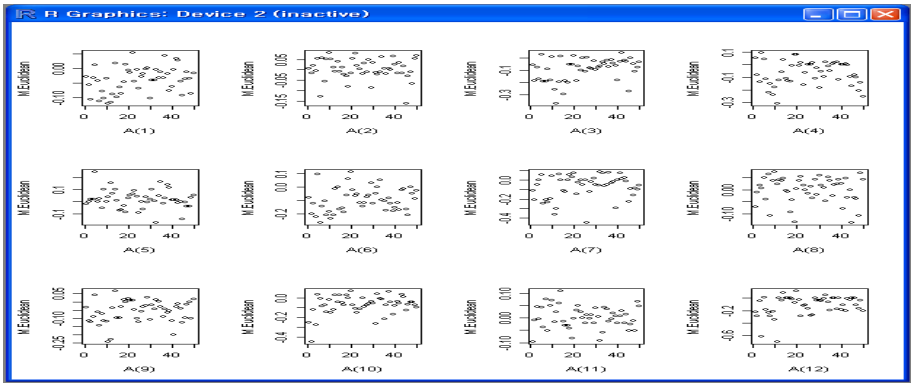
조 합	PC	ED	SC
분류 성능	88.43	86.23	82.41
MSE	0.17	0.19	0.23

기존의 유사성 척도인 피어슨 적률 상관관계 척도를 단일하게 사용하여 50개의 상위 유전자를 선택하고 목록을 생성하여 활용하는 경우 다층퍼셉트론 기반의 분류기를 통해 분류 성능을 확인한 결과 88.43%의 분류 성공률을 보였다. 같은 방법으로 유클리디안 거리 척도를 사용하였을 경우 86.23%의 분류 성능을 보였으며, 스피어만 상관관계 척도를 사용하였을 때에는 82.41%의 분류 성능을 보였다.

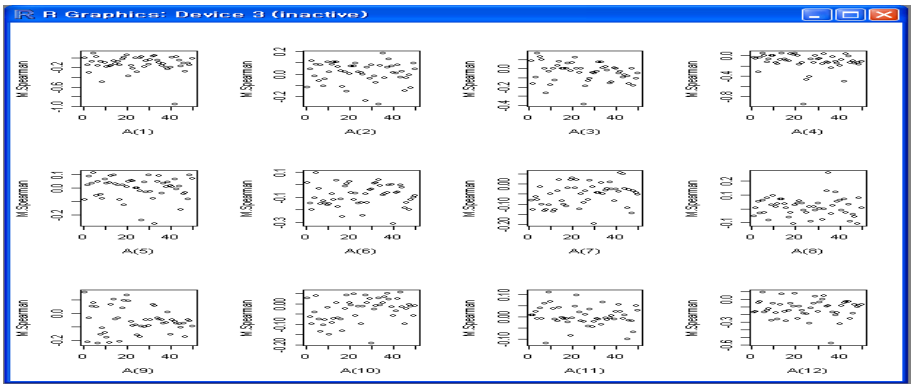
[그림 5.5]은 *Lowess* 표준화 후 기존의 단일 유사성 척도를 이용하여 선택한 상위 50개 유전자의 *M-A plot*의 일부분이다.



(a) 피어슨 적률 상관관계 계수



(b) 유클리디안 거리 계수



(c) 스피어만 상관관계 계수

[그림 5.5] 단일 척도 사용에 따른 상위 50개 유전자 *M-A plot*

그러나 이러한 기존의 단일 유사성 척도를 조합 방법에 의해 조합하여 보다 정보력이 있는 유전자 목록을 생성한 뒤 실험한 경우 분류기에서 향상된 분류 성능을 보였다. 마이크로어레이 데이터에 대해 기존의 유사성 척도를 단일하게 사용한 유전자 선택 방법 중 두 가지를 조합하여 실험한 결과 관찰된 분류 성능은 [표 5.8]과 같다.

