



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2008-0030142
(43) 공개일자 2008년04월04일

(51) Int. Cl.

G06F 17/50 (2006.01) G06F 9/455 (2006.01)

(21) 출원번호 10-2006-0095827

(22) 출원일자 2006년09월29일

심사청구일자 2006년09월29일

(71) 출원인

고려대학교 산학협력단

서울 성북구 안암동5가1 고려대학교 내

(72) 발명자

이재원

서울 성북구 안암동 5-1 고려대학교 정경대학

김철민

부산 서구 아미동1가 10번지 부산대학교 의과대학
(뒷면에 계속)

(74) 대리인

현중철

전체 청구항 수 : 총 7 항

(54) 마이크로어레이 통합 분석 방법

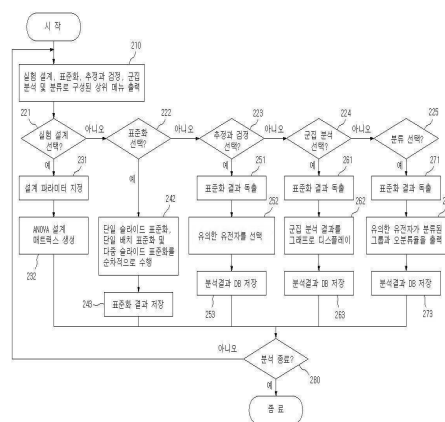
(57) 요약

마이크로어레이 통합 분석 방법이 개시된다.

본 발명은 실험 설계, 표준화, 추정과 검정, 군집 분석 및 분류로 구성된 상위 메뉴의 아이টে에 대하여, 상기 아이টে 중 상기 실험 설계가 선택되면, 염료 교환, 기준설계 또는 루프설계 중 어느 하나의 실험 디자인을 선택하는 인터페이스와 상기 선택된 실험 디자인에 사용될 설계 파라미터를 입력하는 인터페이스를 디스플레이하고, 상기 입력된 설계 파라미터를 이용하여 상기 선택된 실험 디자인에 대한 아노바 설계 매트릭스를 생성하는 단계, 상기 아이টে 중 상기 표준화가 선택되면, 입력된 슬라이드 정보를 이용한 단일 슬라이드 표준화, 단일 배치 표준화 및 다중 슬라이드 표준화를 순차적으로 수행하여 표준화 결과를 생성하는 단계, 상기 아이টে 중 상기 추정과 검정이 선택되면, 상기 표준화 결과에 대해, 소정의 추정과 검정 방법을 수행하여 마이크로어레이 실험에서 유의한 유전자를 선택하는 단계, 상기 아이টে 중 상기 군집 분석이 선택되면, 상기 표준화 결과를 이용하여 소정의 군집 분석을 수행하여 생성된 군집 분석 결과를 그래프로 디스플레이하는 단계 및 상기 아이টে 중 상기 분류가 선택되면, 상기 표준화 결과 및 처리군 간의 변동과 처리군 내에서의 변동의 비를 이용하여 후보 유전자를 선정하고, 상기 후보 유전자에 대해 소정의 분류를 수행하여 상기 설계 파라미터에 의한 관심 대상 그룹을 예측하며, 오분류율을 출력하는 단계를 포함한다.

본 발명에 의하면, 하나의 통합시스템에서 마이크로어레이 자료 분석의 전 과정을 처리함으로써, 데이터베이스를 공유하여 마이크로어레이 스캐닝 이미지 자료에 대한 체계적인 통계분석을 수행할 수 있고, 각 실험의 상황에 따른 최적의 통계분석법을 적용함으로써 부적절한 방법의 적용으로 인한 위양성 및 위음성 오류율을 최소화하고 연구결과의 신뢰성을 높일 수 있으며, 편리하고 사용자 친화적인 인터페이스를 제공하여 분석이 용이하다.

대표도 - 도2



(72) 발명자

전명식

서울 성북구 안암동 5-1 고려대학교 정경대학

송석현

서울 성북구 안암동5가 고려대학교 정경대학

이정복

경기 안산시 단원구 고잔1동 516

손인석

서울 성북구 안암동 5-1 고려대학교 정경대학

박미라

대전 중구 용두동 143-5 을지과학기술 의예과

특허청구의 범위

청구항 1

실험 설계, 표준화, 추정과 검정, 군집 분석 및 분류로 구성된 상위 메뉴의 아이টে에 대하여,

상기 아이টে 중 상기 실험 설계가 선택되면, 염료 교환, 기준설계 또는 루프설계 중 어느 하나의 실험 디자인을 선택하는 인터페이스와 상기 선택된 실험 디자인에 사용될 설계 파라미터를 입력하는 인터페이스를 디스플레이 하고, 상기 입력된 설계 파라미터를 이용하여 상기 선택된 실험 디자인에 대한 아노바 설계 매트릭스를 생성하는 단계;

상기 아이টে 중 상기 표준화가 선택되면, 입력된 슬라이드 정보를 이용한 단일 슬라이드 표준화, 단일 배치 표준화 및 다중 슬라이드 표준화를 순차적으로 수행하여 표준화 결과를 생성하는 단계;

상기 아이টে 중 상기 추정과 검정이 선택되면, 상기 표준화 결과에 대해, 소정의 추정과 검정 방법을 수행하여 마이크로어레이 실험에서 유의한 유전자를 선택하는 단계;

상기 아이টে 중 상기 군집 분석이 선택되면, 상기 표준화 결과를 이용하여 소정의 군집 분석을 수행하여 생성된 군집 분석 결과를 그래프로 디스플레이하는 단계; 및

상기 아이টে 중 상기 분류가 선택되면, 상기 표준화 결과 및 처리군 간의 변동과 처리군 내에서의 변동의 비를 이용하여 후보 유전자를 선정하고, 상기 후보 유전자에 대해 소정의 분류를 수행하여 상기 설계 파라미터에 의한 관심 대상 그룹을 예측하며 오분류율을 출력하는 단계를 포함하는 마이크로어레이 통합 분석 방법.

청구항 2

제 1 항에 있어서,

상기 아노바 설계 매트릭스를 생성하는 단계는

상기 설계 파라미터는 상기 선택된 실험 디자인에서 사용될 어레이 개수, 어레이 간 반복 여부, 어레이 내 반복 여부 및 처리 리스트를 포함하는 것을 특징으로 하는 마이크로어레이 통합 분석 방법.

청구항 3

제 1 항에 있어서,

상기 표준화 결과를 생성하는 단계는

상기 단일 슬라이드 표준화 수행시에 평균, 강도, 프린트 팁 또는 스케일 표준화 중 어느 하나의 표준화 방법을 선택하는 인터페이스를 디스플레이하고, 상기 단일 슬라이드를 상기 선택된 방법으로 표준화시키는 단계; 및

상기 염료 교환 실험에 대해 동일한 조건에서 Cy3와 Cy5를 반대로 하여 실험한 결과들을 비교하여 표준화를 수행하는 단계를 포함하는 것을 특징으로 하는 마이크로어레이 통합 분석 방법.

청구항 4

제 1 항에 있어서,

상기 마이크로어레이 실험에서 유의한 유전자를 선택하는 단계는

상기 표준화 결과에 대해, 뉴턴 분석 또는 B-통계량 분석 중 어느 하나의 분석을 수행하여 마이크로어레이 실험에서 유의한 유전자를 추정하는 단계;

상기 표준화 결과에 대해, 에스에이엠(SAM) 분석 또는 이비에엠(EBAM) 중 어느 하나의 분석을 수행하여 반복된 마이크로어레이 실험에서 유의한 유전자를 탐색하고 상기 탐색된 유전자에 따른 잘못 발견할 확률을 연산하는 단계; 및

상기 표준화 결과에 대해, 에이디피(ADP) 분석을 수행하여 마이크로어레이 실험에서 유의한 유전자를 추정하며 반복된 마이크로어레이 실험에서 유의한 유전자의 조정유의확률을 연산하여 상기 조정유의확률을 디스플레이하는 단계를 포함하는 것을 특징으로 하는 마이크로어레이 통합 분석 방법.

청구항 5

제 1 항에 있어서,

상기 군집 분석 결과를 그래프로 디스플레이하는 단계는

유전자 또는 샘플에 대한 계층적 군집분석, k-평균 군집분석, 자기조직도 분석, 주성분 분석 또는 유전자에 대한 진셰이빙 분석 중 어느 하나의 분석 방법을 선택하는 인터페이스를 제공하는 단계;

상기 계층적 군집분석이 선택된 경우, 상기 군집 분석 결과로서 덴드로 그래를 디스플레이하는 단계;

상기 k-평균 군집분석 또는 자기조직도 분석 중 어느 하나의 분석이 선택된 경우, 상기 군집 분석 결과로서 각 클러스터마다 프로파일의 평균 그래프를 디스플레이하는 단계;

상기 주성분 분석이 선택된 경우, 상기 군집 분석 결과로서 제1 주성분과 제2 주성분을 축으로 하여 각 유전자들을 플롯한 그래프를 디스플레이하는 단계; 및

상기 진셰이빙 분석이 선택된 경우, 상기 군집 분석 결과로서 각 클러스터에서의 실험과 유전자 발현정도를 나타내는 그래프, 실제 주어진 데이터와 순열 데이터를 나타내는 그래프 및 클러스터의 크기에 따른 겹 통계량을 나타내는 그래프를 디스플레이하는 단계를 포함하는 것을 특징으로 하는 마이크로어레이 통합 분석 방법.

청구항 6

제 1 항에 있어서,

상기 설계 파라미터에 의한 관심 대상 그룹을 예측하며 오분류율을 출력하는 단계는

선형판별분석, 이차판별분석, 선형판별분석의 크로스확인 또는 이차판별분석의 크로스 확인 중 어느 하나의 분석 방법을 선택하는 인터페이스를 제공하고, 선택된 분석 방법을 수행하는 단계를 포함하는 것을 특징으로 하는 마이크로어레이 통합 분석 방법.

청구항 7

제 1 항에 있어서,

상기 설계 파라미터에 의한 관심 대상 그룹을 예측하며 오분류율을 출력하는 단계는

상기 표준화 결과에 대해, 의사결정규칙을 도표화하여 예측의 대상이 되는 집단을 복수의 소집단으로 분류하고 의사결정나무를 생성하는 단계를 포함하는 것을 특징으로 하는 마이크로어레이 통합 분석 방법.

명세서

발명의 상세한 설명

발명의 목적

발명이 속하는 기술 및 그 분야의 종래기술

<8> 본 발명은 DNA 칩에 관한 것으로, 특히, 마이크로어레이 통합 분석 방법에 관한 것이다.

<9> 마이크로어레이 실험도 여타의 생물학 실험과 마찬가지로 실험 설계가 우선되어야 하는데, 실험 설계는 실험 목적에 의존적일 뿐 아니라, 분석 방법을 염두에 두고 결정되어야 한다. 실험의 목적은 두 군간 유의적 차이를 보이는 유전자를 검색하고, 이를 기초로 모집단 선별용 바이오마커를 개발하는 연구일 수도 있고, 현재까지의 임상 기술로 판별이 어려운 Small Round Blue Cell Tumor의 네 가지 서브 타입을 판별하는 것이 목적일 수도 있다. 또한, 유전자발현의 차이를 마이크로어레이를 이용하여 광범위하게 분석하는 연구 또는 진단에 사용될 수 있다. 혹은 초파리의 전사체에서 연령, 성별, 유전자형의 세 요인의 교호작용 크기를 파악하는 것이나, 어느 특정 종의 모집단 간, 그리고 모집단 내 유전자 발현의 변이 양상을 파악하는 것이 실험의 목적일 수 있으며, 두개의 항암제를 동시에 투여할 때 반응을 보이는 유전자 검색이 연구의 목적일 수도 있다. 연구의 목적에 따라 실험 설계가 달라지고, 분석 방법이 달라지게 되는 것은 물론이다.

<10> 마이크로어레이는 실험의 속성상 여러 단계를 거치게 되고, 매단계마다 실험 오차가 개입할 여지가 있으므로 일

반적으로 오차가 많은 실험이라고 할 수 있다. 여러 단계에서 누적된 오차는 통계 분석을 어렵게 하는 한 요인이 되고 있지만, 그 보다 더 통계학자들을 '당혹'스럽게 하는 것은 수천 개의 유전자중에서 불과 몇 개를 선별하여야 하는데 분석대상의 관찰치를 구성하는 마이크로어레이는 불과 수십 개에 불과한 이른 바 '큰 p, 작은 n'의 문제를 구성하고 있다는 점이다. 여기서 p는 설명변수의 개수이고, n은 관찰치의 개수를 나타내고 있으며, 전통적으로 통계학자들은 '큰 n, 작은 p' 자료를 주로 다루어 왔으나 마이크로어레이 자료가 대표적인 '큰 p, 작은 n' 자료를 구성하고 있다. 따라서, 마이크로어레이 실험 자료의 통계적 분석은 통계학자들에게도 새로운 도전이 되고 있다.

- <11> 종래의 마이크로어레이 실험 분석 방법은 각 분석 방법을 수행할때마다 별도의 프로그램을 필요로 하고, 각 프로그램마다 데이터 호환이 어렵고, 분석 데이터들을 체계적으로 관리할 수 없으며, 부적절한 방법의 적용으로 인한 위양성 및 위음성 오류율이 증가되는 것을 최소화할 수 없는 문제점이 있다.

발명이 이루고자 하는 기술적 과제

- <12> 따라서, 본 발명이 이루고자 하는 기술적 과제는 데이터베이스를 공유하여 마이크로어레이 스캐닝 이미지 자료에 대한 체계적인 통계분석을 수행할 수 있고, 부적절한 방법의 적용으로 인한 위양성 및 위음성 오류율을 최소화하고 연구결과의 신뢰성을 높일 수 있으며, 하나의 통합시스템에서 마이크로어레이 자료 분석의 전 과정을 처리할 수 있어 분석이 용이한 마이크로어레이 통합 분석 방법을 제공하는데 있다.

발명의 구성 및 작용

- <13> 상기의 기술적 과제를 이루기 위하여, 본 발명은 실험 설계, 표준화, 추정과 검정, 군집 분석 및 분류로 구성된 상위 메뉴의 아이템에 대하여, 상기 아이템 중 상기 실험 설계가 선택되면, 염료 교환, 기준설계 또는 루프설계 중 어느 하나의 실험 디자인을 선택하는 인터페이스와 상기 선택된 실험 디자인에 사용될 설계 파라미터를 입력하는 인터페이스를 디스플레이하고, 상기 입력된 설계 파라미터를 이용하여 상기 선택된 실험 디자인에 대한 아노바 설계 매트릭스를 생성하는 단계, 상기 아이템 중 상기 표준화가 선택되면, 입력된 슬라이드 정보를 이용한 단일 슬라이드 표준화, 단일 배치 표준화 및 다중 슬라이드 표준화를 순차적으로 수행하여 표준화 결과를 생성하는 단계, 상기 아이템 중 상기 추정과 검정이 선택되면, 상기 표준화 결과에 대해, 소정의 추정과 검정 방법을 수행하여 마이크로어레이 실험에서 유의한 유전자를 선택하는 단계, 상기 아이템 중 상기 군집 분석이 선택되면, 상기 표준화 결과를 이용하여 소정의 군집 분석을 수행하여 생성된 군집 분석 결과를 그래프로 디스플레이하는 단계 및 상기 아이템 중 상기 분류가 선택되면, 상기 표준화 결과 및 처리군 간의 변동과 처리군 내에서의 변동의 비를 이용하여 후보 유전자를 선정하고, 상기 후보 유전자에 대해 소정의 분류를 수행하여 상기 설계 파라미터에 의한 관심 대상 그룹을 예측하며, 오분류율을 출력하는 단계를 포함하는 마이크로어레이 통합 분석 방법을 제공한다.

