SMILE : 마이크로어레이 데이터 저장·관리·분석을 위한 통합 LIMS 개발

이정원° 진희정 조환규
부산대학교 컴퓨터공학과
{jwlee°, hjjin, hgcho}@pearl.cs.pusan.ac.kr

SMILE : Development of an Integrated LIMS for Management and Analysis of Microarray Data

Jeong-Won Lee° , Hee-Jeong Jin, Hwan-Gue Cho
Dept. of Computer Engineering, Pusan National University

요 약

마이크로어레이 실험의 등장으로 한 번에 수백 개에서 수천 개의 유전자를 실험할 수 있게 되었다. 이는 기존의 실험과 비교했을 때 질적인 측면과 양적인 측면에서 가히 혁신적이라 할 수 있다. 마이크로어레이 칩을 이용한 실험에서 쏟아져 나오는 엄청난 데이터를 비교, 분석, 관리하기 위해서는 실험실의 마이크로어레이 분석 소프트웨어나 시스템간의 데이터 형식이 호환되어야 하며, 소프트웨어의 지원 또한 획기적이고 효율적이어야 한다. 본 논문에서는 다양한 종류의 마이크로어레이 입력 데이터 및 분석 데이터를 다룰 수 있고, 표준 파일 형식으로의 변환 기능을 제공하며, 마이크로어레이 이미지 분석용 소프트웨어인 ArrayMall[1,2]과 유전자 조절 네트워크 분석 시스템인 GENAW[3]를 통합하고 마이크로어레이 실험 데이터의 분석, 관리 및 데이터 공유를 위한 분산 시스템인 SMILE[4]에 대해 소개한다.

1. 서 론

마이크로어레이 칩은 기능 유전체학의 주요 도구 중 하나로 염기서열을 알고 있는 DNA 분자를 소형 기판위에 고밀도로 배열해 놓은 것이다. 마이크로어레이 칩을 이용하여 유전자에 대한 분석을 함으로써, 많은 양의 유전자에 대한 실험을 한 번에 수행할 수 있게 되었다.

마이크로어레이 실험에서부터 데이터 분석을 통하여 유전자의 기능이나 유전자 간의 관계 등을 밝혀내기 위해서는 여러 가지 단계의 분석 과정이 필요하다. 우선 그 첫 번째로 마이크로어레이 칩을 이용한 실험이 이루어지고, 마이크로어레이 칩을 레이저를 이용하여 이미지로 스캔한다. 스캔된 이미지는 스팟(spot)들의 발현 량을 측정하기 위해 이미지 그리딩(gridding) 작업과 같은 이미지 프로세싱 작업이 수행된다. 이미지 작업 후 측정된 스팟들의 발현 량은 정규화 과정을 통하게 되고, 그 뒤 질병 조기 진단, 유전자 네트워크 구성 등을 위해서 분석 과정이 이루어진다. 이에 따라 실험 결과로 생성되는 많은 양의 데이터를 효율적으로 비교, 분석, 관리할 수 있는 시스템이 필요하게 되었고, 다양한 마이크로어레이 데이터의 관리를 위한 LIMS(Laboratory Information Management System), 마이크로어레이 이미지 분석 툴, 스팟 분석 툴, 유전자 발현 데이터 분석 툴, 유전자 정보 검색 프로그램들이 개발되었다. 연구자들은 하나의 실험 데이터의 결과를 얻기 위해, 이러한 다양한 시스템들을 사용하여 마이크로어레이 데이터를 분석하여야하며, 각각의 분석 단계마다 사용하는 시스템들이 분리 되어있어 각 시스템들 간의 입출력 포맷을 맞추어주어야 했다.

또한 연구자들이 마이크로어레이 데이터의 분석 및 관리를 위해 LIMS 시스템을 사용하고 있는 경우, 마이크로어레이 데이터의 분석을 위한 기존의 여러 시스템들이 사용하고 있는 LIMS 시스템과 연계된 것이 아니므로 연구자 개개인의 개인 PC에서 여러 가지 분석을 수행하고, 이들 결과에서 의미 있다고 판단되는 것들을 연구실 전체의 LIMS에 저장하여 관리한다. 따라서 한 연구실에서 LIMS 시스템을 사용하고 있다고 하더라도, 연구자 개개인을 위한 개인 PC용 LIMS 시스템이 필요하다.