[표 5.8] 상위 50개 유전자에 대한 2개 척도 조합에 따른 분류 성능

조 합	PC-ED	PC-SC	ED-SC
분류 성능	94.04	90.18	88.09
MSE	0.08	0.10	0.17

기존의 유사성 척도인 피어슨 적률 상관관계 척도를 단일하게 사용하여 실험 하였을 때에는 그 분류 성능이 88.43%로 나타났다. 유클리디안 거리 척도를 사용하여 같은 방법으로 실험한 결과 86.23%의 분류 성능을 보였다. 그러나 이 두 가지 기존 유사성 척도를 사용한 유전자 선택 방법을 조합 방법에 의해 조합하여 실험한 결과 *PC-ED*의 경우 94.04%의 향상된 분류 성능을 보이고 있다. 피어슨 적률 상관관계 척도와 스피어만 상관관계 척도를 조합하여 실험한 *PC-SC*의 경우 역시 90.18%의 분류 성능을 나타내면서, 피어슨 적률 상관관계 척도와 스피어만 상관관계 척도를 단일하게 사용하여 실험한 결과보다 높은 분류 성능을 보였다. 기존 유사성 척도를 사용하여 유전자를 선택하는 방법 두 가지를 조합하는 경우 한 가지 척도를 사용하여 실험을 했을 때보다 전반적으로 만족할 만한 분류 성능 향상을 보였다.

[표 5.9] 상위 50개 유전자에 대한 3개 척도 조합에 따른 분류 성능

조 합	PC-ED-SC
분류 성능	82.24
MSE	0.17

기존의 유사성 척도를 사용하여 유전자를 선택하는 방법 세 가지를 모두 조합하는 경우에도 [표 5.9]과 같은 분류 성능 향상을 보인다. 그러나 이러한 분류 성능 향상은 두 가지 척도를 조합하는 경우에 가장 현저하게 나타나고 세 가지 이상의 척도를 조합하는 경우에서는 다소 소극적으로 나타났다. 이는 하나의 유전자 선택 방법만으로는 분류하고자 하는 해 공간을 모두 포함하지 못할 가능성이 있으나, 많은 유전자 선택 방법의 조합이 오히려 포함하지 않아야 할 해 공간까지 포함하는 경우 분류기의 분류 성능을 상대적으로 저하시킬 수도 있을 것으로 추정된다.

마이크로어레이 데이터에 대해 전체적으로 *PC*, *PC-ED*가 포함된 조합이 다층퍼셉트론 기반 분류기의 분류 성능을 상대적으로 크게 향상시키는 것으로 나타났으며, *SC*가 포함된 조합은 상대적으로 분류 성능 향상에 긍정적인 영향을 제공하지 못하는 것으로 나타났다. 즉, 전자에 해당하는 조합이 보다 높은 신뢰성 있고 일관성이 유지된 정보력이 있는 유전자 목록을 생성할 수 있다고 볼 수 있다.

이러한 분류 성능의 향상은 전반적으로 두 종류의 척도를 조합하였을 때 가장 현저하게 나타나고, 세 종류 척도를 조합하였을 때에는 소극적으로 나타나는 것으로 관찰되었다. 이는 많은 유전자 선택 방법의 조합은 오히려 포함하지 않아야 할 해 공간까지 포함하는 경우 분류기의 분류 성능을 상대적으로 저하시킨다는 것을 관찰 할 수 있었다. 또한 선택된 유전자의 수가 50개일 때 상대적으로 가장 높은 분류 성능을 보였다.

결과적으로, 제안한 시스템과 이를 위한 조합 방법이 전체적인 관점에서 신뢰성과 일관성이 유지되는 분리 가능한 정보력이 있는 유전자 목록을 생성하고 있음을

알 수 있다. 기존의 유사성 척도를 단일하게 사용하여 유전자를 선택하고, 목록을 생성해서 실험하는 경우보다 기존 유전자 선택 방법 두 개 이상을 조합 방법에 의해 조합하여 보다 정보력이 있는 유전자 목록을 생성하고 실험을 수행한 경우에 그 분류 성능이 향상되었던 실험 결과도 이를 뒷받침한다. 따라서 마이크로어레이 데이터를 이용한 암 분류에 있어서 암의 분류와 예측을 통한 진단 분야뿐만 아니라 정확한 분류 이후의 치료 분야에도 긍정적인 동기 부여를 하기 위해서 본 논문에서 제안한 시스템과 조합 방법은 효과적으로 적용될 수 있을 것으로 생각된다.

4. 평균 실행 시간

두 개 이상의 조합 방법으로 선택되어진 유전자 수에 따른 분류 성능의 결과가 대부분 좋게 나왔기 때문에 조합 방법에 따른 성능 대비 시간을 계산하기 위해 평균 실행 시간을 계산하였다.