- <14> 본 발명에 따른 마이크로어레이 통합 분석 방법은 마이크로어레이 스캐닝 이미지에 대해 블럭 인텍싱 및 스팟 인텍싱을 수행하여 생성된 이미지 자료를 통계적으로 분석하는 소프트웨어에 관한 것이다. 본 발명에 따른 마이크로어레이 통합 분석 방법은 크게 다섯 분야로 구분된다. 즉, 실험 설계, 표준화, 추정과 검정, 군집 분석 그리고 분류이다.

- <15> 먼저, 실험 설계는 염료 교환, 기준설계(reference design), 루프설계, 분석 방법을 제공한다.

- <16> 표준화는 단일 슬라이드 표준화, 염료 교환 표준화, 다중 슬라이드 표준화 그리고 단일 슬라이드 배치 방법을 제공한다. 단일 슬라이드 표준화는 평균 표준화, 강도 의존 표준화, 프린트 팁 표준화, 스케일 표준화 방법이 가능하고, 다중 슬라이드 표준화는 다중 스케일 표준화(Multiple scale normalization)와 순위수 정규화(quantile normalization) 방법이 가능하다. 추정과 검정에서는 유의한 유전자를 추정하고 검정하는 방법을 제공한다. 군집 분석은 유의한 유전자를 각 군집별로 분류하는 방법이다. 마지막으로, 분류에서는 고전적 분류 방법(DLDA, DQDA)과 Tree 방법이 가능하다.

- <17> 이하에서는 도면을 참조하여 본 발명의 바람직한 실시 예를 설명하기로 한다. 그러나, 다음에 예시하는 본 발명의 실시 예는 여러 가지 다른 형태로 변형될 수 있으며, 본 발명의 범위가 다음에 상술하는 실시 예에 한정되는 것은 아니다.

- <18> 다음은 실험 설계에 대한 설명이다. 염료 편향은 표준화가 필요한 여러 사유중 하나를 구성한다. 그러나 표준화가 모든 슬라이드에서 모든 스팟에 대하여 염료 편향을 완벽하게 제거하여 주는 것은 아니기 때문에 실험자들은 염료교체 실험을 권장한다. 암 연구에서 흔히 활용되는 설계는 (공통) 준거(reference) 설계, 블럭 설계 (block

design), 그리고 루프설계(loop design)가 있다. 준거(reference)설계는 상술한 공통준거(reference)설계와 같은 말로서 각 어레이마다 같은 내부 준거(reference) 표본을 사용하여 스왑 간 생길 수 있는 변동을 통제한다. 이를 도식으로 표현하면 도 1a와 같다. 간단한 형태의 환설계는 다음 도 1b와 같다. 물고기를 대상으로 종間 변동과 종內 변동의 크기를 비교하기 위하여 환(loop)설계로써 마이크로어레이 실험을 실시할 수 있다.

<19> 도 2는 본 발명에 따른 마이크로어레이 통합 분석 방법의 흐름도이다.

<20> 본 발명에 따른 마이크로어레이 통합 분석 방법은 대용량의 데이터베이스를 처리할 수 있는 컴퓨터 시스템, 서버 및 클라이언트를 포함하는 네트워크 시스템 또는 이와 유사한 처리 능력을 갖는 시스템 상에서 구현되는 것을 가정한다.

<21> 먼저, 실험 설계, 표준화, 추정과 검정, 군집 분석 및 분류로 구성된 상위 메뉴의 아이টে를 화면으로 출력한다(210 과정).

<22> 아이টে를 중 실험 설계가 선택되면(221 과정), 염료 교환(dye-swap), 기준설계(reference design) 또는 루프설계(loop design) 중 어느 하나의 실험 디자인을 선택하는 인터페이스와 선택된 실험 디자인에 사용될 설계 파라미터를 입력하는 인터페이스를 디스플레이하여, 사용자의 파라미터 입력에 따라 설계 파라미터드을 지정시킨다(231 과정). 다음, 입력된 설계 파라미터를 이용하여 선택된 실험 디자인에 대한 아노바(ANOVA) 설계 매트릭스를 생성한다(232 과정).

<23> 한편, 아이টে를 중 표준화가 선택되면(222), 슬라이드 정보를 입력받고, 입력된 슬라이드 정보를 이용한 단일 슬라이드 표준화, 단일 배치 표준화 및 다중 슬라이드 표준화를 순차적으로 수행하여 표준화 결과를 생성한다(242 과정). 생성된 표준화 결과는 파일로 저장하여 추정과 검정, 군집 분석 및 분류의 수행과정에서 독출되도록 한다(243 과정).

<24> 한편, 아이টে를 중 추정과 검정이 선택되면(223 과정), 표준화 결과를 독출한다(251 과정). 다음, 독출된 표준화 결과에 대해, 소정의 추정과 검정 방법을 수행하여 마이크로어레이 실험에서 유의한 유전자를 선택한다(252 과정). 이때, 수행하는 분석 방법에 따라 선택된 유전자에 따른 잘못 발견할 확률을 연산할 수 있다. 이때, 소정의 추정 방법은 뉴턴 분석(Newton's method) 방법, B 통계량(B-statistic) 방법, 조정유의확률 방법(ADP), 에스에엠(Significant Analysis of Microarray, SAM) 그리고 이비에이엠(Empirical Bayes Analysis of Microarray, EBAM) 등의 분석방법을 포함한다. 다음, 추정된 유전자 정보, 잘못 발견할 확률(연산한 경우) 등의 데이터들을 데이터베이스에 저장하여 이하의 다른 분석 과정에서 참조되도록 한다(253 과정).

<25> 한편, 아이টে를 중 군집 분석이 선택되면(224 과정), 표준화 결과를 독출한다(261 과정). 다음, 독출된 표준화 결과를 이용하여 소정의 군집 분석을 수행하여 생성된 군집 분석 결과를 그래프로 디스플레이한다(262 과정). 이때, 소정의 군집 분석은 계층적 군집 분석(hierarchical clustering), K-평균 군집 분석(k-means clustering), 주성분 분석(PCA: Principal Component Analysis), 자기조직도(SOM: Self-Organizing Map), 진셰이빙(geneshaving) 등을 포함한다. 다음, 군집 분석 결과를 데이터베이스에 저장하여 이하의 다른 분석 과정에서 참조되도록 한다(263 과정).

<26> 한편, 아이টে를 중 분류가 선택되면, 표준화 결과를 독출한다(271 과정).

<27> 다음, 독출된 표준화 결과를 이용하여 처리군 간의 변동과 처리군 내에서의 변동의 비를 연산하고, 연산된 변동의 비를 이용하여 후보 유전자를 선정한다(272 과정). 또한, 선정된 후보 유전자에 대해 소정의 분류를 수행하여 상기 설계 파라미터에 의한 관심 대상 그룹을 예측하며, 이에 따른 오분류율을 출력한다(272 과정). 특히, 각각의 처리군에 유의한 유전자에 의해 분류된 그룹 즉, 관심 대상 그룹과 오분류율을 출력한다(272 과정). 이때, 관심 대상 그룹은 실험에 사용된 Test Set 중에서 실험의 결과로 알아내고자 하는 질병이 발현된 그룹으로서, 어떠한 유전자를 가진 그룹이 간암과 같은 특정 질병과 연관성이 있는지 파악하기 위한 정보이다. 이때, 소정의 분류는 고전적 분류 방법(DLDA, DQDA)과 Tree 방법(또는 의사결정 나무(Decision Tree))를 포함한다. 다음, 유의한 유전자가 분류된 그룹 정보 및 오분류율을 데이터베이스에 저장하여 다른 분석 과정에서 참조되도록 한다(273 과정).

<28> 마지막으로, 모든 분석이 종료되었는지 판단한다(280 과정). 이 과정(280 과정)은 사용자에게 모든 과정을 종료시킬 것인지 아니면 다른 분석을 연이어 수행할 것인지 선택을 요청하는 메시지를 출력하고, 사용자로부터 선택 결과를 입력받는 과정을 포함할 수 있다. 이때, 사용자가 모든 분석의 종료를 선택하면 모든 과정을 종료하고, 그렇지 않으면 위의 과정들(210-280 과정)을 반복한다.

<29> 이하에서는 염료 교환(dye-swap), 기준(reference), 루프(loop) 실험자료 에 대한 ANOVA 분석 소프트웨어를 설명한다.

<30> 입력 데이터 형식은 다음의 표 1과 같이 구성한다.

표 1a

열	변수명	비고
col 1	gene id	유전자를 구분할 수 있는 공백 없는 문자열 형태
col 2	R	붉은색 염료(Cy5)로 염색된 처리의 발현도 log(foreground red-background red)
col 3	G	녹색 염료(Cy3)로 염색된 처리의 발현도 log(foreground green-background green)
col 4	flag	에러가 발생한 유전자는 1, 아닌 경우는 0으로 표시

<32> 출력 데이터 형식은 다음의 표 2와 같이 구성된다.

표 1b

열	변수명	데이터 형	비고
col 1	Gene	integer	각 유전자별로 고유의 숫자를 부여
col 2	intensity	double	R, G값을 한 컬럼으로 병합
col 3	Array	integer	1~어레이 개수
col 4	Dye	integer	1,2
col 5	Variety	integer	1~Variety 개수

<34> 도 3에서 왼쪽 메뉴의 실험 설계(Experimental Design)을 선택하고 나타나는 하위메뉴에서 염료 교환(dye-swap), 기준(reference), 루프(loop) 중 하나를 선택한다.

<35> 이때, Experiment Type은 실험 디자인 선택, Array file은 Array 파일 리스트, Number of Arrays은 분석에 이용될 어레이 개수 입력, Array Replication은 어레이 간 반복 여부, Multiple Spotting은 어레이 내 반복 여부, RNA Species List은 처리(Variety) 리스트, Array information은 각 어레이 별 처리 정보 입력, Name of the ANOVA Design Matrix는 저장될 파일 이름, Generate a Design Matrix은 데이터 셋 생성 버튼, Test sample 은 비교 대상 처리 선택, Bootstrap Replications은 붓스트랩 반복 연산 횟수, Significance Level은 유의수준, RUN은 ANOVA 실행 버튼을 나타낸다.

<36> 먼저, 염료 교환(dye-swap), 기준설계(Reference Design), 루프설계(Loop Design) 중에서 실험 디자인을 선택한다.

<37> Array file은 Array 파일 리스트로서, 각 파일은 하나의 어레이 자료(gene id, Red intensity, Green intensity, flag 로 구성)이다, Array file은 리스트를 보여주는 기능일 뿐 특별한 조작이 필요 없다.

<38> 다음, Number of Arrays에서 분석에 이용될 어레이 개수를 입력한다. 실험된 어레이 중 일부분만 사용할 수도 있으므로 분석에 이용할 어레이 개수를 직접 입력한다. 또한, Array Replication은 염료 교환(dye swap) 실험을 두 번 반복한 경우와 같이, 실험에 어레이 간 반복이 있는 경우 체크한다. Multiple Spotting은 칩 제작시 유전자를 동일 횟수(2회 이상) 반복하여 점적한 경우에 체크한다. RNA Species List에는 실험에 이용된 처리를 차례대로 입력한다. 예를 들어, 시간에 따른 발현의 차이를 보는 경우에는 time0, time1, time2, time3 등으로 입력하고, 약품 투여 양에 따른 차이를 보는 경우에는 0mg, 10mg, 20mg, 30mg 등으로 입력한다.

<39> Name of the ANOVA Design Matrix은 설계 매트릭스가 저장될 파일 이름이다.

<40> Generate a Design Matrix는 데이터 셋 즉, 아노바 설계 매트릭스를 생성하기 위한 버튼이다. 실험 정보를 모두 입력하고 이를 선택하면 데이터셋을 생성한다.

<41> Test sample은 비교 대상 처리를 선택하는 버튼이고, Bootstrap Replications는 붓스트랩 반복 연산 횟수이며, Significance Level는 유의수준을 나타낸다.

<42> 실험의 오차를 줄이기 위하여 염료 교환(dye swap) 실험을 3회에 걸쳐 반복 실험하는 경우, 표 1c와 같이, 총 6개의 어레이가 생성된다.

표 1c

<43>

	Red(Cy5)	Green(Cy3)
Array1	Variety2	Variety1
Array2	Variety2	Variety1
Array3	Variety1	Variety2
Array4	Variety2	Variety1
Array5	Variety1	Variety2
Array6	Variety1	Variety2

<44> 다음은 표준화에 대한 설명이다.

<45> 표준화는 마이크로어레이 실험 후 통계분석을 행함에 앞서 계통적인 변동(systematic variation)을 제거하는 작업으로 시스템에서 구현된 표준화 방법은 다음과 같다.

<46> cDNA 마이크로어레이 실험에서 표준화는 Cy5, Cy3의 비를 보정하는데 사용한다. cDNA 마이크로어레이 분석을 할 경우 각각의 $\log_2 R$, $\log_2 G$ 로 하는 것보다 $M = \log_2(R/G)$ 로 변환하여 하는 장점은 절대적 강도는 마이크로어레이의 고유한 스팟-스팟(spot-spot)변동에 의해서 교란(confound)될 수 있기 때문에 여러 슬라이드에서 절대적 강도보다 M 으로 분석하는 것이 더 안정적이기 때문이다. 그래서 표준화도 각 절대적 강도의 로그비를 보정하는데 중점을 두는 것이다.

<47> 표준화 중 단일 슬라이드 표준화(Single Slide normalization)는 한 장의 마이크로어레이 슬라이드(Microarray slide)를 분석하기 위한 방법으로 뒤에서 사용할 단일 배치에서 사용할 방법중 어떤 것이 가장 잘 적용될 수 있을지를 사전에 살펴보는 단계이다. 즉, 한 슬라이드에서 Cy3, Cy5의 비를 맞추는 방법이다.

<48> 입력 파일 형식은 [gene_id] [Cy5] [Cy5 Background] [Cy3] [Cy3 Background]으로서, gene_id은 문자, 숫자로 구성되고, Cy5, Cy5 Background, Cy3, Cy3 Background는 숫자로 구성되며, 테이터는 탭 분리되어야 한다. 이 때, 결과 파일 형식은 [geneid] [$A = \log_2 \sqrt{Cy5/Cy3}$] [$M = \log_2(Cy5/Cy3)$]이다.

<49> 단일 슬라이드 메뉴를 선택하면 나타나는 분석 대상 중 하나를 선택하면 도 4a와 같은 화면이 나타나게 된다. 화면은 Base Info 테이블과 분석을 하기위한 정보가 들어있는 'Statistical Analysis' 테이블로 구성되어 있다.

<50> 분석은 'Statistical Analysis' 테이블의 'Standardization Input File' 항목에 있는 'Execute'를 선택하면 분석이 시작된다. 이하의 분석 방법 수행을 위해, Java Web start를 실행시키도록 소프트웨어를 구성할 수 있다.

<51> 도 4b는 분석을 위한 선택사항을 입력하는 화면이다. 단일 슬라이드는 Mean, Intensity, Print-tip, Scale의 4가지 표준화 방법을 지원한다. 이 방법중 사용하기를 원하는 방법을 하나 선택하여서 밑의 'ok'를 선택하면, 분석이 시작된다. 이 때 분석 방법중 Print-tip이나 scale을 입력하면 'Number of Tip' 항목과 'Number of genes in block' 옵션이 활성화 되도록 소프트웨어를 구성할 수 있다.

<52> Global 표준화는 Cy5의 강도와 Cy3의 강도가 상수배로 되어있다고 가정하는 경우이다. 강도 의존 표준화(Intensity dependent normalization)는 Cy5의 강도와 Cy3의 강도가 A-dependent한 관계가 되어 있다고 가정하는 경우이다. 프린트 팁 표준화(print-tip normalization)는 프린트 팁 그룹(print-tip-group) 효과가 있다고 가정하는 경우이다. 스케일 표준화(scale normalization)는 프린트 팁 표준화한 후 프린트 팁 그룹(print-tip-group)간에 분산을 보정하는 방법이다.