본 논문에서는 다양한 종류의 마이크로어레이 입력 데이터 및 분석 데이터를 다룰 수 있고, 표준 파일 형식으로의 변환 기능을 제공하며, 마이크로어레이 이미지 분석용 소프트웨어인 ArrayMall[1,2]과 유전자 조절 네트워크 분석 시스템인 GENAW[3]를 통합하고 마이크로어레이 실험 데이터의 분석, 관리 및 데이터 공유를 위한 분산 시스템인 SMILE[4]에 대해 소개한다. 또한 SMILE은 GenePix나 Imagene과 같은 기존의 마이크로어레이 분석 시스템들과의 연계를 위해서 마이크로어레이 데이터의 모든 정보들을 MIAME 표준안에 맞추어서 저장, 관리하고 있다. 따라서 본 시스템에서 제공하고 있는 ArrayMall과 GENAW 외의 기존의 시스템들과의 연계도 가능하다.

2. 기존의 마이크로어레이 LIMS와의 비교

마이크로어레이 실험에서 나온 데이터의 체계적인 관리를 위해서 개발된 LIMS에는 Acuity[6], Argus[7]와 같이 상업용으로 제공되는 시스템과, BASE[8]와 같이

open project로 공개되어 있는 시스템이 있다. 기본적으로 마이크로어레이 LIMS가 제공하는 기능은 다음과 같다.

- 마이크로어레이 실험 단계에 따른 데이터의 계층적 관리 기능
- 데이터 검색 기능
- 사용자 관리 및 보안 기능
- 외부 DB 연결 기능

이외에도 이미지 분석 기능이나 정규화 기능과 같은 부가적인 기능을 제공하기도 한다. 표 1에서 보듯이, 대부분의 마이크로어레이 LIMS는 기본적인 데이터의 계층적 관리, 반복 실험 처리, 정보 검색 기능, 정규화 기능, 사용자 관리 및 보안, 외부 DB 연결 기능을 제공하고 있으며, 시스템에 따라 부가적인 기능을 제공하는 것도 있다. SMILE은 다른 마이크로어레이 LIMS들과는 달리, 프로젝트의 작업 흐름을 관리하기 위한 스케줄링 및 메모 항을 포함하는 스케줄 관리 기능과 유전자 조절 네트워크 분석 기능, 공동 연구 데이터들을 관리하기 위한 공유 데이터 관리 기능들을 추가적으로 제공한다.

	Acuity	BASE	SMILE
분석 파일 업로드 기능	O	O	O
계층적 관리	O	O	O
메타 파일 생성 기능	X	O	O
반복 실험 데이터 처리	X	O	O
정규화 기능	O	O	O
3D 가시화	O	O	O
이미지 분석	O	O	O
유전자 조절 네트워크 분석	X	X	O
스케줄러 및 메모 기능	X	X	O
레포팅 기능	O	O	O
MIAME-ML export	O	O	O
MAGE-OM import	X	O	X
다양한 데이터 파일 지원	O	O	O
데이터 백업 및 복구	O	X	O

표 1. 기존 LIMS와의 비교 : SMILE은 기존의 LIMS들과 비교하여, 유전자 조절 네트워크 분석 기능과 프로젝트 스케줄 관리, 공유 데이터 관리 기능들을 제공한다.

3. SMILE(Small and solid Microarray LIMS Experimentors)

3-1. SMILE의 기능

SMILE은 마이크로어레이에 관련된 실험, 분석 데이터를 효율적으로 관리할 수 있는 리눅스 기반의 LIMS 시스템으로, 단순한 데이터의 저장, 관리 기능뿐만이 아니라 마이크로어레이 데이터를 분석, 관리, 저장하기 위한 다양한 기능들을 지원한다.

① 데이터베이스 관리 : 다양한 마이크로어레이 데이터의 효율적인 저장, 관리를 위한 데이터베이스를 구축하고, 저장된 데이터의 백업/저장 모듈 기능을 제공하며, 데이터베이스를 사용하는 연구자들의 계층의 구분을 통한 접근 권한을 관리한다.

② 그룹웨어 기능 제공 : LIMS를 사용하는 연구자들 간의 원활한 의사소통을 위해, 쪽지 함과 게시판, 메일 보내기, 프로젝트의 스케줄을 관리할 수 있는 기능을 제공한다. 그림 1은 SMILE에서 제공하는 스케줄러 기능을 보여준다.