이는 조합 방법을 통해 선택된 유전자 수에 따라 모델을 구축하는 시간과 *10-fold cross-validation*하는 시간을 합한 총 평균시간이다. [표 5.10]은 선택된 유전자 수에 따라 조합 방법 별로 측정한 것이다. 그 단위는 (초)이다.

선택된 유전자의 수가 50개 일 때 모든 조합 방법의 평균 실행 시간이 짧게 걸렸다. 또한 선택된 유전자의 수가 많아질수록 시간이 더 많이 걸리는 것으로 나타났다. *PC-ED*의 조합의 경우 다른 척도 조합보다 실행시간이 빨랐으며, 세 종류 척도를 조합하였을 때에는 두 가지 척도를 조합하였을 때보다 더 많은 시간이 걸리는 것으로 나타났다.

[표 5.10] 평균 실행 시간

	PC	ED	SC	PC-ED	PC-SC	ED-SC	PC-ED-SC
Gene = 50	12.38	13.11	15.22	12.36	13.08	13.24	13.94
Gene = 90	70.01	71.38	74.24	69.04	71.94	72.42	80.23
Gene = 200	102.95	111.03	111.1	102.88	110.77	110.78	110.89

Ⅵ. 결론 및 향후 연구과제

오늘날 인간 genome 프로젝트와 같은 종합적인 연구의 궁극적 목적을 달성하기 위해서는 이들 연구로부터 획득한 대량의 관련 데이터에 대해 새로운 현실적 의미를 부여할 수 있어야 한다. 이러한 맥락에서 유전자 발현 분석 시스템과 염기 서열 분석 시스템의 구축이 새롭게 주목을 받고 있다. 이러한 응용의 현실적 시도를 가능하게 했던 것은 마이크로어레이 기술의 발전이다. 그러나 마이크로어레이 실험에서 얻어진 원 자료에는 다양한 종류의 잡음이 포함되어 있다. 표준화 과정은 본격적인 마이크로어레이 자료의 통계적 분석 이전에 실시되는 가장 중요한 전처리(*pre-processing*) 분석 방법이다. 초기에는 주로 간단한 *Global* 표준화 방법이 많이 사용되었지만 점차 비선형 *Lowess* 방법 같은 복잡한 방법이 많이 사용되고 있다.

최근 마이크로어레이는 유전자 발현 정보를 기반으로 암의 분류와 예측을 통한 진단 분야에도 활용되고 있다. 그러나 이의 현실적 구현을 지원하기 위한 전산학적 관점에서 볼 때, 마이크로어레이로부터 얻어지는 데이터 집합은 기존 데이터 마이닝에서 사용하던 데이터 집합과는 달리 샘플의 수는 극히 적고, 각 샘플의 특징 즉 유전자 수는 매우 많다는 특징을 지닌다. 따라서 현재의 마이크로어레이 기술을 이용해서 효과적으로 암을 정확하게 분류하기 위해서는 특정 암의 분류와 밀접하게 관련이 있는 유용한 유전자를 선택하는 과정이 필수적이다. 이러한 유용한 유전자 선택 작업은 분류 성능을 높이는 데에도 기여하지만, 그 자체로서도 생물학적으로 매우 중요한 의미를 지닌다. 마이크로어레이 데이터를 이용한 암 분류에 있어서, 별도로 분리된 정보력이 있는 유전자들은 암의 분류와 예측을 통한 진단 분야뿐만 아니라 정확한 분류 이후의 치료 분야에도 긍정적인 동기를 부여를 할 수 있다는 점에서 매우 중요한 의미를 지닌다. 따라서 그 활용을 위해 이러한 정보력이 있는 유전자들의 목록을 별도로 추출해 낼 수 있어야 한다.

이에 본 논문에서는 Lowess 표준화 후 정보력이 있는 유전자 목록을 조합하는 시스템을 고안하고 보다 분류 성능을 향상 시킬 수 있는 조합 방법을 제안하였다. 그리고 제안한 시스템과 방법론의 가능성을 실험을 통해 확인하였다. 조합 방법은 *Lowess* 표준화 후 다수의 기존 유전자 선택 방법으로부터 유용하다고 평가된 유전자 데이터들만을 재조합하여 정보력이 있는 유전자 목록을 생성하도록 디자인되었다. 다층퍼셉트론을 이용하여 마이크로어레이 데이터에 대해 실험을 한 결과 제안한 시스템과 이를 위한 조합 방법이 전체적인 관점에서 신뢰성있는 분리 가능한 정보력이 있는 유전자를 선택하고 목록을 생성해서 실험을 수행한 경우보다 기존 유전자 선택 방법 중 두 개 이상을 조합 방법에 의해 조합하여 보다 정보력이 있는 유전자 목록을 생성하고 실험을 수행한 경우에 그 분류 성능이 향상되었던 실험 결과도 이를 뒷받침한다.