<53> 도 4c는 도 4b의 옵션에 따른 분석 결과를 보여주는 화면이다. 이 결과는 입력한 자료를 선택한 방법으로 표준화 시킨 화면이다. 화면에는 각 유전자의 분포와 노란색 선으로 표시된 lowess 선을 표시된다.

<54> 표준화 중 염료 교환은 실험시에 동일한 조건에서 Cy3와 Cy5만을 반대로 하여 실험한 두 결과를 비교하여 표준화하는 방법이다. 염료 교환 표준화(Dye-Swap normalization)는 염료 교환(Dye swap)된 실험에서 표준화하는 방

법이다.

- <55> 입력 파일로서, 염료교환은 아래와 같은 형식의 파일을 2개 입력 받게 된다. 이 두 파일은 동일한 조건에서 Cy3와 Cy5만을 반대로 하여서 실험한 결과에 대한 파일이다. 입력 파일 형식은 [gene_id] [Cy5] [Cy5 Background] [Cy3] [Cy3 Background]으로서, gene_id는 문자나 숫자 Cy5, Cy5 Background, Cy3, Cy3 Background는 숫자, 데이터는 탭 분리 되어야 한다. 분석 결과 파일은 [geneid] [$A = \log_2 \sqrt{Cy5/Cy3}$] [$M = \log_2 (Cy5/Cy3)$]의 형식이다.
- <56> 도 4d는 단일 염료 교환(single dyeswap) 분석 장면이다. 염료 교환 분석은 화면에 출력되는 슬라이드들 중 분석에 사용할 2개의 슬라이드를 선택하고 실행버튼 즉, 'execute' 를 선택하면 결과 파일이 생성되도록 소프트웨어를 구성할 수 있다.
- <57> 바람직하게는, 왼쪽에서 슬라이드를 선택한후 가운데 있는 '-'> 화살표를 선택하면 선택이 되고 오른쪽에서 슬라이드를 선택한뒤에 '<-'를 선택하면 해제가 되도록 소프트웨어를 구성할 수 있다.
- <58> 표준화 중 단일 배치는 단일 슬라이드에서 분석했던 결과를 기준으로 하여서 여러개의 슬라이드를 동일한 방법으로 표준화 하는 서버 프로그램이다. 여기서 생성된 결과 파일은 자동으로 다중 슬라이드 분석 방법의 목록으로 넘어가게 된다. 단일 슬라이드 배치 표준화는 모든 슬라이드를 동일한 표준화 방법으로 한번에 처리하는 기능이다.
- <59> 입력 파일 형식은 단일 배치 는 염료 교환 표준화와 같은 형식으로서, 입력 파일을 여러개 입력 받게 된다. 분석 결과 파일은 다음과 같은 형식으로 나타난다. 즉, [geneid] [$M_1 = \log_2 (Cy5/Cy3)$] [$M_2 = \log_2 (Cy5/Cy3)$] 이다. 여기서 M 값은 선택한 슬라이드 들의 M값이며 각 슬라이드에 따른 M값이 연속으로 나타나게 된다.
- <60> 도 4e는 단일 배치 분석화면이다. 단일 배치 분석화면은 단일 염료 교환과 비슷한 인터페이스를 가지고 있다. 단, 선택할 수 있는 슬라이드의 개수에 제한이 없다. 자신이 적용하고 싶은 슬라이드를 선택해서 프로그램을 실행하면 된다.
- <61> 단일 배치는 표준화 분석 방법을 지정해야 한다. 이것은 단일 슬라이드에서 사용하던 방법과 동일한 방법이다. 이때, Global 표준화 는 Cy5의 강도와 Cy3의 강도가 상수배로 되어있다고 가정하는 경우이다. 강도 의존 표준화(Intensity dependent normalization)는 Cy5의 강도와 Cy3의 강도가 A-dependent한 관계가 되어 있다고 가정하는 경우이다. 프린트 팁 표준화(print-tip normalization)는 프린트 팁 그룹(print-tip-group) 효과가 있다고 가정하는 경우이다. 화면의 슬라이드 선택 부분 바로 위에서, 프린트 팁을 선택할 경우 도 4e처럼 선택부분 오른쪽에 추가 정보를 입력하는 부분이 나타나게 된다.
- <62> 표준화 중 다중 슬라이드는 뒤의 군집 분석, 추정과 검정, 분류에서 사용될 실질적인 자료를 만드는 방법이다. 앞의 단일 배치 에서 결과로 나온 파일을 이용하여 작동된다. 다중 슬라이드 표준화(Multiple slide normalization)는 반복된 슬라이드에서 분산이나 슬라이드의 변위치를 맞추는 방법이다.
- <63> 입력 파일 형식은 단일 배치 결과로 나온 파일을 입력 받게 된다. 분석 결과 파일 역시 단일 배치 의 출력 형식과 동일하다.
- <64> 도 4f는 다중 슬라이드 분석 화면이다. 화면상에 나와있는 Multiscale, Quantile 중 하나를 선택하고, 'execute' 를 선택하면 실행이 된다.
- <65> 분석옵션은 2가지가 있다. 먼저, 다중 스케일 표준화는 각 슬라이드의 분산을 조정하는 방법이다. 순위수 정규화는 각 슬라이드의 변위치 를 조정하여 표준화하는 방법이다. 'execute' 를 선택하면, 결과파일이 생성된다.
- <66> 다중 슬라이드는 슬라이드의 순서가 분석 결과에 영향을 미치는 방법이다. 따라서 사용자가 자료에 들어있는 슬라이드의 정보를 직접 변경 시킬 수 있다. 이것은 'modify'를 선택하면, 엑셀과 같은 편집 프로그램을 작동시켜 편집하고 저장하도록 할 수 있다.
- <67> 다음은 추정과 검정에 관한 설명이다.
- <68> 추정과 검정을 실시하고자 할 때는 분석 방향을 설정하여야 한다. 이 부분은 크게 두 부분으로 나뉜다. 첫째, 어떤 통계량을 사용하여 발현에 대한 유전자의 순위를 매길 것인가 하는 추정부분, 둘째 해당 통계량이 주어지고 그에 따라 추정을 실시하여 발견한 유전자들에 대한 신뢰수준을 결정하는 검정부분이다.
- <69> B 통계량(Lonnstedt 등의 베이지안 B 통계량)은 다르게 발현되는 각 유전자에 대한 사후 log-odds를 추정하는

통계량이다. B 통계량 역시 S 통계량처럼 t 통계량에 일정한 패널티를 부여한 방식이다. 특히 t-값이 높은 자료에 상당한 패널티를 가하는 방식이다. B 통계량을 통한 log-odd 순위는 bayesian 프로그램으로 얻을 수 있다.

- <70> 에이디피(ADP) 프로그램으로 전통적인 t 통계량, 에스에이엠(SAM) 과 이비에이엠(EBAM) 으로 S 통계량, Bayesian 프로그램으로 B 통계량을 사용하여 가능한 모든 통계량의 순위를 결정할 수 있다. 특히 EBAM에서 S 통계량의 비율을 사용자가 대표적 몇 개를 입력하여 순위를 산출한다.
- <71> 조정유의확률(Adjusted p-value)은 다음과 같다. Family-wise error rate 는 유전자의 어떤 구성이 실제로 다르게 발현되는지 상관없이, 다르게 발현된 유전자들 중에서 적어도 한 유전자가 거짓으로 나타날 확률을 의미한다. Family-wise error rate를 추정하는 방법은 상당히 다양한 버전 이 있으며, 가장 보수적인 방법으로 조정유의확률을 들 수 있다. ADP 프로그램에서 조정유의확률을 얻을 수 있다.
- <72> FDR은 유의적으로 다르게 발현된 것으로 선택된 유전자들 중에서 오차비율의 기대값으로 정의되며 "잘못 발견할 비율"을 의미한다. Tusher 의 SAM 방법은 임의순열 방법에 의해 각 유전자의 기대순위를 연산하고 실제 유전자와 차이를 통해 FDR을 추정한다. FDR 은 SAM 프로그램에서 얻을 수 있다.
- <73> SAM과 거의 유사한 방법으로 EBAM을 통해 FDR을 추정할 수 있다. 다만, 사용되는 S 통계량에 차이가 있으므로 에스에이엠(SAM)과 다른 FDR 결과를 유도한다. 이 방법에서 FDR 추정은 보통 사후확률 0.9에서 정하며 연구자가 사후확률 선택을 반대로 여러 FDR과 기각유전자 수를 고려하여 경험적으로 역선택할 수도 있다. 다양한 S 통계량에 대한 FDR은 EBAM 프로그램에서 얻을 수 있다.
- <74> 추정과 검정 중 뉴턴 분석(Newton method)은 단일 슬라이드 마이크로어레이 실험에서 유의한 유전자를 추정하는 서브 프로그램이다. 입력 파일 형식은 [유전자이름] [A(_)] [M(_)]와 같다.
- <75> 분석 결과 파일은 Log posterior odds가 0보다 큰 유전자 리스트, Log posterior odds가 0보다 작거나 같은 유전자 리스트, 최대값과 최소값을 이용해서 등고선을 그리기 위한 Log posterior odds 값, 첸(Chen) 그래프를 그리기 위한(95%, 99%) 4개의 절편 값을 포함한다.
- <76> 도 5a의 실행 결과 그래프에서, 검정색 등고선은 posterior odds의 변화가 1:1, 10:1, 100:1인 등고선을 나타내고, 하늘색 직선은 첸(Chen) 방법의 95%와 99% 신뢰구간을 나타내며, 빨강색 점은 Log posterior odds가 0보다 큰 유전자를 나타내고, 녹색 점은 Log posterior odds가 0보다 작거나 같은 유전자를 나타낸다.
- <77> 추정과 검정 중 B 통계량(B-statistic)은 반복된 마이크로어레이 실험에서 유의한 유전자를 추정하는 서브 프로그램이다. 입력 파일의 데이터 형식은 다음의 표 2와 같이 탭 분리한다.

표 2

<78>

gene1	1.269	0.277	-0.343	0.659
gene2	0.606	0.838	-0.405	1.241
gene3	0.852	0.627	1.918	-1.077
gene4	-0.065	-0.570	-0.446	0.364
.....				

- <79> 분석 결과는 각 유전자의 M(_)값 평균, 각 유전자의 log posterior odds 값을 포함한다.
- <80> 도 5b의 옵션 입력 화면에서, Data는 *.txt 형식이고, Number of rows는 유전자의 개수이며, Number of replication은 슬라이드의 수, P는 임의의 유전자가 유의하게 발현될 확률이다.
- <81> 도 5c는 도 5b의 옵션에 따른 결과 그래프의 일 예이다.
- <82> 이때, 빨강색은 유의한 유전자 (log posterior odds 값이 0보다 큰 유전자), 초록색은 유의하지 않은 유전자 (log posterior odds 값이 0보다 작은 유전자)를 나타낸다.
- <83> 추정과 검정 중 에이디피(ADP)는 4 종류의 반복된 마이크로어레이 실험에서 유의한 유전자의 조정유의확률을 연산하는 서브 프로그램이다. 입력 파일 형식은 B-통계량의 경우와 같이 탭 혹은 공백문자로 분리한다.
- <84> 분석 결과는 10종류의 조정유의확률과 3 종류의 유의확률을 포함하고, 에이디피(ADP) 실행과 관련된 각종 정보를 제공한다.

- <85> 5d의 옵션 입력 화면에서, Open data는 data는 텍스트 형식이고, Response type은 4종류의 실험유형(1그룹, 2그룹, 3그룹 이상, 반응형식)을 나타내고, Number of permutations는 임의순열반복횟수, Number of bootstraps는 부트스트랩반복횟수, Seed Numbers는 임의난수를 발생하는 초기값, Number of groups는 각 그룹의 반복횟수 입력, Quantitative response file는 반응형식이 있는 실험에 대한 반응형식 지정파일, p-value는 10종류의 조정 유의확률, Run Result Capture는 실행과정 표시, Raw p-value는 임의순열과 부트스트랩 유의확률, Tmp Outfile Capture는 Tmp file에 대한 내용 출력을 나타낸다.
- <86> 도 5e는 도 5d의 옵션에 따른 실행결과 그래프의 일 예이다. 이때, 파랑색은 raw p-value를, 빨강색은 adjusted p-value를 나타낸다.
- <87> 도 5e에서, 특정 그림에 있어 유의한 유전자를 마우스로 탐색하면 도 5f와 같은 화면을 볼 수 있도록 소프트웨어를 구성할 수 있다.
- <88> 추정과 검정 중 에스에이엠(SAM)은 4 종류의 반복된 마이크로어레이 실험에서 유의한 유전자를 탐색하고 FDR을 연산하는 서브 프로그램이다. 입력 파일 형식은 B-통계량의 경우와 같이 탭 혹은 공백문자로 분리한다.
- <89> 분석 결과는 Delta 값에 따른 유의한 유전자 수와 3종류의 선택기준에 의한 FDR을 제공하고, 에스에이엠(SAM) 실행과 관련된 각종 정보와 초기 Delta 값을 제공하며, 초기 Delta 값에 대한 에스에이엠(SAM) 도면의 정보, 다른 Delta 값에 대한 조정된 에스에이엠(SAM) 도면의 정보, 최종 Delta 값에 대한 유의한 유전자 리스트 및 Q-value 제공한다.
- <90> 도 5g의 옵션입력 화면에서, Import data는 data는 텍스트 형식이고, Response type은 4종류의 실험유형(1그룹, 2그룹, 3그룹 이상, 반응형식), Number of permutations는 임의순열 반복횟수, Seed Numbers는 임의난수를 발생하는 초기값, Number of groups는 각 그룹의 반복횟수 입력, Quantitative response file은 반응형식이 있는 실험에 대한 반응형식 지정파일, Gene Observed score는 Gene 과 Observed score 와의 도면, Gene Expected score는 Gene 과 Expected score 와의 도면, Gene plot은 에스에이엠(SAM) 통계량의 분모(표준 오차)와 Expected score 와의 도면, 에스에이엠(SAM) plot은 Observed score 와 Expected score 와의 도면을 나타낸다. FDR plot은 Mean FDR, Median FDR, 90th percentile FDR plot 제공하고, FDR table list는 FDR 결과 파일을 제시한다.
- <91> 도 5h는 도 5g의 옵션에 따라 연산된 FDR을 보여주는 화면이다. 도 5h의 FDR을 보고 Delta를 결정할 수 있다.
- <92> 도 5i는 도 5g의 옵션에 따른 SAM plot의 일 예를 도시한 것이다. 이때, Input Delta value를 통해 Delta 값을 선택할 수 있다. 도 5j는 도 5h 또는 도 5i에 따라 입력된 Delta에 따른 SAM plot의 일 예를 도시한 것이다. 이때, 빨강부분은 Positive 하게 유의한 유전자, 파랑부분은 Negative 하게 유의한 유전자를 나타낸다.
- <93> 추정과 검정 중 이비에이엠(EBAM)은 4 종류의 반복된 마이크로어레이 실험에서 Empirical Bayesian 방법을 통한 유의한 유전자 탐색하고 FDR을 연산하는 서브 프로그램이다. 입력 파일 형식은 B-통계량의 경우와 같이 탭 혹은 공백문자로 분리한다.
- <94> 분석 결과는 사후확률에 따른 유의한 유전자 수와 3종류의 선택기준에 의한 FDR을 제공하고, 이비에이엠(EBAM) 실행과 관련된 각종 정보와 초기 사후확률을 제공하며, 초기 사후확률에 대한 이비에이엠(EBAM) 도면의 정보, 다른 사후확률에 대한 이비에이엠(EBAM) 도면의 정보, 로짓회귀분석에 의한 사후확률과 spline basis 값을 제공하고, 최종 사후확률에 대한 유의한 유전자 리스트 및 Q-value를 제공한다.
- <95> 도 5k의 옵션입력 화면에서, Import data에서 data는 텍스트 형식이고, Response type은 4종류의 실험유형(1그룹, 2그룹, 3그룹 이상, 반응형식), Spline df는 Spline의 자유도를 입력(보통 5), Fudge Factor Percentile은 표준오차를 보정해줄 표준오차의 위치값 (0~1의 값), Number of permutations는 임의순열 반복횟수, Seed는 임의난수를 발생하는 초기값, Number of groups는 각 그룹의 반복횟수 입력, Quantitative response file는 반응형식이 있는 실험에 대한 반응형식 지정파일, Standard Z score는 Gene 과 보정되지 않은 Z score 도면, Expression score는 Gene 과 보정된 Z score 도면, Numerator:Denominator는 이비에이엠(EBAM) 통계량의 분모(표준오차)와 분자 도면, Natural spline LOGIT prob..는 LOGIT 결과 도면, 이비에이엠(EBAM) plot은 Expected score 와 사후확률간의 도면이다. FDR plot은 Mean FDR, Median FDR, 90th percentile FDR plot 제공하고, FDR table list는 FDR 결과 파일을 제시한다.
- <96> 도 5l은 도 5k의 옵션에 따라 연산된 FDR의 출력화면이다. 이때, 출력된 FDR을 보고 사후확률을 결정할 수 있다.