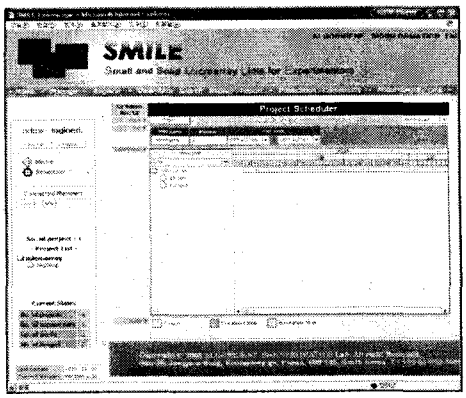

그림 1. 스케줄러 실행 화면 : 스케줄러 화면에서는 현재 로그인한 연구자가 해야 하는 작업 내용을 볼 수 있으며, 관리자는 프로젝트에 포함된 연구자들의 작업을 할당하고 관리할 수 있다.

③ 표준화 : 마이크로어레이 데이터에 대한 표준화 작업을 위해서 MIAME 표준안 분석하고 MAML 포맷으로 export하는 기능을 제공하며, MAML 포맷은 연구자의 설정에 따라 제공되는 포맷을 달리한다.

④ 외부 데이터베이스와의 연계 : LIMS 시스템에 저장된 유전자들의 보다 다양한 정보를 연구자들에게 제공하기 위해서, GeneCard를 링크시켜 정보를 제공한다.

⑤ 검색 : 저장되어있는 데이터에서 연구자들에게 많은 정보로부터 필요한 정보만을 검색하여 볼 수 있도록, 단순검색, 여러 필드에 대한 and/or 추가 검색, 전체 검색 등의 기능을 제공한다.

⑥ 가시화 기능 : 마이크로어레이 이미지 뷰어, 스팟 2D·3D 뷰어, 멀티 스팟 2D·3D 뷰어, histogram/scatter plot, 정규화 가시화, 패턴 가시화 등의 가시화 기능을 제공한다. 그림 2는 SMILE에서 제공하는 다양한 이미지 뷰어들을 보여준다.

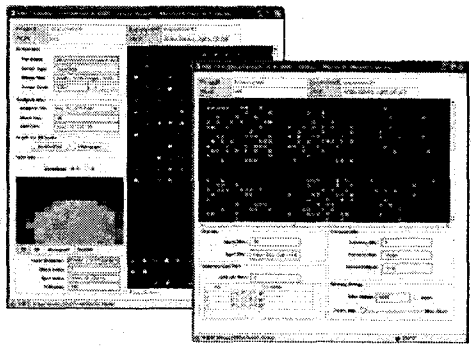

그림 2. SMILE에서 제공하는 이미지 뷰어의 예 : 마이크로어레이의 스팟 뷰어와 HeatMap

⑦ SMILE들 간의 데이터 공유 : Friend-to-friend(F2F 혹은 P2P)[9] 관계를 기반으로 하여, "친구(friend)" 관계에 있는 SMILE 시스템들 간의 데이터 공유를 지원해주는 분산 LIMS 기능을 제공한다. 데이터 공유에 대한 권한은 "읽기"와 "저장"의 두 가지로 나누어져있다. 그림 3은 SMILE들 간의 데이터 공유 관계를 나타낸 것이다. 데이터를 제공하는 SMILE을 Service Provider, 데이터를 공유 받는 SMILE을 Client로 정의한다. 데이터 공유는 Project, Experiment, Work, Shop의 마이크로어레이 데이터 각각의 구조 단계에서 허용을 할 수 있으며, 허용된 단계 의 하위 데이터들을 모두 제공하는 것이다.