따라서 마이크로어레이 데이터를 이용한 암 분류에 있어서 암의 분류와 예측을 통한 진단 분야뿐만 아니라 정확한 분류 이후의 치료 분야에도 긍정적인 동기 부여를 하기 위해 본 논문에서 제안한 시스템과 조합 방법이 효과적으로 적용될 수 있을 것으로 생각된다.

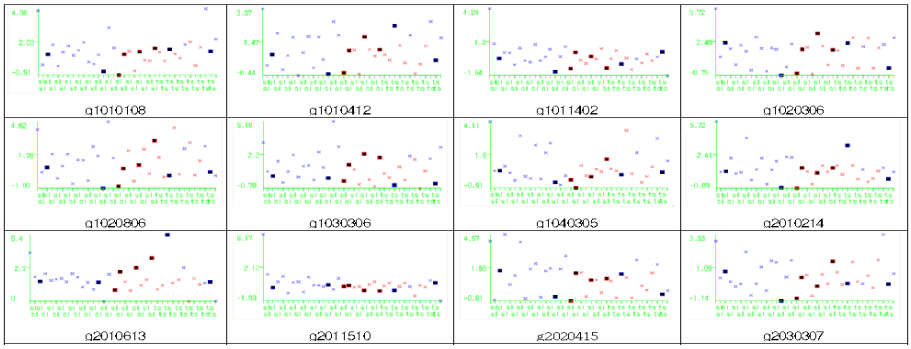
향후 연구 과제는 다양하고 보다 체계적인 많은 데이터의 획득과 분석을 통해 좀 더 효율적인 조합을 찾는 연구가 계속되어야하고, 이와 더불어 실험하는 사람이 그들의 데이터를 면밀히 조사하고 강도에 의존하는 정규화 방법을 일상적으로 적용할 것을 고려해야 할 것이다.

이에 아직 사용해보지 못한 또 다른 정규화 방법과 유의한 유전자 선택 방법을 추가하여 더 많은 연구를 진행하고자 한다.

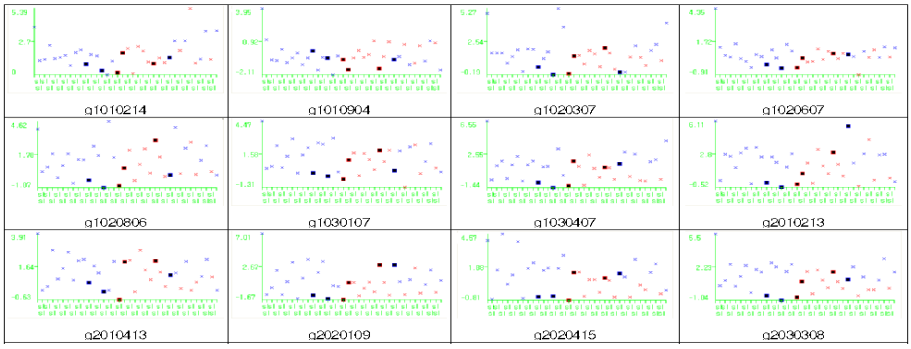
부 록

조합 방법에 의한 유전자들의 각 슬라이드에서 잘못 분류된 정보의 일부분이다. 그림에서 기호 ■는 잘못된 분류를 나타내고, 기호 × 는 올바른 분류를 나타낸다.