- <97> 도 5m은 도 5k의 옵션에 따른 EBAM) plot의 일 예이다. 이때, Posterior probability setting을 통해 사후확률 값을 선택할 수 있다. 이때, 빨간색선은 EBAM plot, 파란색선은 조정되지 않은 EBAM plot, 초록부분은 유전자의 expression score를 나타낸다.
- <98> 도 5l 또는 도 5m에 따라 사후확률이 입력되면, 입력된 사후확률에 따른 EBAM plot이 출력되도록 소프트웨어를 구성할 수 있다.
- <99> 도 5n은 도 5l 또는 도 5m에 따라 입력된 사후확률에 따른 유의한 유전자 리스트의 일 예이다.
- <100> 다음은 군집 분석에 관한 설명이다.
- <101> 계층적 군집분석(Hierachical clustering)에는 병합적 방법과 분할적 방법의 두가지가 있다. 병합적 방법은 모든 개체를 각각의 클러스터로 본 후에 유사한 클러스터들을 모아 나가는 것이고 분할적 방법은 모든 개체를 묶어서 하나의 클러스터로 본후에 그룹을 쪼개나가는 것이다. 이때 개체의 유사성이 트리 구조의 도면으로 표현된다.
- <102> K-평균 군집분석 (k-means clustering)은 개체들을 미리 정해진 수(k)의 클러스터에 할당하는 것으로 클러스터의 초기값을 규정하여 결정된 초기클러스터에 각 개체들을 할당한 후 정해진 기준에 따라 최적화될때까지 재할당하는 과정을 거쳐 최종 클러스터가 정해지게 된다. 클러스터의 중심을 기준으로 개체를 할당하게 되는데 중심으로는 평균(k-means)이나 중앙값(k-medoids)을 쓸 수 있다.
- <103> 자기조직도 (SOM :Self Organizing Map)는 신경망모형의 한 종류로서 k-평균 군집분석과 비슷하지만 결과로 나오는 유전자들의 그룹에서 위치상 이웃한 그룹들은 떨어진 그룹보다 유사한 패턴을 보여주게 된다. 분석을 위해서는 사용자가 노드의 형태와 그리드(grid)의 행렬수를 미리 정해야 한다. 바람직하게는, 직사각형과 6각형모양의 노드를 정할 수 있다.
- <104> 주성분 분석(PCA:Principal Component Analysis)은 구조가 복잡하고 서로 상관되어 있는 다차원 자료를 변환시켜 주성분이라는 서로 독립적인 인공의 변수들을 만들어냄으로써 몇 개의 의미있는 주성분으로 많은 부분을 설명할 수 있도록 하는 것이다. 이는 자료의 단순화나 요약, 특이값이나 클러스터의 탐색에 적절하다. 주성분은 상관행렬이나 공분산 행렬로부터 연산할 수 있다.
- <105> 진셰이빙(Geneshaving) 방법은 제 1주성분으로 "Super gene"을 연산한 다음 이와 가장 관련이 없는 유전자들(하위 10%)을 잘라내는 과정을 마지막 유전자 하나만이 남을 때까지 반복한다. 이 절차는 연쇄적인 유전자 블록을 생성하게 되는데 이때 각 클러스터는 클러스터간의 분산은 최대화하고 클러스터내의 분산은 최소화하도록 구성된다. 이 방법을 이용할 때, 유전자들은 단 하나의 클러스터로만 할당되지 않고 하나 이상의 클러스터로 할당될 수 있다.
- <106> 계층적 군집분석(hierachical clustering) 메뉴를 선택하면, 유전자와 샘플에 대한 계층적 군집분석을 실행하고 결과로서 덴드로그램(트리)을 제공한다.
- <107> 분석에 사용될 데이터는 탭으로 분리된 텍스트 파일(*.txt)형식이어야한다. 아래의 예제와 같이 각 유전자들은 행방향으로, 실험(샘플,어레이)는 열방향으로 입력된다. 첫 열에는 해당 유전자의 이름이 입력되고 첫 행에는 실험(샘플, 어레이)의 이름이 들어간다.
- <108> 도 6a에서, Import data에는 업로드한 데이터가 지정된다.
- <109> Linkage Method에서는 군집간의 거리를 정의하는 척도를 선택한다. 이때, 척도로서, Average linkage는 각 군집내에서 하나씩의 개체를 택해 연결가능한 모든 가능한 경우의 거리를 평균한 것을 두 군집간 거리로 정의한다. Complete linkage는 각 군집에 속하는 두 개체 사이의 거리들 중에서 최장 거리를 두 군집간의 거리로 정의한다. Single linkage는 각 군집에 속하는 두 개체 사이의 거리들중에서 최단거리를 두 군집간의 거리로 정의한다.
- <110> Similarity measure는 두 개체간 상사성의 척도를 정한다. correlation(uncentered)은 피어슨의 상관계수를 사용한다. correlation(centered)은 두 개체들의 평균을 0으로 한 피어슨 상관계수를 사용한다. absolute correlation(uncentered)은 상관(비중심화) 척도의 절대값을 사용한다. absolute correlation(centered)은 상관(중심화) 척도의 절대값을 사용한다. Euclidean distance는 유클리드 거리를 이용한다.
- <111> 도 6b의 덴드로그램에서, 화면 왼쪽으로는 유전자클러스터링, 화면 아래쪽으로는 실험(샘플, 어레이)클러스터링

결과가 보인다. 가장 오른쪽의 색상표를 기준으로 데이터값을 볼 수 있다.

- <112> 확대하여 보고 싶은 부분의 트리를 선택하면 도 6c와 같은 새 창이 뜨면서 확대 도면을 보여준다. 유전자 이름과 실험이름이 나타나 있다. 이 화면에서, 그림과 글씨 크기를 조절할 수 있도록 소프트웨어를 구성할 수 있다.
- <113> 도 6c의 화면에서, 찾고 싶은 유전자 이름을 입력하면, 새 창에 그 유전자만 확대된 도면이 표시되도록 소프트웨어를 구성할 수 있다.
- <114> k-평균 군집분석(k-means clustering)에서는 유전자 또는 샘플에 대한 k-평균 군집분석(k-means clustering)을 실행하고 주어진 k수에 따른 유전자 또는 샘플의 군집결과를 그래프와 함께 제공한다.
- <115> 분석에 사용될 데이터는 탭으로 분리된 텍스트 파일(*.txt)형식이어야 한다. 아래의 예제와 같이 각 유전자들은 행방향으로, 실험(샘플, 어레이)는 열방향으로 입력된다. 첫 열에는 해당 유전자의 이름이 입력되고 첫 행에는 실험(샘플, 어레이)의 이름이 들어간다.
- <116> 도 6d의 화면에서, Import data 부분에 분석할 데이터가 지정된다. Target에서 유전자클러스터링을 할지 실험에 대해 클러스터링할지 선택한다. 유전자클러스터링을 위해서는 Gene 을, 실험에 대해 클러스터링하려면 Experiments 선택한다. Number of clusters에 원하는 클러스터의 수를 지정한다. centroid definition에서는 평균정의 방법을 선택한다. K-means는 클러스터의 중심점을 평균으로 정한다. K-medoids는 클러스터의 중심점을 중간값으로 정한다. Number of iteration에는 분석을 실행하기 위한 최대 반복수를 지정한다.
- <117> 도 6e는 도 6d에 따른 결과 그래프의 일 예이다. 각 클러스터마다 프로파일의 평균을 그린 그래프이다.
- <118> 한 클러스터를 선택하면 도 6f와 같은 화면이 나타나 자세한 정보를 알 수 있다. 왼쪽은 이 클러스터에 속하는 유전자(실험) 이름이 써있고 오른쪽은 그들의 프로파일들을 그린 그래프가 나타난다. 왼쪽 유전자(실험) 이름 중에서 하나를 선택하면 해당 프로파일이 오른쪽에 나타난다. 이때, 여러 개의 유전자를 한꺼번에 지정할 수 있도록 소프트웨어를 구성할 수 있다.
- <119> 자기조직도(Self Organizing Map:SOM)에서는 유전자 또는 샘플에 대한 자기조직도(Self Organizing Map:SOM)분석을 실행하고 주어진 노드의 수 및 형태에 따른 유전자 또는 샘플의 군집결과를 그래프와 함께 제공한다.
- <120> 분석에 사용될 데이터는 탭으로 분리된 텍스트 파일(*.txt)형식이어야한다. 아래의 예제와 같이 각 유전자들은 행방향으로, 실험(샘플, 어레이)는 열방향으로 입력된다. 첫 열에는 해당 유전자의 이름이 입력되고 첫 행에는 실험(샘플, 어레이)의 이름이 들어간다.
- <121> 도 6g의 화면에서, Import data 부분에는 업로드한 데이터가 지정된다. Target에서 유전자클러스터링을 할지, 실험에 대해 클러스터링할지 선택한다. 유전자클러스터링을 위해서는 Gene 을, 실험에 대해 클러스터링하려면 Experiments 선택한다. Number of rows에 원하는 행의 수를 지정한다. Number of columns에 원하는 열의 수를 지정한다. 도면은 3*3 SOM을 가정하여 행과 열의 수를 모두 3으로 지정한 것이다. Number of iteration에는 분석을 실행하기 위한 최대 반복수를 지정한다. Initial learning rate에 0과 1사이의 값인 최초의 학습률을 지정한다. Radius에 Map에서 이웃의 크기를 지정한다. 이웃의 크기는 1보다 커야 한다. Neighborhood function에는 커널함수의 형태를 지정한다. Bubble 또는 Gauss를 지정한다. Topology에는 격자의 형태를 정한다. rectangular 또는 hexagonal 형태의 격자 모양을 지정한다.
- <122> 도 6h는 도 6g에 따른 결과 그래프의 일 예이다. 각 클러스터마다 프로파일의 평균을 그린 그래프이다.
- <123> 도 6h에서, 한 클러스터를 선택하면 도 6i와 같은 화면이 나타나 자세한 정보를 알 수 있다. 왼쪽은 이 클러스터에 속하는 유전자(실험) 이름이 써있고 오른쪽은 그들의 프로파일들을 그린 그래프가 나타난다. 왼쪽 유전자(실험) 이름 중에서 하나를 선택하면 해당 프로파일이 오른쪽에 나타난다. 이때, 여러 개의 유전자를 한꺼번에 지정할 수 있도록 소프트웨어를 구성할 수 있다.
- <124> 주성분분석(Principal Component Analysis:PCA)에서는 유전자 또는 샘플에 대한 주성분분석을 실행하고 결과를 그래프와 함께 제공한다.
- <125> 분석에 사용될 데이터는 탭으로 분리된 텍스트 파일(*.txt)형식이어야 한다. 각 유전자들은 행방향으로, 실험(샘플, 어레이)는 열방향으로 입력된다. 첫 열에는 해당 유전자의 이름이 입력되고 첫 행에는 실험(샘플, 어레이)의 이름이 들어간다.
- <126> 도 6j의 화면에서, [Import data]에는 업로드된 데이터 파일이 표시된다. [PC based on]옆의 선택바에서 PCA분

석을 할 때 어떤 데이터값에 근거하여 고유값과 고유벡터를 구할 것인지를 지정한다. 사용행렬에는 R[Correlation]행렬과 V[Covariance]행렬 그리고 S[SSCP] 행렬이 있다.

- <127> 도 6k는 도 6j에 따른 PCA 그래프의 일 예이다. 이 도면은 제1 주성분과 제2 주성분을 축으로 하여 각 유전자들을 플롯한 것이다. 창 아래쪽에는 제1 주성분과 제2 주성분 각각의 고유치와 비율이 나타난다. 각각의 점들에 마우스 포인터를 갖다 대면 해당하는 유전자의 이름이 뜨고 오른쪽에 해당하는 1축과 2축의 좌표값이 나타난다. 특정 유전자들을 보고 싶으면 마우스로 영역을 드래그 할 수 있다. 도 6l과 같이 해당하는 유전자들의 리스트와 프로파일이 새창으로 디스플레이되도록 소프트웨어를 구성할 수 있다.
- <128> 이 중 어떤 한 유전자의 프로파일을 보고 싶으면 '유전자 리스트' 창에서 원하는 유전자를 선택하면 'Gene plot' 창과 'Gene Profile' 창에 빨간색으로 해당 유전자가 표시된다. 이때, 여러개의 유전자를 선택할 수도 있다.
- <129> 도 6m은 View Experiment plot을 선택했을 때 나타나는 실험에 대한 화면으로 유전자에 관한 도 6l과 마찬가지로 해석한다.
- <130> 진셰이빙(Geneshaving)에서는 유전자에 대한 진셰이빙 분석을 실행하고 결과를 그래프와 함께 제공한다.
- <131> 분석에 사용될 데이터는 탭으로 분리된 텍스트 파일(*.txt)형식이어야 한다. 각 유전자들은 행방향으로, 실험(샘플, 어레이)는 열방향으로 입력된다. 첫 열에는 해당 유전자의 이름이 입력되고 첫 행에는 실험(샘플, 어레이)의 이름이 들어간다.
- <132> 도 6n의 화면에서, [Data file]에는 업로드된 데이터파일이 표시된다. [Number of clusters]에는 원하는 클러스터의 개수를 지정한다. [Number of permutation]에는 값을 추정하기 위해 필요한 순열의 수를 선택한다. 도 6n은 클러스터의 수를 3으로, 순열의 수를 5로 지정한 경우이다.
- <133> 도 6o는 도 6n에 따른 결과 그래프 화면의 일 예이다. 각 클러스터당 3개의 그래프가 제공된다.
- <134> 첫 번째 열의 그래프는 첫 번째 클러스터에서의 실험과 유전자발현정도를 보여준다. 가로축에 실험이, 세로축에는 해당 유전자가 있다. 실험의 순서는 각 클러스터별로 유전자의 평균 발현수치순으로 정렬되어 있다. 분산비율($\%variance=VB/VT*100$)과 그룹간-그룹내분산비(VB/VW)와 그룹간 분산(VB) 및 총분산(VT)의 수치를 함께 제공한다. 분산비율이 높을 수록 클러스터가 더 뭉쳐져 있다고 해석할 수 있다. 바람직하게는, 해당 그래프를 더블 선택하면 새로운 창으로 확대된 그래프를 제공할 수 있다.
- <135> 두 번째 그래프는 각 클러스터에 해당되는 유전자의 수, 즉 클러스터의 크기(Cluster Size)를 결정하기 위한 방법으로 실제 주어진 데이터(청색그래프)와 순열 데이터(적색그래프)에서의 분산비율 ($VB/VT*100$)을 보여준다. 순열 데이터는 유전자와 실험이 아무 관계가 없을 때 우연히 나오게 되는 분산비율을 연산하기 위한 것이다. 이 값과 실제 주어진 데이터에서의 분산비율의 차이가 가장 크게 되는 경우를 그 클러스터의 크기로 정하게 된다. 프로그램에서는 이 값에 따라 자동으로 연산된 크기에 의해 각 클러스터의 유전자수가 정해진다. 이 예에서는 1500개의 유전자중 첫 클러스터에 18개, 두 번째 클러스터에 2개, 세 번째 클러스터에 6개의 유전자가 할당되었다. 역시 해당 그래프를 선택하면 별도의 창으로 볼 수 있다. 이 예는 전체 유전자수가 1500개인 데이터로서 가로축은 0부터 1500까지의 수를 의미한다.
- <136> 세 번째 그래프는 두 번째 그래프와 같은 정보를 주는 것으로 클러스터의 크기(Cluster Size)에 따른 값 통계량(실제데이터와 순열 데이터의 분산비율의 차)을 보여준다. 따라서 이 값이 가장 큰 지점을 각 클러스터의 크기로 정하게 된다.
- <137> 다음은 분류에 관한 설명이다.
- <138> 마이크로어레이 자료의 분석에서 분류분석(또는 판별분석)의 목적은 이미 정해진 각 처리(treatment, 예를 들어 정상 조직 및 암종 조직에서 얻은 표본)를 가장 잘 분별해주는 유의한 유전자를 찾아내는 것이다.
- <139> 전통적 분류기법(Classical Discriminant Methods)은 다음과 같다. 가장 대표적인 전통적 분류방법으로 FLDA(Fisher's Linear Discriminant Analysis)는 다차원에 펼쳐진 자료들에 대해 처리간의 차이를 가장 극명하게 나타낼 수 있도록 일차원(즉, 선형결합)으로 축소하는 방법이다. 만약 각 처리마다 분산이 다를 경우 즉, 각 처리군에서 자료의 퍼진 정도가 상이하게 되면 사용될 수 있는 좋은 방법이 DLDA(Diagonal Linear Discriminant Analysis)가 된다. 마지막으로 자료의 분산-공분산이 이차형식(quadratic form)의 형태를 보이게 되면, 이를 DQDA(Diagonal Quadratic Discriminant Analysis)라고 한다.