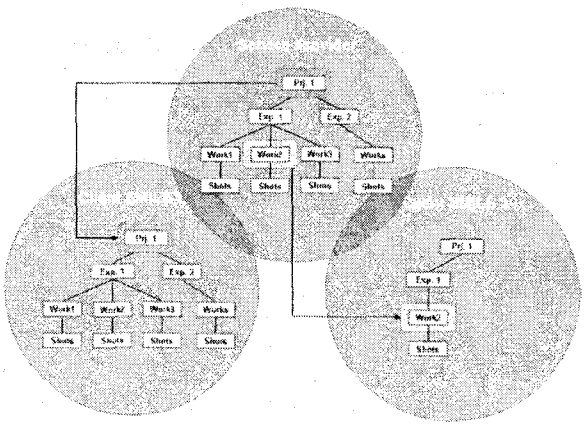

그림 3. SMILE들 간의 데이터 공유 : 친구 관계가 설정된 SMILE들 간의 데이터 공유가 이루어진다. Service Provider에서 Client SMILE1에게는 Project 1의 단계에서 데이터 공유를 제공하고 있으므로, Client SMILE 1에서는 Service Provider의 Project 1의 모든 데이터에 대한 접근이 가능하며, Project 1을 자신의 프로젝트처럼 쉽게 접근할 수 있다. 이와는 달리 Service Provider에서 Client SMILE 2에는 Work 2단계에서부터 데이터 공유 권한을 제공함으로 Client SMILE 2에서는 Work 2가 속한 데이터들에 대해서만 접근할 수 있다.

3-2. 데이터 저장 구조

SMILE에서는 마이크로어레이 실험 단계에 따른 데이터를 계층적으로 저장한다. 이는 마이크로어레이 분석 소프트웨어 및 시스템간의 데이터 호환 및 공유를 위한 기본 틀을 제공해준다. 데이터 관리는 그림 4에서 보듯이 project, experiment, work, shot으로 계층화되며, 각각의 정보는 다음과 같다.

- Project : 생물학자들이 실험에서 최종적인 결과를 얻기 위해 하는 일련의 모든 실험 과정으로, 하나 또는 그 이상의 experiment의 집합을 나타낸다.
- Experiment : 마이크로어레이실험에서 실험 대상과 마이크로어레이 디자인이 같고, 실험 조건이 다른 Work들의 집합을 나타낸다.
- Work : 실험 방법과 조건이 같고, 마이크로어레이 실험에서 사용된 실험 대상과 마이크로어레이 디자인이 같은 Shot의 집합을 나타낸다.
- Shot : 마이크로어레이 실험의 결과로 생성된 이미지로, 이미지가 생성된 시간 순서에 따라 등록된다.

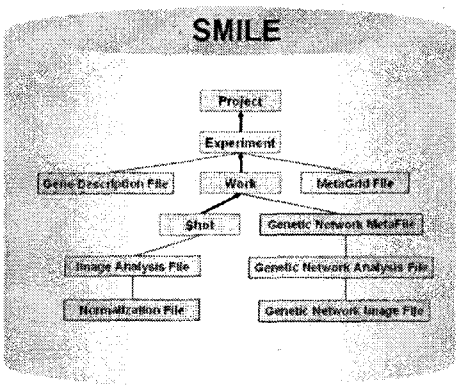

그림 4. SMILE의 데이터 관리 계층도

예를 들어, 벼의 생장에 영향을 미치는 유전자를 알아내고자 하는 연구를 수행할 때 이를 하나의 Project로 생성하고, 벼의 생장에 영향을 주는 요인으로 "빛"에 대한 연구를 수행한다면 이를 하나의 Experiment로 생성하며, 빛의 양에 따른 실험을 수행하게 되면 그 각각의 실험을 Work 단위를 생성할 수 있다. 마지막으로 각 실험에서 얻은 이미지들을 Shot에 등록하여 저장, 관리, 분석할 수 있다. 데이터 관리 계층과 함께 실험의 각 단계에서 몇몇 데이터 파일들이 생성될 수 있다(그림 4).

- Gene Description File : 실험에 사용된 유전자들에 대한 정보 파일이다.
- MetaGrid File : 동일한 유전자 집합을 이용하여 하나의 실험을 하는 경우, 마이크로어레이의 배열이 대부분 같게 된다. 이러한 마이크로어레이의 배열을 저장한 파일이며, MetaGrid File을 이용하여 이미지 분석을 보다 쉽게 수행 할 수 있다.

- MetaFile : 마이크로어레이 실험을 통하면 각각의 유전자에 대한 여러 분석 값들(발현양의 평균값, 분산 값, 각 채널에 대한 값 등)을 얻을 수 있다. 이 때 여러 실험 결과에서 사용자가 원하는 필드 값들만을 선택하여 생성한 파일이다.
- Genetic Network MetaFile : 유전자 조절 네트워크 분석 프로그램의 입력 파일로, 여러 개의 이미지 분석 파일을 가지고 유전자 발현 강도의 중앙값들만을 모아 놓은 메타 파일로 만든 파일이다.
- Genetic Network Analysis File : 유전자 조절 네트워크 프로그램을 수행한 결과로 생성된 분석 파일이다.
- Genetic Network Image File : 유전자 조절 네트워크 프로그램을 수행한 결과로 생성된 유전자 조절 네트워크의 구성 그림이다.
- Image Analysis File : 실험에서 얻어진 이미지에 대해 이미지 분석 프로그램을 통해 분석한 파일이다.
- Normalization File : 하나 이상의 이미지 분석 파일에 대해 정규화를 수행하여 만들어진 파일이다.