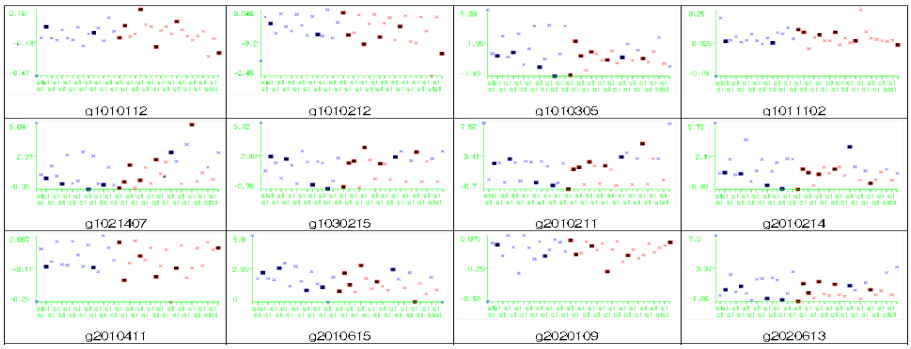
PC-ED



(a) 200개 유전자

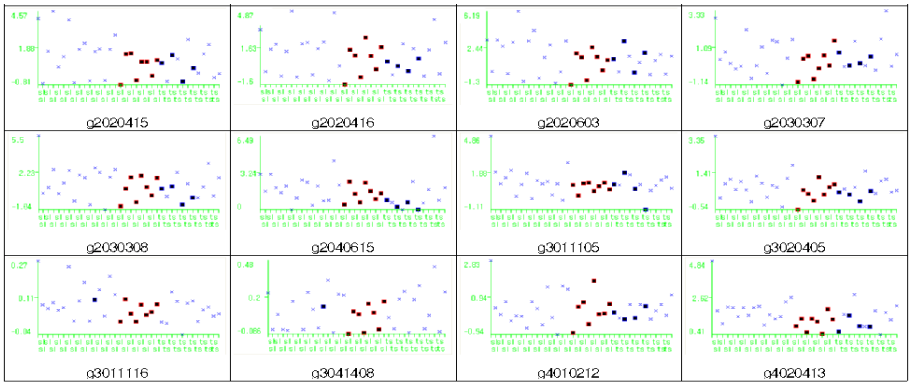


(b) 90개 유전자

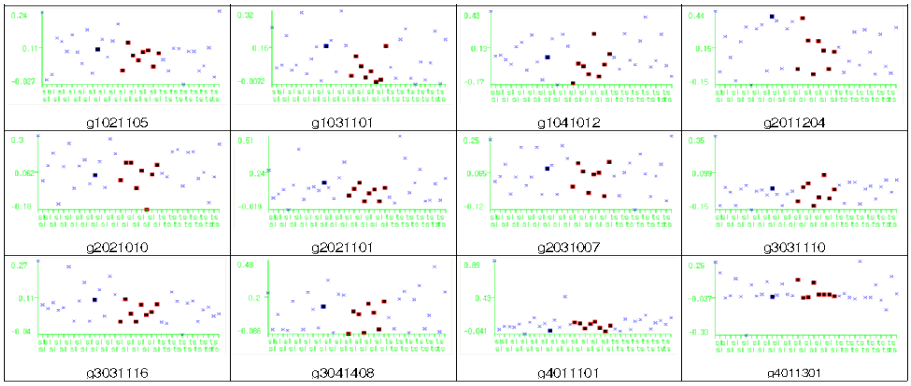


(c) 50개 유전자

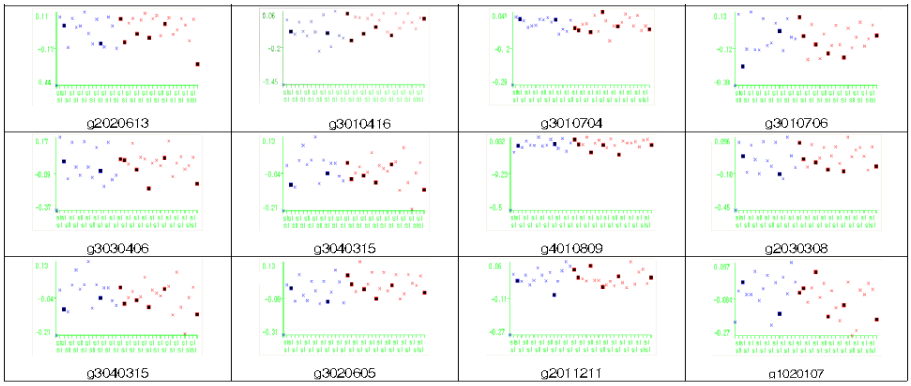
PC-SC



(a) 200개 유전자

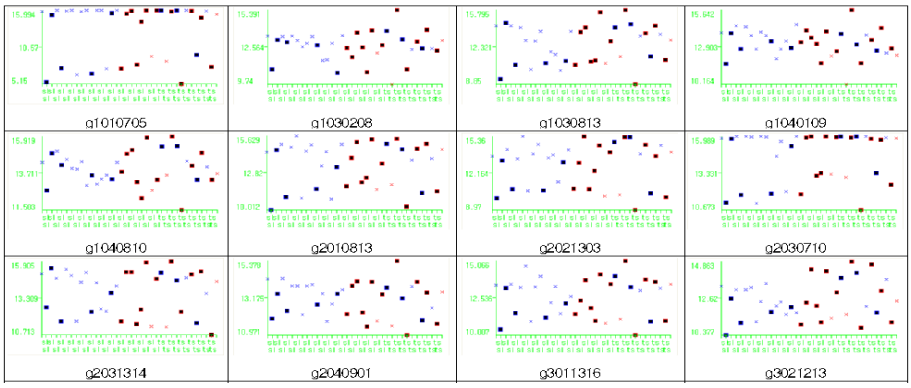


(b) 90개 유전자

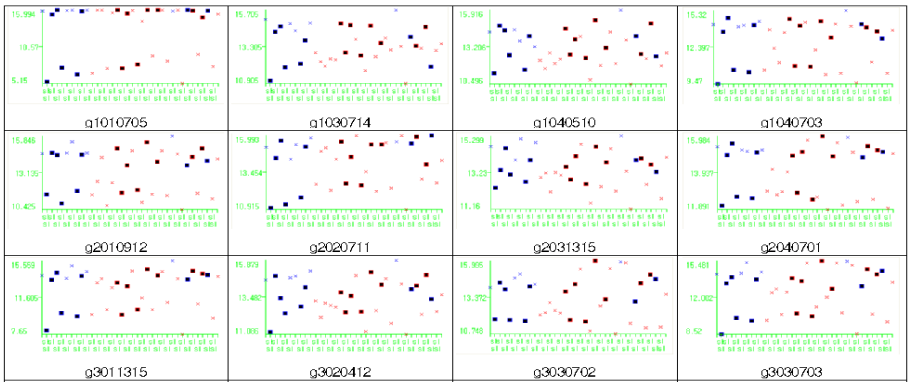


(c) 50개 유전자

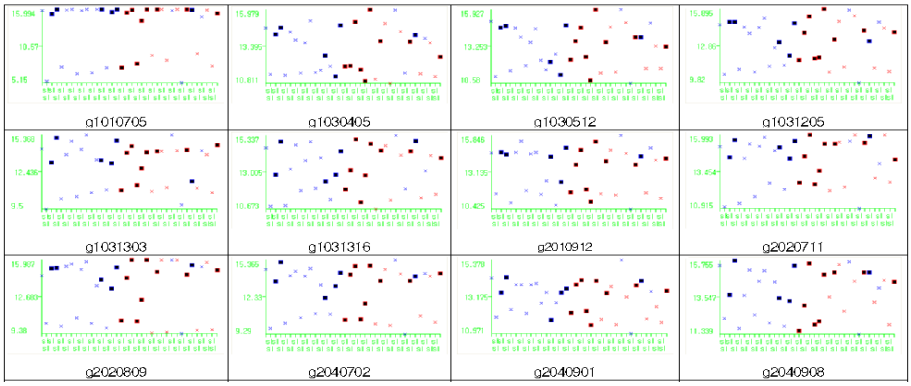
ED-SC



(a) 200개 유전자

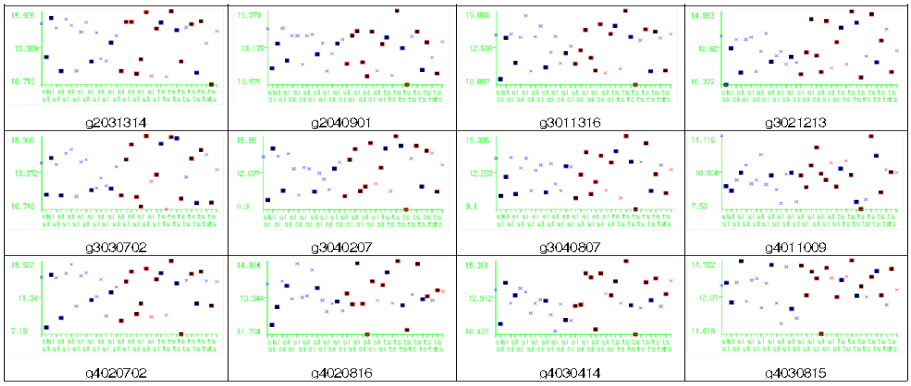


(b) 90개 유전자

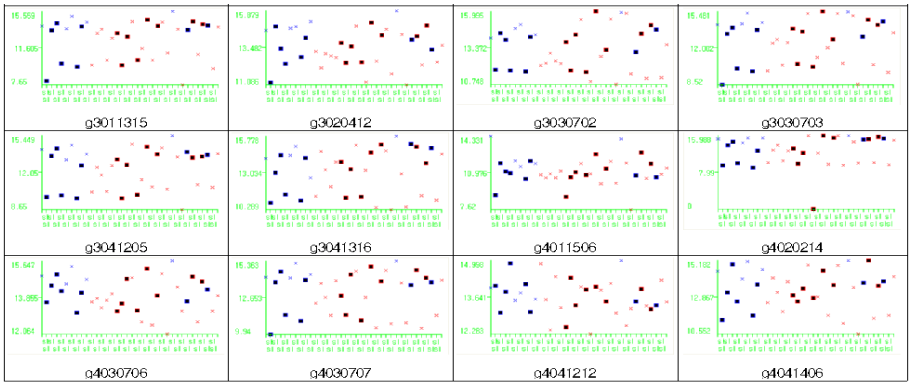


(c) 50개 유전자

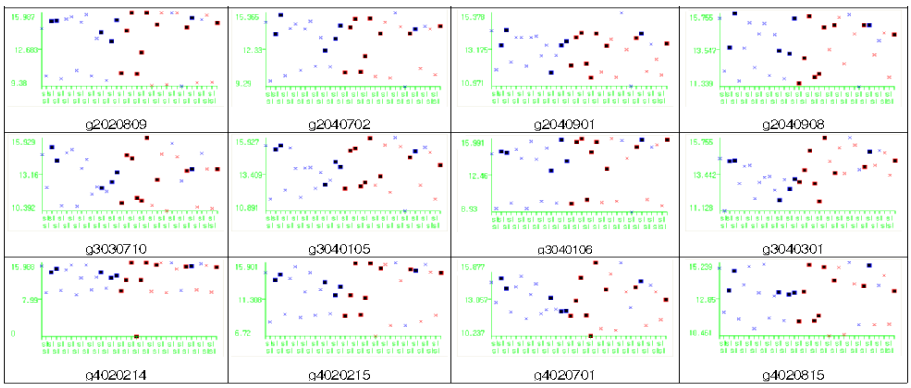
PC-ED-SC



(a) 200개 유전자



(b) 90개 유전자



(c) 50개 유전자

참 고 문 헌

[1] M. Brown, W. Grundy, D. Lin, N. Christianini, C. Sugnet, M. Ares Jr., and D. Haussler, "Support vector machine classification of