- <140> 의사결정나무(Decision Tree)방법은 처리를 가장 잘 설명할 수 있는 후보 유전자 하나를 찾아 그 기준에 따라 자료를 나누어 주고(가지를 뺌), 순차적으로 나누어진 각 자료에서 다시 가장 유의한 후보 유전자를 선택하여 가지를 뺌어가는 방법으로, 일정한 규칙에 따라 계속 가지를 뺌어가는 과정을 취하여, 전체 자료를 분류해 내는 유용한 방법중 하나이다. 나무 도면은 계층적 클러스터링의 도면과 매우 유사한 결과를 제공하고 있어 사용자에게 직관적으로 결과를 이해 할 수 있게 하는 장점을 가지고 있으며 마이크로어레이 자료뿐만 아니라 다양한 분야에서 사용되고 있다.
- <141> 변수선택 방법(Variable Selection)은 다음과 같다. 유의하다고 예상되는 후보 유전자를 미리 선택해야 하는 과정이 마이크로어레이 자료의 분류분석에는 필수적이다. 본 발명에 적용되는 변수선택 방법은 처리군 간의 변동(Within SS)과 처리군내에서의 변동(Between SS)의 비(ratio)를 이용한 방법(BSS/WSS)이다. 즉 처리군 간에는 수치값 차이가 크고, 동일 처리군에서는 그 수치값이 유사하여 변동이 작은 후보 유전자를 선택하는 방법이다.
- <142> 마이크로어레이 자료에 대한 분류분석은 표준화된 자료를 이용하여 다음과 같은 과정을 거치게 된다. 첫번째, 결측치 보정이다. 어떤 분석을 실시하느냐에 따라 결측치를 사전에 정리해야 한다. 의사결정나무의 경우는 결측치를 보정하지 않아도 분석이 가능하나, 나머지 방법은 결측치가 보정이 되지 않은 경우 그 유전자는 분석에서 제외된다. 두번째, 변수선택법을 이용한 후보 유전자 선택이다. BSS/WSS에 의하여 유의할 것으로 예상되는 후보 유전자를 선정한다. 이때 후보 유전자의 수는 전체 관찰치의 수보다 작게 한다. 세번째, 크로스 확인(Cross Validation:CV)이다. 분류분석에서는 얻어진 결과가 얼마나 정확한 예측을 하는지 확인할 평가 도구가 필요한데, 가장 직관적인 것이 오분류율(misclassification rate)이다. 오분류율은 연산된 판별결과에 따라 잘못 분류된 개체의 비율을 의미하는데, 통상적으로 전체 자료를 n등분하여, n-1개로 분류분석 모형을 만들고, 나머지 하나를 이용하여 그 분류모형의 오분류율을 연산한다. 이 때 모형을 구성하기 위해서 필요한 자료를 training set이라고 하고, 평가를 위한 자료를 test set이라 한다. 통상적으로 n-fold cross validation이라 함은 위의 오분류율을 n회 반복하여 얻는 과정을 말한다. 따라서 분류분석의 평가를 위해 cross validation을 수행하여야 하고, 이를 위해 적절한 n수를 정해야 한다. 마지막으로, 분석방법 선택이다. 각각의 방법을 통하여 분석을 실시한다.
- <143> BSS/WSS gene selection method는 처리군 간의 변동(Within SS)과 처리군내에서의 변동(Between SS)의 비(ratio)를 이용하여 유의할 것으로 예상되는 후보 유전자를 선정하는 방법이다.
- <144> 입력 파일 형식은 다음의 표 3a과 같이 구성할 수 있다.

표 3a

<145>

[유전자 이름]	[class]	[class]	[class]	[class]	[class]	[class]	[class]	[class]	[class]
one3	1	1	1	1	2	2	3	3	
GENE0000		-0.549	-0.16135		-0.1183	-0.07699		-0.41637	-0.53994
-1.88314		-1.26389							
GENE0001		-0.65766		-0.16635		-0.19691		-0.12792	-0.32658
-0.38762		-1.55486		-0.75601					
GENE0002		-0.75556		-0.24525		-0.08846		-0.13201	-0.50722
-0.44909		-1.64757		-0.70395					
GENE0003		-0.57216		-0.2546	-0.04408		-0.17327	-0.4554	-0.66926
-1.63925		-1.08307							
GENE0004		-0.40403		-0.08962		-0.07734		-0.13194	-0.04299
-0.34925		-1.49124		-0.57431					
GENE0005		-0.46961		0.04602		0.00438	0.03012	-0.26229	-
0.41479	-1.41275		-0.55466						
GENE0006		-0.49563		-0.05849		0.09863	-0.07333	-0.48383	-
0.34593	-1.39221		-0.31024						
GENE0007		-0.51306		-0.01078		0.01036	-0.15065	-0.4402	-0.4574
-1.2619	-0.76481								

- <146> 분석 결과 파일의 형식은 다음의 표 3b와 같이 구성할 수 있다.

표 3b

<147>

[유전자이름] [BSS/WSS]	
GENE0000	7.252297
GENE0001	2.127878
GENE0002	1.411586
GENE0003	4.922214
GENE0004	2.015062
GENE0005	1.860405
GENE0006	0.877066
GENE0007	3.195392
GENE0008	4.270624
GENE0009	1.747254
GENE0010	6.200477

<148>

도 7a에서, Number of Classes는 그룹의 개수를 의미한다.

<149>

Classical method는 마이크로어레이 자료의 분석에서 분류분석(또는 판별분석)의 목적은 이미 정해진 각 처리(treatment, 예를 들어 정상 조직 및 암종 조직에서 얻은 표본)를 가장 잘 분별해주는 유의한 유전자를 찾아내는 것이다. 본 발명에 따라 구현된 판별분석 프로그램은 고전적 방법으로서 (D)LDA - 선형판별분석, (D)QDA - 이차판별분석이 CV(cross validation)과 같이 구현된다.

<150>

입력 파일 형식은 train 데이터와 test 데이터를 두개가 필요하며 입력 형태는 다음의 표 4a와 같이 구성할 수 있다.

표 4a

<151>

[index1] [index2] ... [indexN] [class]
32.2 43.2 ... 54.3 1
65.4 72.3 ... 24.3 2

<152>

분석 결과 파일은 Misclassification Error(오분류율)와 각 유전자가 분류된 Group이 나타나며, CV로 연동될 경우는 오분류율만 출력된다.

<153>

옵션 입력 화면에서 DLDA는 Diagonal Linear Discriminant Analysis 방법 실행하는 것이고, DLDA(CA)는 Diagonal Linear Discriminant Analysis 방법의 모형을 평가를 위해 크로스 확인을 수행하는 것이며, DQDA는 Quadratic Discriminant Analysis 방법을 실행하는 것이고, DQDA(CA)는 Quadratic Discriminant Analysis 방법의 모형을 평가를 위해 크로스 확인을 수행하는 것이다.

<154>

Decision tree는 의사결정나무를 생성하는 서브 프로그램이다. 의사결정나무는 의사결정규칙(decision rule)을 도표화하여 관심대상이 되는 집단(class)을 몇 개의 소집단으로 분류(classification)하거나 예측(prediction)을 수행하는 분석방법이다.

<155>

입력 형태는 다음의 표 4b와 같이 구성할 수 있다. 각 유전자는 공백이나 ,(콤마)로 구분되어야 하며 맨 마지막 column에는 예측의 대상이 되는 class(target 변수)가 있어야 한다. class는 1부터 시작하는 정수이어야 한다.

표 4b

<156>

32.2 43.2 ... 54.3 1
65.4 72.3 ... 24.3 2
1번째Gene 2번째Gene class

<157>

도 7b의 옵션 입력 화면에서, Import data 부분에는 업로드한 데이터가 지정된다. Numer of class는 클래스의 개수, Number of Genes는 유전자의 개수로서 자료의 column수에서 클래스를 제외한(column수-1)를 나타낸다. N-fold cross validation는 교차타당성 검토시 data를 N개로 나누게 된다. Pruning rate에 따라 tree가 불필요하

게 크게 만들어진 경우 가지치기(Pruning)를 수행된다. 예를 들어 Pruning rate가 0.3으로 지정되었다면 70%의 자료만 tree를 만드는데 사용되고, 30%의 자료는 가지치기에 사용된다.

<158> 도 7c의 결과는 클래스가 3개인 경우이다. 첫 번째 분류는 1671번째 유전자가 59.8를 기준으로 분기가 일어나고 있다. 이 첫 번째 분기에 의해서 1번째 클래스를 모두 찾아낼 수 있다. 두 번째 분기는 첫 번째 분기의 오른쪽(1671번째 유전자가 59.8보다 큰 쪽)에서 일어난다. 1727번째 유전자가 324.8을 기준으로 분기하고 있다. tree의 결과 중 원은 분기가 일어나는 곳을 의미하며 원 아래의 []안의 숫자는 분기가 일어나는 유전자를 의미한다. 경계가 되는 값은 오른쪽 아래에 부등호와 함께 적혀있다. 더 이상 분기가 없는 곳은 정사각형으로 표현되어있다. 원과 정사각형내의 숫자는 각 클래스의 분포를 나타낸다.

<159> 본 발명이 소프트웨어로 구현될 때, 실험 설계, 표준화, 추정과 검정, 군집 분석 및 분류를 표시하는 상위 메뉴, 기타 인터페이스는 자바 서버 페이지(JSP)를 이용하여 구현될 수 있다.

<160> 본 발명이 소프트웨어로 구현될 때, 각 구성 요소인 실험 설계, 표준화, 추정과 검정, 군집 분석 및 분류는 데이터베이스를 공유하는 별도의 자바 애플릿으로 구현될 수 있다. 또한, 표준화에 포함된 단일 슬라이드 표준화, 염료 교환 표준화, 다중 슬라이드 표준화 그리고 단일 슬라이드 배치 방법을 별도의 자바 애플릿으로 구현할 수 있다. 마찬가지로, 추정과 검정에 포함된 뉴턴(Newton) 방법, B 통계량(B-statistic) 방법, 조정유의확률 방법(ADP), SAM 그리고 EBAM의 분석방법을 별도의 자바 애플릿으로 구현할 수 있다. 또한, 군집 분석에 포함된 계층적 군집 분석, k-평균 군집 분석, 주성분 분석, 자기조직도, 진웨이빙을 별도의 자바 애플릿으로 구현할 수 있다.

<161> 본 발명은 소프트웨어를 통해 실행될 수 있다. 소프트웨어로 실행될 때, 본 발명의 구성 수단들은 필요한 작업을 실행하는 코드 세그먼트들이다. 프로그램 또는 코드 세그먼트들은 프로세서 판독 가능 매체에 저장되거나 전송 매체 또는 통신망에서 반송파와 결합된 컴퓨터 데이터 신호에 의하여 전송될 수 있다.

<162> 컴퓨터가 읽을 수 있는 기록매체는 컴퓨터 시스템에 의하여 읽혀질 수 있는 데이터가 저장되는 모든 종류의 기록 장치를 포함한다. 컴퓨터가 읽을 수 있는 기록 장치의 예로는 ROM, RAM, CD-ROM, DVD±ROM, DVD-RAM, 자기 테이프, 플로피 디스크, 하드 디스크(hard disk), 광데이터 저장장치 등이 있다. 또한, 컴퓨터가 읽을 수 있는 기록매체는 네트워크로 연결된 컴퓨터 장치에 분산되어 분산방식으로 컴퓨터가 읽을 수 있는 코드가 저장되고 실행될 수 있다.

<163> 본 발명은 도면에 도시된 일 실시 예를 참고로 하여 설명하였으나 이는 예시적인 것에 불과하며 당해 분야에서 통상의 지식을 가진 자라면 이로부터 다양한 변형 및 실시 예의 변형이 가능하다는 점을 이해할 것이다. 그러나, 이와 같은 변형은 본 발명의 기술적 보호범위내에 있다고 보아야 한다. 따라서, 본 발명의 진정한 기술적 보호범위는 첨부된 특허청구범위의 기술적 사상에 의해서 정해져야 할 것이다.

발명의 효과

<164> 상술한 바와 같이, 본 발명에 의하면, 하나의 통합시스템에서 마이크로어레이 자료 분석의 전 과정을 처리함으로써, 데이터베이스를 공유하여 마이크로어레이 스캐닝 이미지 자료에 대한 체계적인 통계분석을 수행할 수 있고, 각 실험의 상황에 따른 최적의 통계분석법을 적용함으로써 부적절한 방법의 적용으로 인한 위양성 및 위음성 오류율을 최소화하고 연구결과의 신뢰성을 높일 수 있으며, 편리하고 사용자 친화적인 인터페이스를 제공하여 분석이 용이한 효과가 있다.

도면의 간단한 설명

<1> 도 1a 및 도 1b는 본 발명에 적용되는 실험 설계 방법들의 개념도이다.

<2> 도 2는 본 발명에 따른 마이크로어레이 통합 분석 방법의 흐름도이다.

<3> 도 3은 도 2를 위한 서버 프로그램인 실험 설계의 화면 구성의 일 예를 도시한 것이다.

<4> 도 4a 내지 도 4f는 도 2를 위한 서버 프로그램인 표준화의 화면 구성의 일 예를 도시한 것이다.