3-3. 마이크로어레이 데이터의 표준화

마이크로어레이 데이터는 개개 유전자 및 조직표본에 대한 정보와 실험조건, 실험 방법들에 대한 상세한 주석이 있어야 해석이 가능하기 때문에 다차원적인 자료구조를 필요로 한다. 이런 데이터의 복잡성 때문에 체계적인 데이터 관리나 데이터 공유에 어려움이 있다. 이를 해결하기위해 MGED(Microarray Gene Expression Data) 단체에서는 마이크로어레이 데이터를 공유할 수 있도록 MIAME(Minimal Information About a Microarray Experiment)와 같은 표준화 작업을 진행해왔다. MIAME는 마이크로어레이 데이터와 실험에 대한 환경 정보들을 정의하는 것이다. 이는 데이터를 잘못 해석하지 않도록 최소한의 정보를 정의하고 XML 문서의 형태로 저장하여 데이터의 교환과 데이터 처리의 자동화, 그리고 검색을 쉽게 할 수 있도록 하는데 목적이 있다.

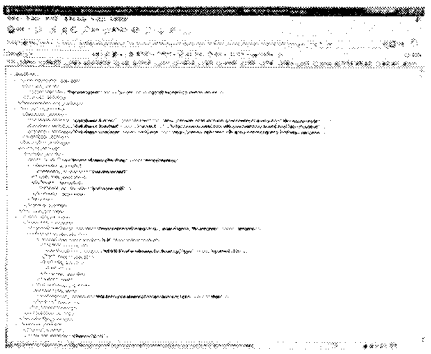

그림 5. MAGE-ML 포맷의 표준화 기능

SMILE은 MIAME 명세에 따라 그림 3의 Image Analysis File 및 MetaFile에 대해 MAGE-ML로 변환시켜 준다(그림 5). 즉, SMILE의 MAGE-ML은 XML기반으로 마이크로어레이 디자인, 마이크로어레이 제조 정보, 실험 셋업과 실행의 정보, 유전자 발현 데이터와 데이터 분석 결과들에 대한 정보를 표현하며, MAGE-ML의 DTD는 XML 문서 구조를 정의한 것으로 사용자가 문서에 사용될 수 있도록 허용하는 엘리먼트의 목록과 Cardinality 등을 나타낸다. 또한 MAGE-ML 포맷은 그 크기가 아주 크므로, 연구자의 설정에 따라 몇 개의 항목만 저장하도록 제공되는 포맷을 달리한다. 이를 통하여 기존의 마이크로어레이 분석 시스템들인 GenePix, ImaGene과 같은 시스템들의 결과 파일을 SMILE에서 저장, 관리, 분석할 수 있다.

3-4. 기존 개발 시스템들과 SMILE의 연계 구성도

SMILE 시스템은 기존의 마이크로어레이 시스템들과 연계하기 위해서 데이터를 표준화하고 다양한 분석파일들을 허용한다. 본 장에서는 현재 SMILE과 연계되어있는 마이크로어레이 시스템인 ArrayMall과 GENAW를 보여준다.

① ArrayMall : 연구자 개개인이 개인 PC에서 사용할 수 있는 마이크로어레이 저장, 관리, 분석 시스템으로 마이크로어레이 데이터의 저장, 관리를 위한 LIMS 시스템인 DataShop, 이미지 분석을 위한 ArrayShop, 스팟들의 발현 값을 정규화해주는 NormalShop, 유전자의 발현 패턴을 분석하는 ExpreView로 이루어져있다.

② GENAW : GENAW는 일련의 마이크로어레이 실험 데이터들의 결과를 이용하여 유전자 네트워크를 구성하는 시스템이다.

그림 6은 3가지 시스템인 SMILE, ArrayMall, GENAW의 구성을 보여준다. ArrayMall의 LIMS 시스템인 DataShop과 SMILE의 저장되어있는 마이크로어레이 데이터들의 정보 파일을 이용하여 서로 연계되며, SMILE과 GENAW는 SMILE에서 GENAW에 사용할 수 있도록 메타 파일을 생성하여 서로 연계된다. 따라서 본 3가지 시스템을 연계하여 사용함으로써 연구자들은 마이크로어레이 실험 이미지 데이터의 분석에서부터 다양한 마이크로어레이 데이터의 저장, 관리, 유전자 네트워크의 구성까지 모두 수행할 수 있다.