microarray gene expression data," UCSC-CRL 99-09, Department of Computer Science, University California Santa Cruz, CA, 1999.

[2] S. Dudoit, Fridlyand, and T. i. Speed, "Comparison of discrimination method, for the classification of tumors using gene expression data," Journal of the American Statistical Association, vol. 97, pp. 77-87, 2002.

[3] 원홍희, 조성배, "암 분류를 위한 기계학습 분류기의 성능평가", 한국정보처리학회 추계학술대회, vol. 09, no. 02, 2002.

[4] Dov Stekel, Micorarray Bioinformatics, Cambridge University Press, 2003.

[5] Taesung Park, Sung-Gon Yi, Sung-Hyun Kang, SeungYeoun Lee, Yongsung Lee and Richard Simon, "Evaluation of normalization methods for microarray data," BMC Bioinformatics, Vol. 33, no. 04, pp. 1471-2105, 2003.

[6] Pierre Baldlnd G. Wesley Hatfield, "DNA Microarrays and gene expression," Cambridge University Press, 2002.

[7] M .B. Eisen and P. O. Brown, "DNA arrays for analysis of gene expression," Methods Enzymbol, vol. 303, pp. 179-205, 1999.

[8] G. K., Yang, Y. H., T. P. Speed., "Statistical issues in microarray data analysis," Functional Genomics, Methods and Protocols 224, 111-136, 2003.

[9] Y. Chen, E. R. Dougherty and M. L. Bittner, "Ratio-Based Decision and the Quantitative Analysis of cCNA Microarray Images," Journal of Biomedical Optics, Vol. 2, no. 4, pp. 364-374, 1997.

[10] Y. H. Yang, S. Dudiot, P. Luu, 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 Research, 2002.

[11] Pan, W., "A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments," Bioinformatics, no. 18, pp. 546-554, 2002.