<5> 도 5a 내지 도 5n는 도 2를 위한 서버 프로그램인 추정과 검정의 옵션 입력의 화면 구성 예 및 실행 결과 그래프의 일 예를 도시한 것이다.

<6> 도 6a 내지 도 6o는 도 2를 위한 서버 프로그램인 군집 분석의 옵션 입력의 화면 구성 예 및 실행 결과 그래프

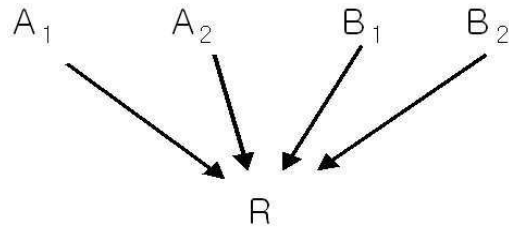
의 일 예를 도시한 것이다.

<7>

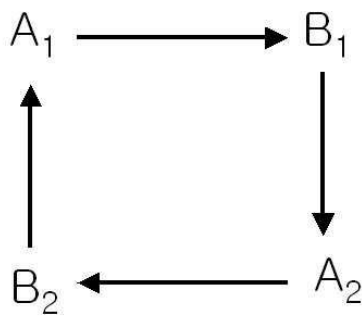
도 7a 내지 도 7c는 도 2를 위한 서브 프로그램인 분류의 옵션 입력의 화면 구성 예 및 실행 결과 그래프의 일 예를 도시한 것이다.

도면

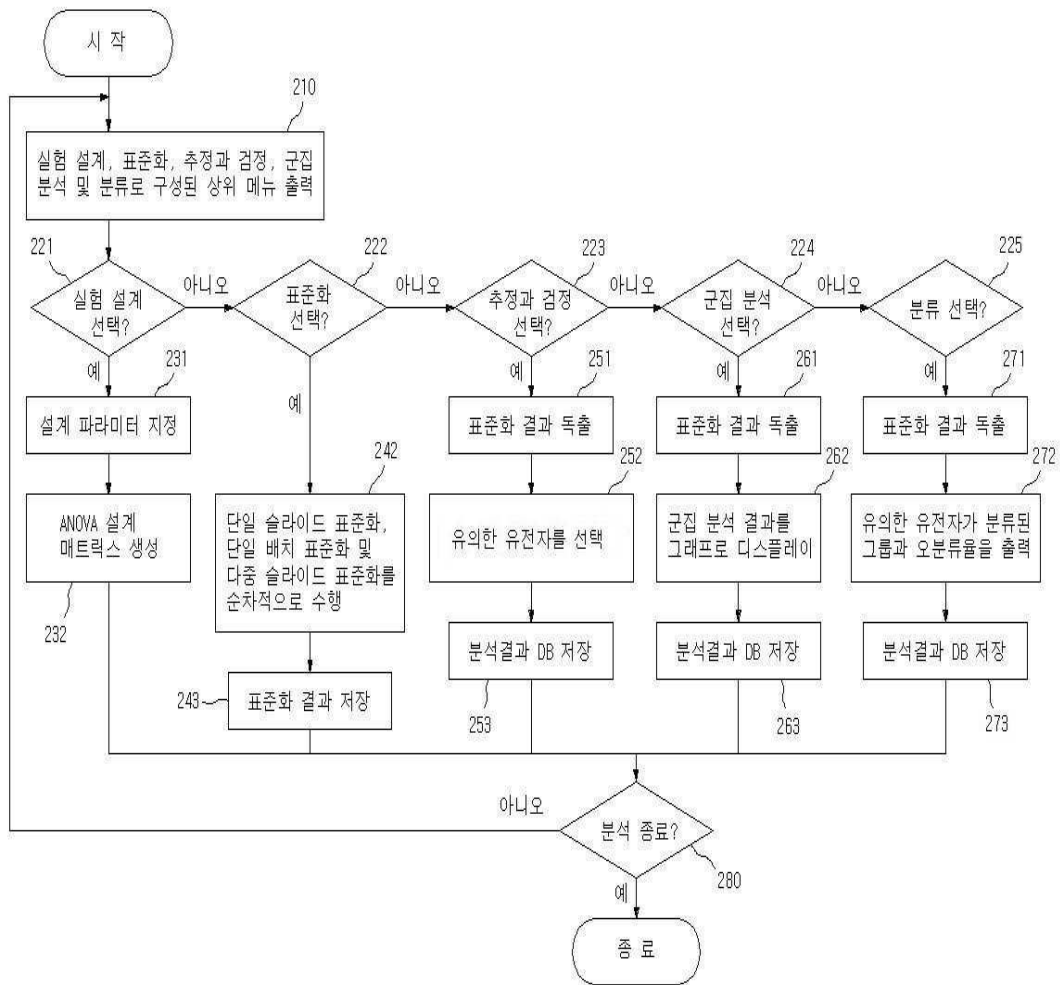
도면1a



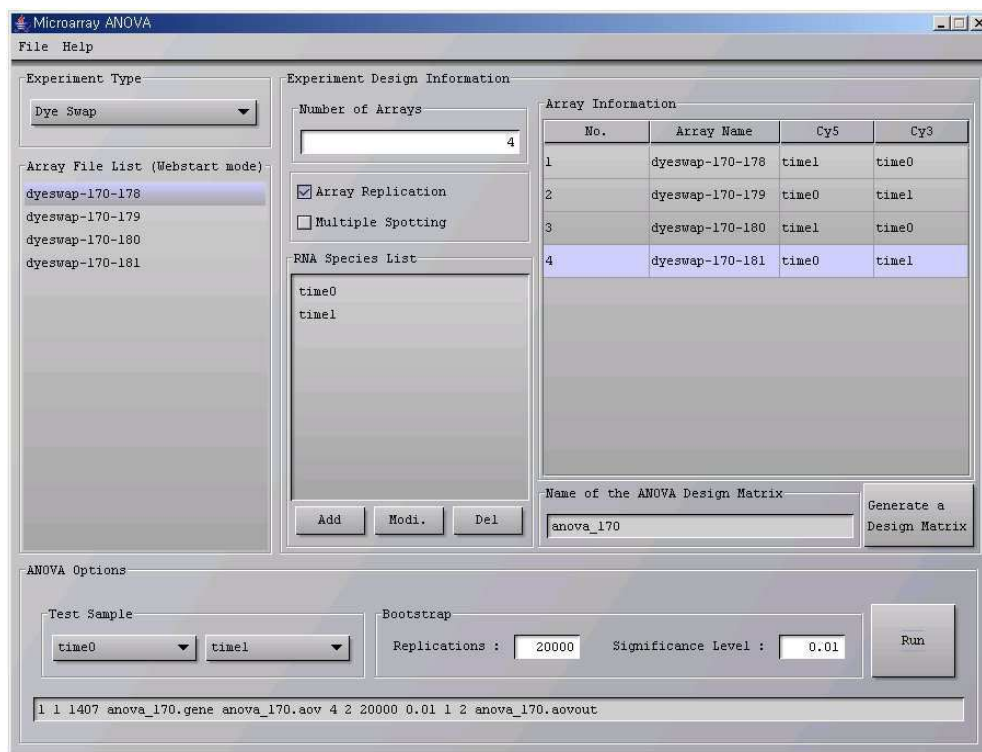
도면1b



도면2



도면3



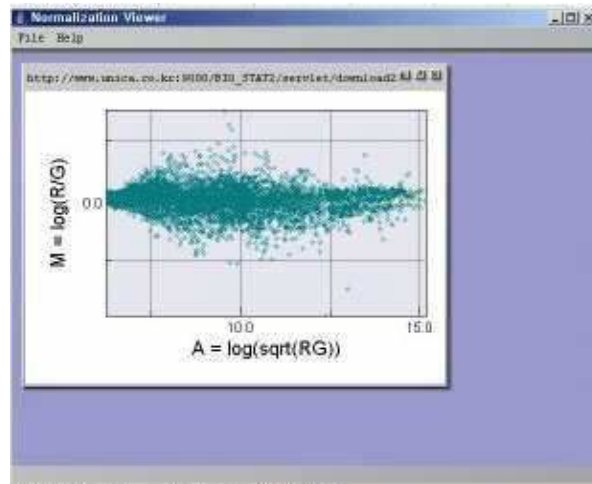
도면4a



도면4b



도면4c



도면4d

Experimental Design

- dyeswap
- reference
- loop

Normalization

- single-slide
- single-dyeswap
- single-batch
- multi-slide

Estimation & Testing

Clustering

Classification

HELP

Base Info.

Statistical analysis division: [single-dyeswap]

Project: [MT2000] Experiment: [reference chip 9c]

ETC: [single-dyeswap]

Statistical Analysis

INPUT File

Standardization Input File

- 24.1(image-analysis-1.txt)
- 24.2(image-analysis-2.txt)
- 48.1(image-analysis-5.txt)
- 48.2(image-analysis-7.txt)
- 48.3(image-analysis-8.txt)
- 48.4(image-analysis-9.txt)

Selection Input File

- 24.1(image-analysis-1.txt)
- 24.3(image-analysis-3.txt)

EXECUTE

OUTPUT File: [reference chip 9c_dyeswap.txt]

LIST

도면4e

Experimental Design

- dyeswap
- reference
- loop

Normalization

- single-slide
- single-dyeswap
- single-batch
- multi-slide

Estimation & Testing

Clustering

Classification

HELP

Base Info.

Statistical analysis division: [single-batch]

Project: [MT2000] Experiment: [reference chip 9c]

ETC: [single-batch]

Statistical Analysis

INPUT Option: ☒ Mean ☐ Standby ☐ Profile

INPUT File

Standardization Input File

- 24.1(image-analysis-1.txt)
- 24.2(image-analysis-2.txt)
- 24.3(image-analysis-3.txt)
- 24.4(image-analysis-4.txt)
- 48.1(image-analysis-5.txt)
- 48.2(image-analysis-7.txt)
- 48.3(image-analysis-8.txt)
- 48.4(image-analysis-9.txt)

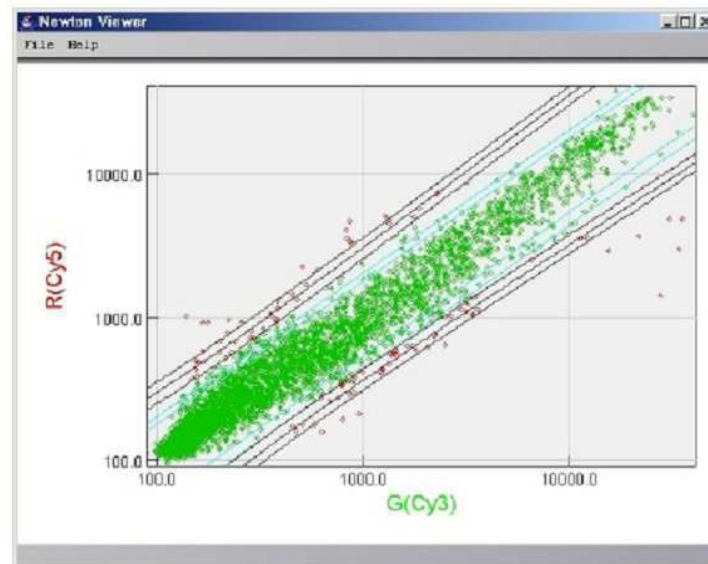
Selection Input File

EXECUTE

LIST

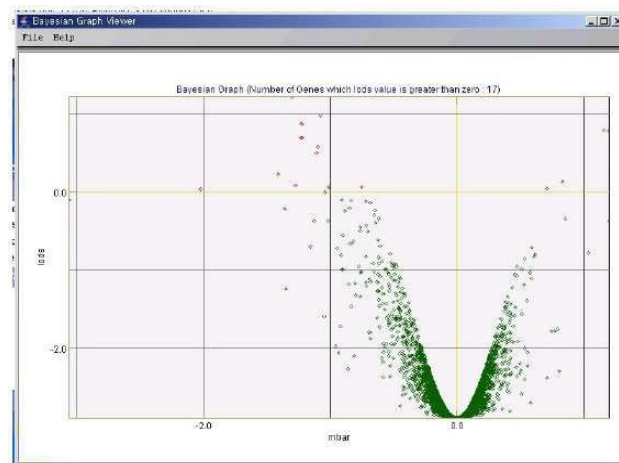
도면4f

도면5a



도면5b

도면5c



도면5d

ADP Execute Dialog

Adjusted p-value of Microarrays

Open data:

Response type:
☒ One Class Sample
☐ Two Class Sample
☐ Multiclass Sample
☐ Quantitative Response Sample

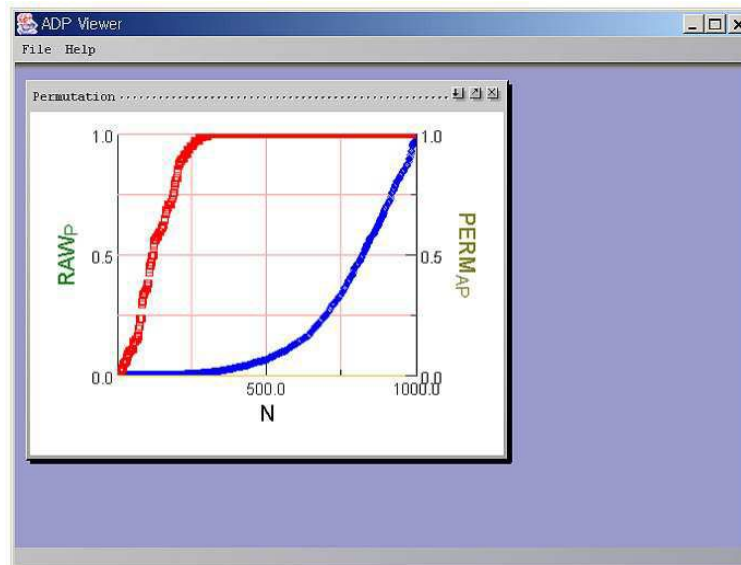
Data structure:
 Number of Genes:
 Number of Columns:

Quantitative response file:
 Number of groups:
 Group 1: Group 2: Group 3:
 Group 4: Group 5: Group 6:
 Group 7: Group 8: Group 9:

Number of permutations:
 Number of bootstraps:
 SEED Number:

p-value:
☒ Bonferroni adjusted ☒ Bonferroni Step-down adjusted ☒ Run Result Capture
☒ Sidak adjusted ☒ Sidak Step-down adjusted ☒ Raw p-value: permutation p-value
☒ Permutation adjusted ☒ Permutation Step-down adjusted ☒ Raw p-value: Bootstrap p-value
☒ Bootstrap adjusted ☒ Bootstrap Step-down adjusted ☒ Tmp Outfile Capture
☒ FDR ☒ Hochberg Step-up adjusted

도면5e



도면5f



도면5g

Execute

Significant Analysis of Microarrays

Import data: sam-127-140.txt Find...