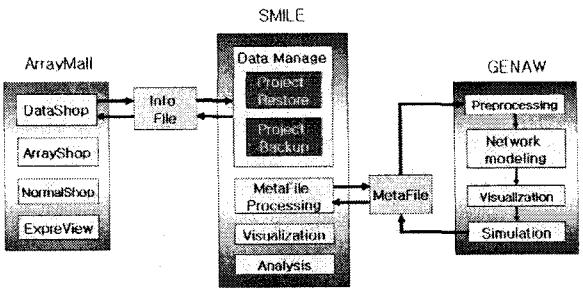

그림 6. 시스템의 구성도 : 마이크로어레이의 실험 결과 생성되는 다양한 데이터들의 저장, 분석, 관리를 위해 ArrayMall, SMILE, GENAW의 세 가지 시스템을 연계하였다.

4. 결론

마이크로어레이 실험에서부터 생물학적 의미를 찾기 위해서는 다양한 분석 작업이 수행되며, 이를 수행하는 동안 다양하고 많은 분석 파일들이 생성된다. 하나의 마이크로어레이 실험에는 여러 연구자들이 함께 연구를 하게 됨으로 생성되고 분석되는 마이크로어레이 데이터 파일들은 더욱 많아지게 된다. 또한, 마이크로어레이 데이터의 분석 방법은 다양하며, 각 분석 방법들은 순차적으로 수행되게 된다. 연구자들은 각각의 분석 작업에 기존에 개발되어 있는 다양한 시스템들도 다양하므로, 각 단계별로 분석을 진행할 때 사용하는 시스템들이 상이해지고, 이에 따라 각 시스템의 입출력 포맷을 연구자가 맞추어서 사용해야했다.

본 논문에서는 다양한 종류의 마이크로어레이 입력 데이터 및 분석 데이터를 다룰 수 있고, 표준 파일 형식으로의 변환 기능을 제공하며, 마이크로어레이 이미지 분석용 소프트웨어인 ArrayMall과 유전자 조절 네트워크 분석 시스템인 GENAW를 통합하고 마이크로어레이 실험 데이터의 분석, 관리 및 데이터 공유를 위한 분산 시스템인 SMILE에 대해 소개한다.

본 논문에서 소개한 시스템에 관한 보다 자세한 내용은 http://neobio.cs.pusan.ac.kr:8080/smile에서 찾을 수 있다.

5. 참고 문헌

[1] 천봉경, 장철진, 진희정, 이평준, 김혜정, 조환규. "ToMAS : 마이크로어레이 이미지 분석용 컴포넌트 도구", *한국정보과학회논문지*. 241-243. 2004

[2] Bong-Kyung Chun, Pyung-Jun Lee, H.J. Jin, M.J. Jun, J.H. Yoon, C.J. Jang, K.S. Lee, H.J. Kim, Hwan-Gue Cho, "ToMAS : software development Toolkits fot Microarray Analysis System,", Proc. of ISMB, Poster, Glasgow, Scotland, July 31- August 4, 2004

[3] 이경신, 조환규, 박선희. "유전자 조절 네트워크 분석을 위한 통합 시스템 개발", *한국정보과학회논문지*. 283-285. 2004

[4] 김혜정. "마이크로어레이 데이터 공유를 위한 분산 LIMS 개발", *한국정보과학회 학술 발표집*. 2005.

[5] David J. Duggan, Michael Bitter, Yidong Chen, Paul Meltzer, and Jeffery M.Trent. "Expression profiling using cDNA microarrays". *Nature Genetics*, 12 : 10-14, 1999

[6] Acuity Microarray Analysis, Visualization and Database software, http://www.moleculardevices.com/

[7] Lao H. Saal et al., "Bioarray software environment: A platform for comprehensive management and analysis of microarray data. *Genome Biology*, 8, 2002

[8] Jason Goncalves, Wojciech L.Marks, and Iobion Informatics LLC. "Roles and Requirements for a Research Microarray Database", *IEEE Eng. Med. Biol. Mag.*, 2002

[9]Friend-to-friend. http://en.wikipedia.org/wiki/Friend-to-friend