[12] Draghicic S, "Statistical intelligence: effective analysis of high-density microarray data," Drug Discovery Todey, Vol. 7, no. 11, pp. S55-S63, 2002.

[13] Efron, B. , Tibshirani, R., Storey, J. D., Tusher, V. G., "Empirical Bayes analysis of a microarray experiment," J Am Stat Assoc, no. 9, pp. 1151-1160, 2001.

[14] Tusher, V. G., Tibshirani, R., Dhu, G. "Significance analysis of microarrays applied to the ionizing radiation response," Proc natl Acad Sci USA, no. 98. pp. 5116-5121, 2001.

[15] Lonnstedt, I., T. P. Speed., "Replicated microarray data," Stat Sinica no. 12, pp. 31-46, 2002.

[16] Scholkopf, Bernhard, Chris Burges, and Alex J. Smola, "Advences In kernel methods: Support Vector Machines," Cambridge: MIT Press, 1999.

[17] Vladimir N. Vapnik, "The Nature of Statistical Learning Theroy," 1999.

[18] Cristianini, Nello, and John Shawe-Taylor., "An introduction to support Vector machines and other kernel-based learning methoe," Cambridge: Cambridge University Press, 2000.

[19] Cortes, C., Vladimir Vapnik, "Support Vector Networks," Machine Learning, no. 20, pp. 273-297, 1995.

[20] R. Hecht-Nielsen, Neurocomputing, Addison-Wesley Publishing Co., 1989.

[21] J. Hertx, A. Krogh, and R. G. palmer, Introduction to The Theory of Neuralcomputation, Vol. 1, Addison-Wesley Publishing Co., 1991.

[22] Heng Guo and Saul B. Gelfand, "Analysis of Gradient Descent Learning Algorithms for Multilayer Feedforward Neural networks", IEEE Transactions on Circuits and Systems, Vol. 38, no. 8, pp. 883-894, 1991.

[23] Marco Gori and Alberto Tesi, "On the Problem of Local Minima in Backpropagation," IEEE Transactions on Pattern Analysis and Machine

Intelligence, Vol. 14, no. 1, pp. 318-362, 1986.

[24] D. E. Rumelhart, G. E. Hilton, R. J. Williams, "Learning Internal Representation by Error Backpropagation," in Parallel Distributed Processing, Vol. 1, pp.318-362, 1986.

[25] D. E. Rumelhart, R. Durbin, R. Golden, and Y. Chauvin, "Backpropagation: The basic theory," in Backpropagation-Theory, Architectures, and Applications, Applications, Y. Chauvin and D.E. Rumelhart(Eds.), Lawrence Erlbaum Associates, Publishers, 1995.

[26] P. D. Wasserman, Neural Computing: Theory and Practice, New York, Van Nostrand Reinhold, 1989.

[27] D. E. Rumelhart, G. E. Hilton, R. J. Williams, "Learning Representation by Back-Propagation Errors," Nature, Vol. 3239, pp. 533-536, Oct. 1986.

[28] Golub, T. R., Slonim, D. K., Tamayo, P., Huard, D., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., and Lander, E. S., "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," Science, vol. 286, no. 5439, pp. 531-537, 1999.

[29] 강석복, 김영민, 최진갑, 김봉선, 양원섭, “응용통계,” 경문사, 1993.

[30] 이훈영, “통계학," 청람, 2004.

[31] Martin T. Hagan, Howard B. Demuth, and Mark beale, "Neural network design," PWS Publishing Company, 1996.

[32] Robert Gentleman, Vincent J.carey, Wolfgang Huber, Rafael A Irizarry, Sandrin Dudoit, "Bininformatics and Computational Biology Solutions Using R and Bioconductor," Springer.

[33] Ian H. Witten, Eibe Frand, "DATA MINING-Practical Machine Learning Tools and Techniques," Morgan Kaufmann.

[34] Sorin Draghici, "DATA ANALYSIS TOOLS FOR DNA MICROARRAYS," Chanpmain & Hall/CRC

[35] Peter Dalgaard, "Statistics and Computing," Springer.
[36] 오오무라 히도시, “통계를 알면 인생이 달라진다,” 자음과모음.
[37] 서울대학교 통계학과 생물정보통계연구실, “마이크로어레이 자료의 통계적 분석,” 자유아카데미.
[38] John Fox, "An R and S-PLUS COMPANION to APPLIED REGRESSION," SAGE.