Data structure: Number of genes: 1000, Number of columns: 9

Response type: One Class Sample, Two Class Sample, Multiclass Sample, Quantitative Response Sample

Number of permutations: 100, Seed number: 12345

Quantitative Response File: Number of groups: Group1: 8, Group2: , Group3: , Group4: , Group5: , Group6: , Group7: , Group8: , Group9:

기본 Plot: ☒ Gene Observed score, ☒ Gene Expected score, ☒ Gene plot (Observed score)

SAM Plot: ☒ SAM plot

FDR Plot: ☒ Mean FDR, ☒ Median FDR, ☒ 90th percentile FDR, ☒ FDR Table List

Ok Cancel

도면5h

Significant Analysis of Microarrays

File Graph Source Help

View Source

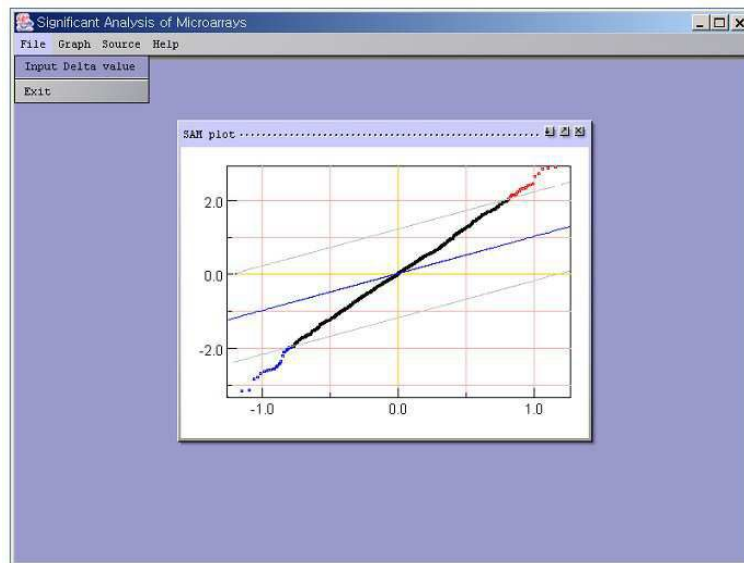
FDR.OUT

Content:

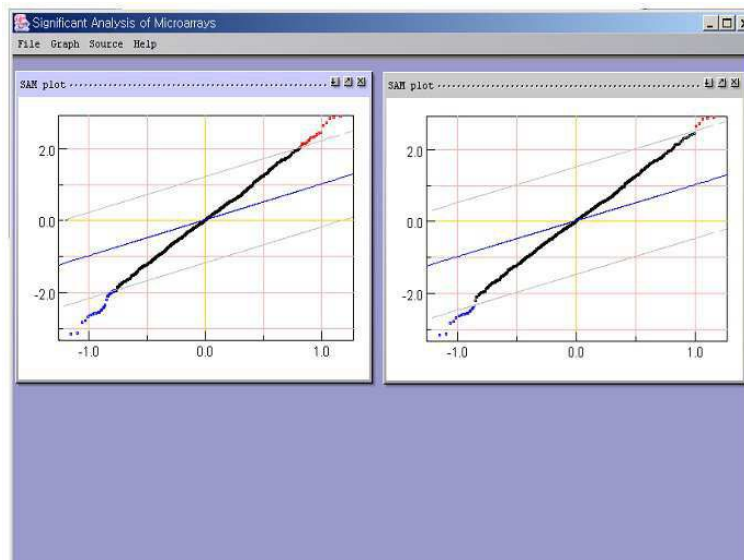
Delta	# of Sig.	# of Falsely Sig.			FDR		
		MEAN	Median	90th	MEAN	Median	90th
0.00272	996	379.708	379.860	384.150	0.38123	0.38139	0.38569
0.00461	996	379.708	379.860	384.150	0.38123	0.38139	0.38569
0.00576	996	379.708	379.860	384.150	0.38123	0.38139	0.38569
0.00637	994	378.608	378.300	383.409	0.38089	0.38058	0.38572
0.00666	993	377.391	377.325	382.980	0.38005	0.37998	0.38568
0.01430	991	373.782	373.620	380.250	0.37719	0.37701	0.38370
0.01697	986	367.626	366.990	375.648	0.37285	0.37220	0.38098
0.02598	981	359.755	358.800	371.280	0.36672	0.36575	0.37847
0.03483	975	351.066	349.245	365.898	0.36007	0.35820	0.37528
0.03622	975	351.066	349.245	365.898	0.36007	0.35820	0.37528

Close

도면5i



도면5j



도면5k

Execute

Empirical Bayesian Analysis of Microarrays

Import data:

Response type:
☐ One Class Sample
☐ Two Class Sample
☐ Multiclass Sample
☐ Quantitative Response Sample

Data structure:
 Number of genes:
 Number of columns:

Quantitative response file:
 Number of groups:
 Group1: Group2: Group3:
 Group4: Group5: Group6:
 Group7: Group8: Group9:

Option:
 Spline df:
 Fudge Factor Percentile:
 Number of permutations:
 SEED:

Data plot:
☒ Standard Z score
☒ Expression score
☒ Numerator:Denominator

EBAM Plot:
☒ Natural spline LOGIT prob...
☒ EBAM Plot

FDR Plot:
☒ Mean FDR
☒ Median FDR
☒ 90th percentile FDR
☒ FDR Table List

도면5l

Empirical Bayes Analysis of Microarrays

File Graph Source Help

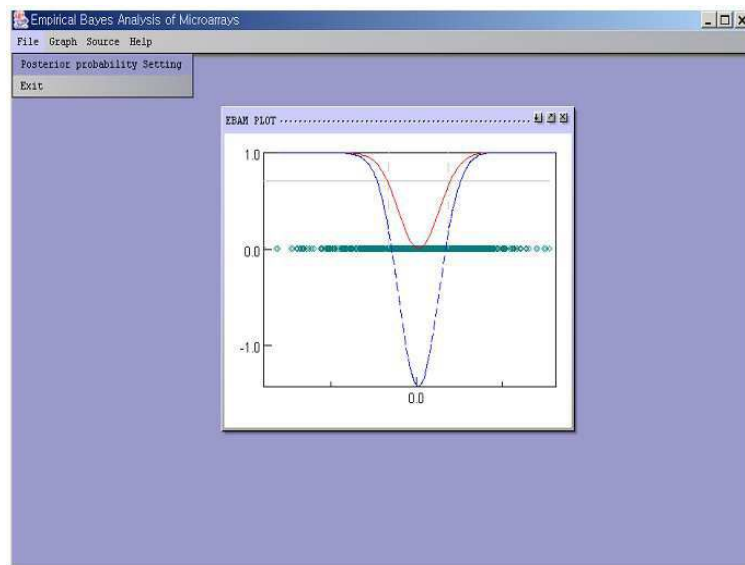
View Source

FDR.OUT

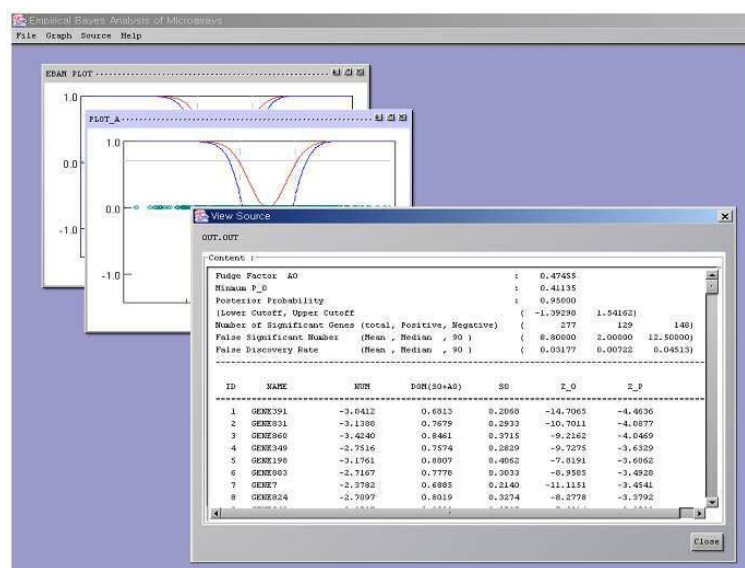
Content :

P	# of Sig.	N of Falsely Sig.			FDR		
		MEAN	Median	90th	MEAN	Median	90th
0.01000	947	853.150	858.000	902.800	0.90090	0.90602	0.95333
0.02000	941	822.000	825.000	885.500	0.87354	0.87673	0.94102
0.03000	920	780.000	782.500	848.400	0.84783	0.85054	0.92217
0.04000	901	739.200	744.000	818.300	0.82042	0.82575	0.90821
0.05000	890	711.100	710.500	800.200	0.79899	0.79831	0.89910
0.06000	879	685.800	678.000	784.200	0.78020	0.77133	0.89215
0.07000	868	663.650	656.000	771.100	0.76457	0.75576	0.88836
0.08000	862	643.250	635.500	752.100	0.74623	0.73724	0.87251
0.09000	846	621.450	611.000	739.000	0.73457	0.72222	0.87352
0.10000	838	601.850	586.500	725.900	0.71820	0.69988	0.86623

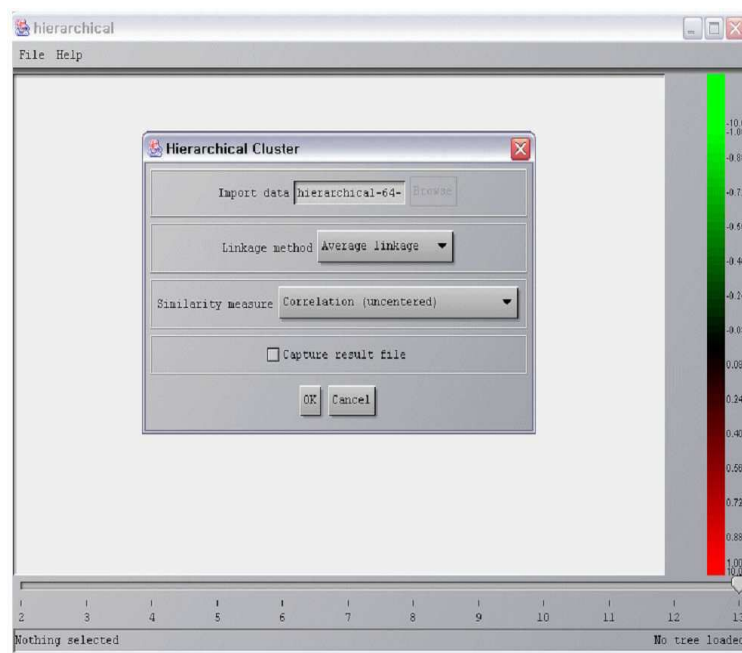
도면5m



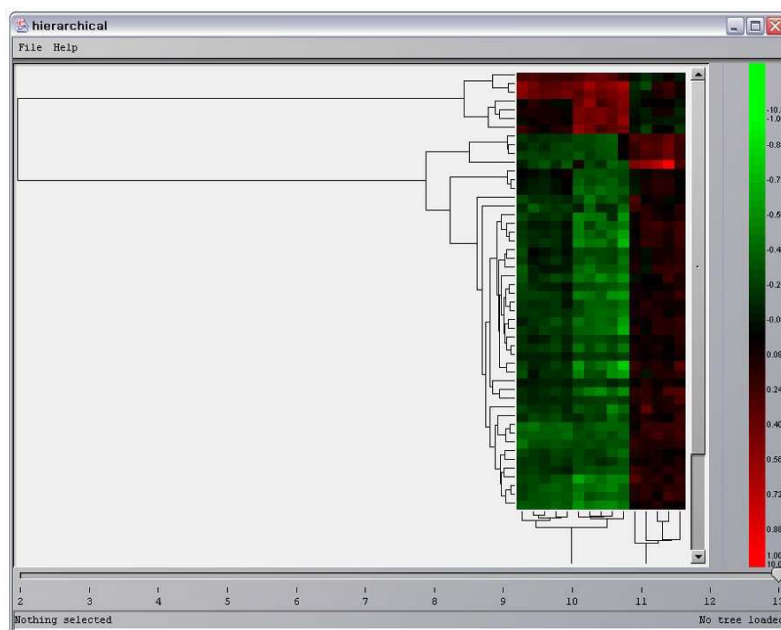
도면5n



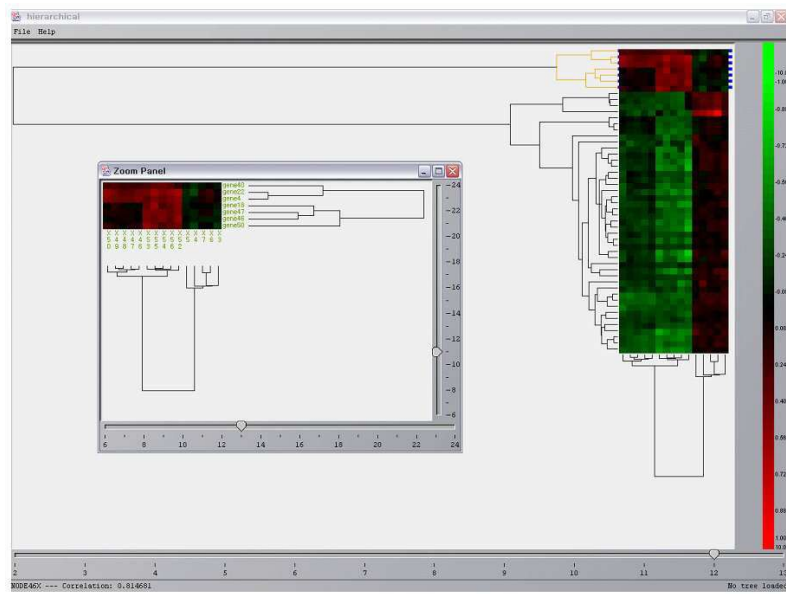
도면6a



도면6b



도면6c



도면6d

K-Means Clustering

Import data:

Target: ☒ Gene ☐ Experiment

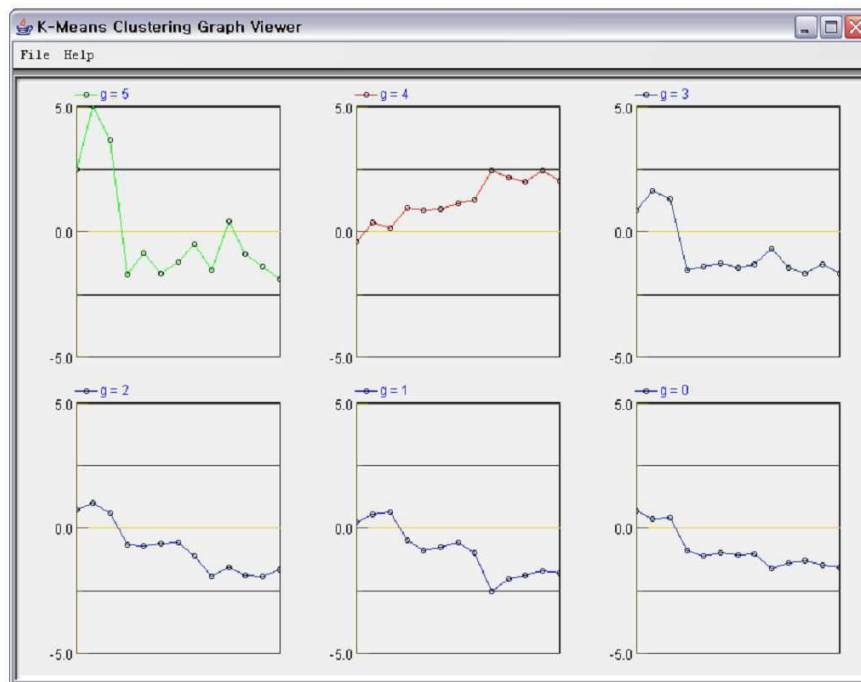
Number of clusters:

Centroid definition:

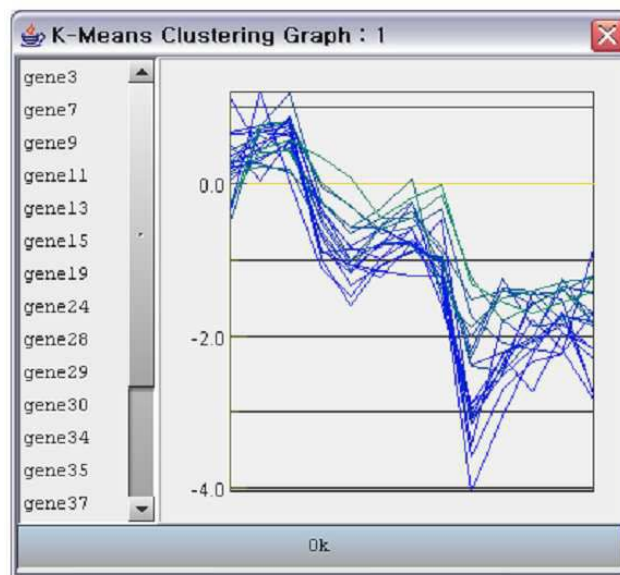
Number of iterations:

☐ Capture result file

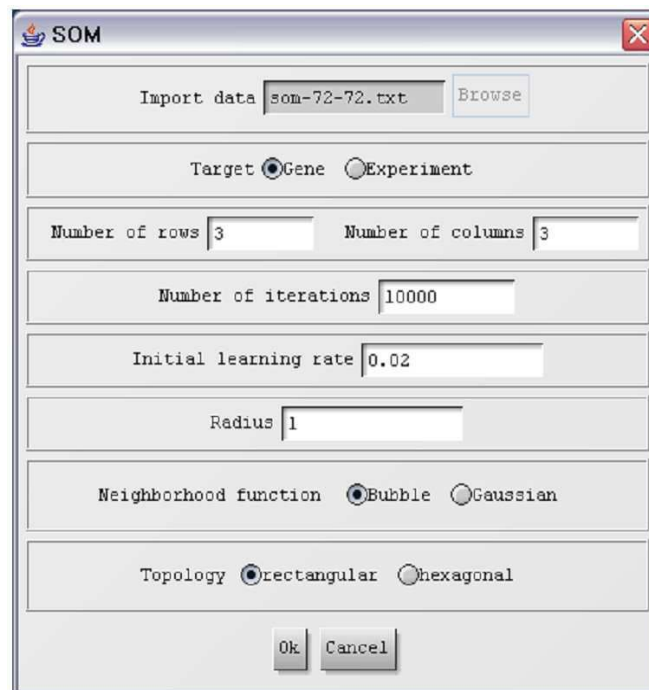
도면6e



도면6f

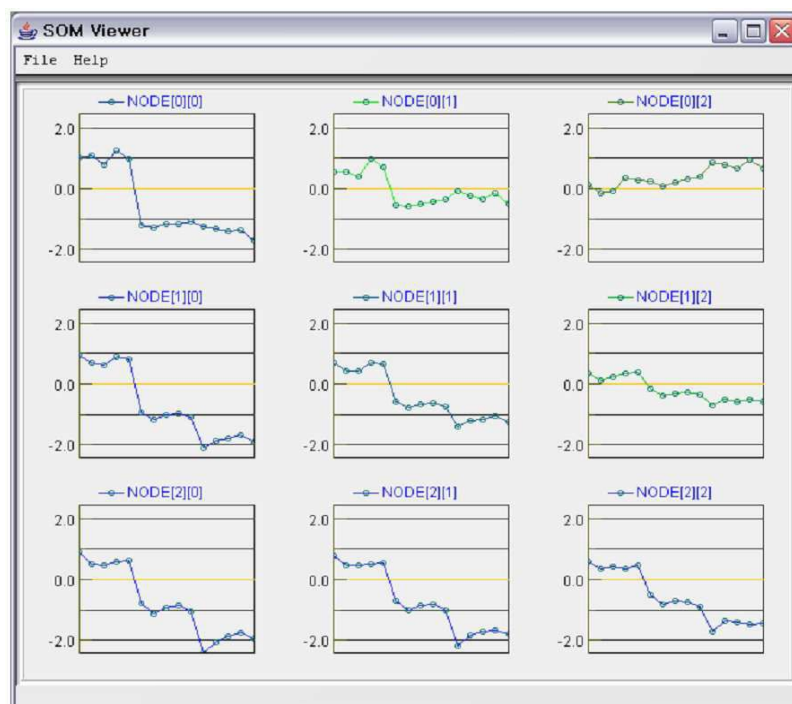


도면6g

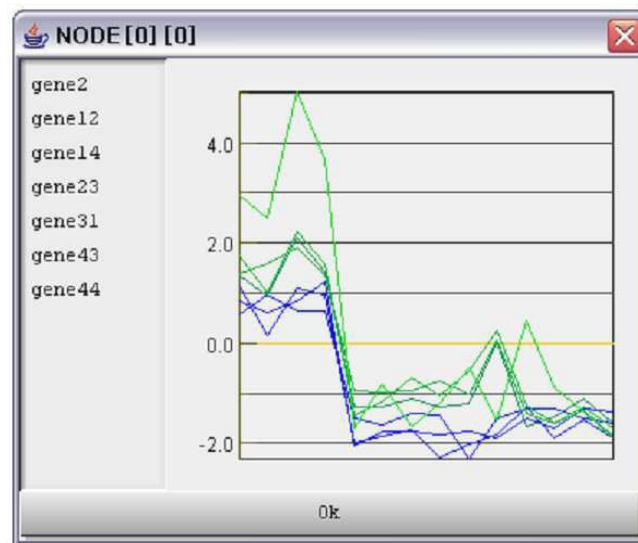


The image shows a configuration window for a Self-Organizing Map (SOM). The window has a title bar with a logo and the text 'SOM'. Inside, there are several input fields and radio buttons. The 'Import data' field contains 'som-72-72.txt' with a 'Browse' button next to it. The 'Target' section has two radio buttons: 'Gene' (selected) and 'Experiment'. The 'Number of rows' and 'Number of columns' fields both contain the value '3'. The 'Number of iterations' field contains '10000'. The 'Initial learning rate' field contains '0.02'. The 'Radius' field contains '1'. The 'Neighborhood function' section has two radio buttons: 'Bubble' (selected) and 'Gaussian'. The 'Topology' section has two radio buttons: 'rectangular' (selected) and 'hexagonal'. At the bottom, there are 'Ok' and 'Cancel' buttons.

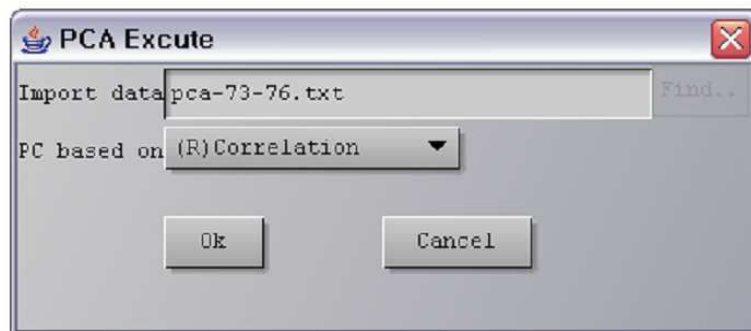
도면6h



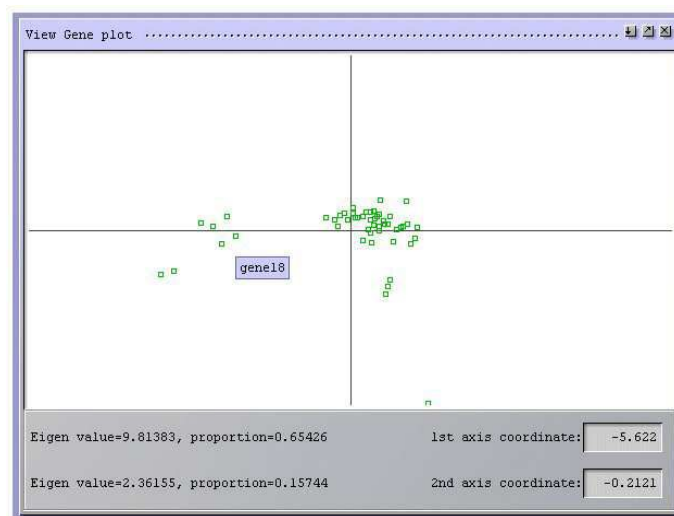
도면6i



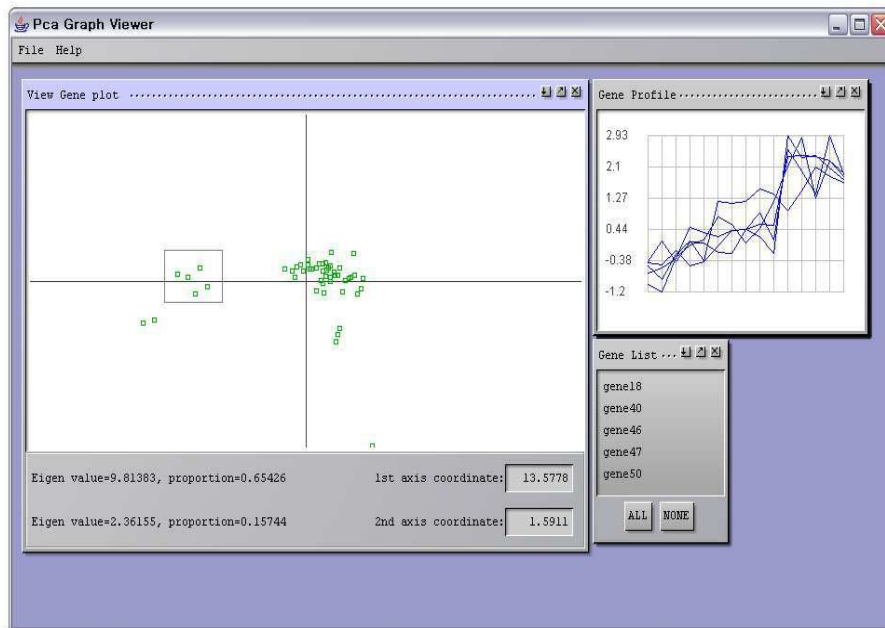
도면6j



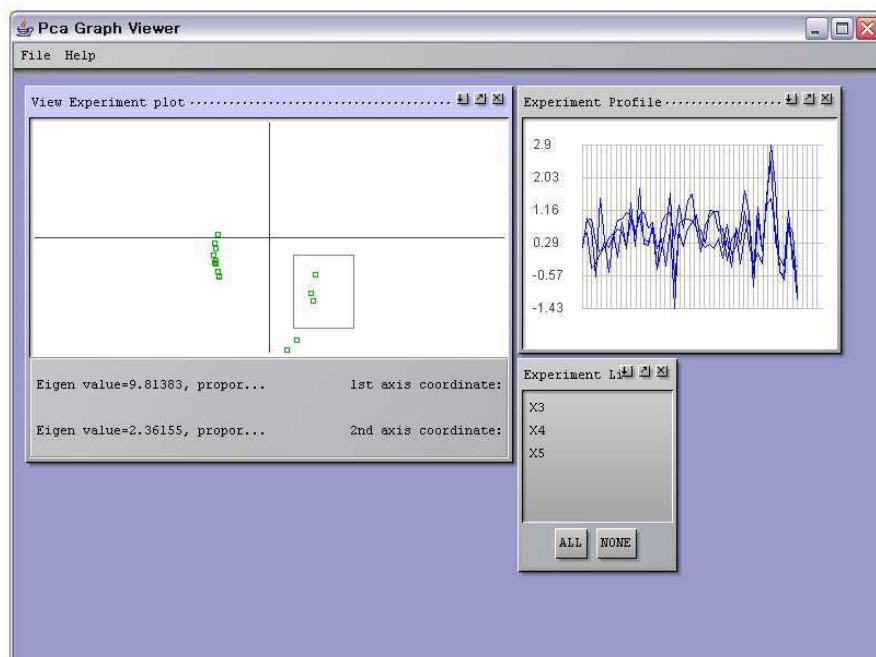
도면6k



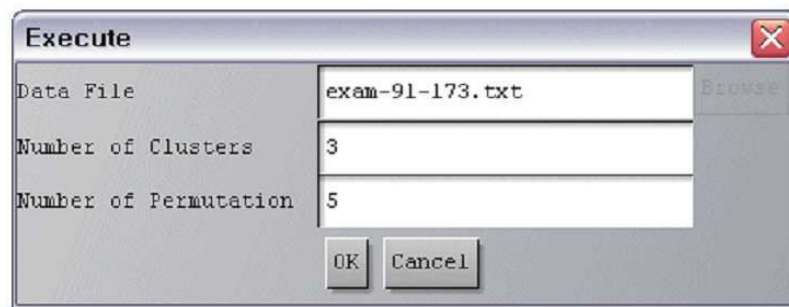
도면61



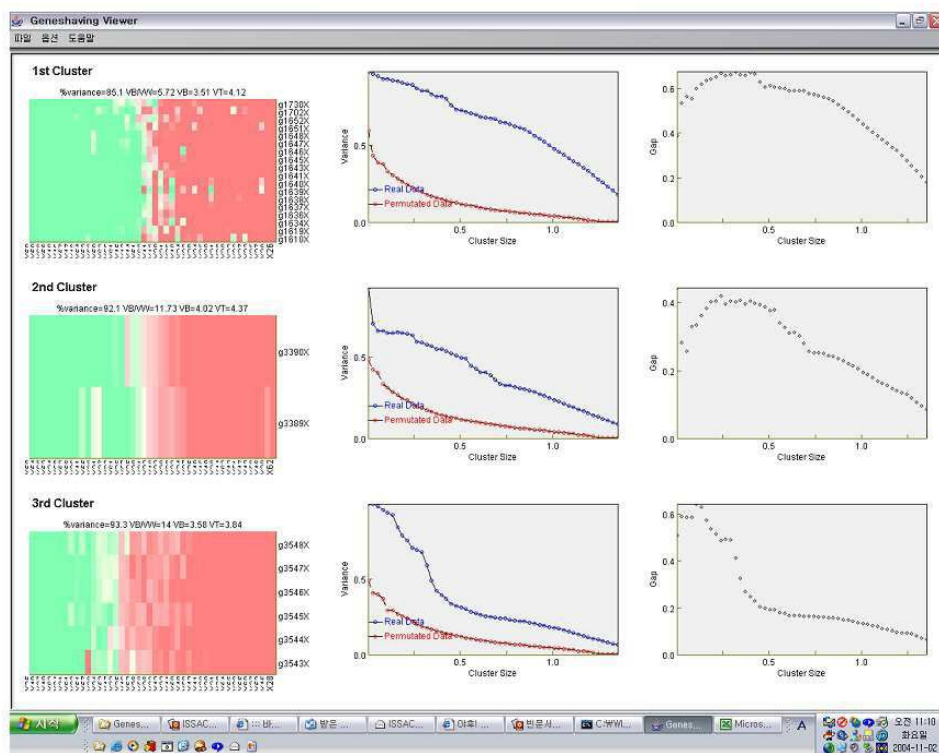
도면6m



도면6n



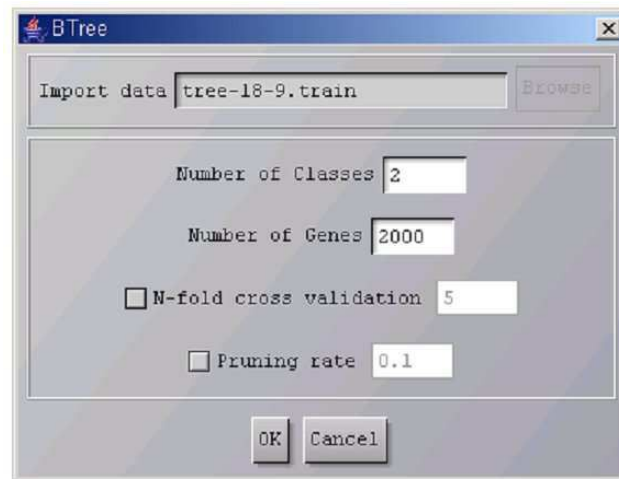
도면6o



도면7a



도면7b



도면7c

