마이크로어레이 데이터 공유를 위한 분산 LIMS 개발

김혜정° 조환규
부산대학교 컴퓨터 공학과, ALGORIGENE Laboratory
{hjkim°, adagio}@pearl.cs.pusan.ac.kr

Development of Distributed LIMS for Microarray Data Share

Hye-Jung Kim°, Hwan-Gue Cho
Dept. of Computer Engineering, Pusan National University

요 약

마이크로어레이 기술로 인해 많은 양의 유전자 발현 데이터가 생겨났다. 데이터의 양이 많아짐에 따라 이를 체계적으로 관리해야 할 필요가 생겼고 이를 위해 LIMS(Laboratory Information Management System)가 만들어졌다. LIMS는 실험 데이터를 체계적으로 관리할 뿐만 아니라 같은 실험실에 있는 연구자 간에 데이터를 공유할 수 있는 장점을 갖고 있다. 만일 여러 연구실 간에 공동연구를 한다면 데이터의 공유가 필요하게 된다. 이를 좀 더 편리하게 하기 위해 다른 연구실의 LIMS에 저장된 데이터 중에 허용된 데이터에 한해서 자신의 LIMS에서 바로 접근할 수 있도록 하는 분산 LIMS를 소개하고자 한다.

1. 서 론

마이크로어레이 기술로 인해 많은 양의 유전자 발현 데이터가 나오게 되었고 이러한 데이터를 이용해서 많은 연구가 이루어지고 있다. 이렇게 데이터가 많아짐에 따라 이를 효율적으로 저장하여 관리할 수 있는 시스템이 필요하게 되었고 이러한 역할을 하는 것이 LIMS(Laboratory Information Management System)이다. 마이크로어레이 실험을 위한 LIMS에는 실험에서 얻어진 이미지 뿐만 아니라 이를 분석한 분석 파일 및 실험의 기본 정보가 저장된다. 이를 통해 연구실 내부의 연구자 간에 데이터를 공유할 수 있어서 불필요한 중복 실험을 방지할 수 있는 장점이 있다. 하지만 이러한 연구가 연구실 간에 공동으로 이루어지는 경우에 서로 간의 데이터 공유가 필요하다. 이를 위해서는 LIMS에 저장된 데이터를 메일 등을 통해 다른 연구실에 전해주어야 한다든지 이렇게 한다하더라도 서로 간의 포맷이 달라 데이터를 분석하기 힘든 불편함이 있다. 만일 자신의 LIMS에서 다른 연구실의 LIMS에 바로 접근하여 통일된 포맷의 데이터를 교환하는 것이 가능하다면 훨씬 편리하게 공유할 수 있을 것이다.

본 논문에서는 연구실 내의 데이터 관리에 사용되는 LIMS와 LIMS에 저장된 데이터를 실험실 간에 공유하기 위한 분산 LIMS에 대해 소개하고자 한다. 데이터를 공유하기 위한 실험실들에서 사용하는 하드웨어나 분석 프로그램 등은 모두 다르므로 분산 LIMS에서는 업로드한 후 파일 포맷을 통일하도록 하였다.

2. 이전 연구

2.1 SOA(Service Oriented Architecture)

SOA는 네트워크 상에서 어떤 기능을 하는 서비스 노드가 그 기능이 필요한 서비스 노드를 대신해 기능을 수행하도록 하는 방식이다. 이렇게 제공되는 기능을 "서비스"라고 하며, 이 서비스 상호작용들은 서술 언어를 사용하여 정의된다. 이러한 SOA의 구조는 그림 1과 같다. [1]

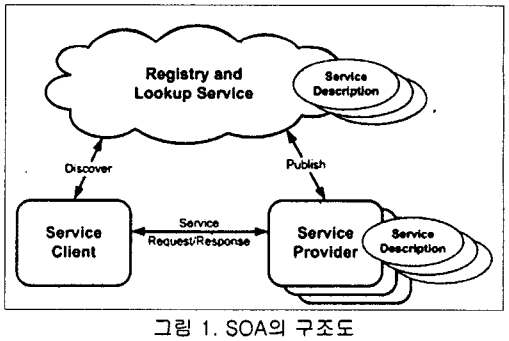

그림 1. SOA의 구조도

그림 1에서 보듯이 SOA는 등록 및 검색 서비스(Registry and Lookup Service), 클라이언트(Service Client), 서비스 제공자(Service Provider)로 구성된다. 일단 서비스 제공자가 자신이 제공하는 서비스에 대한 설명(Service Description)을 등록 및 검색 서비스에 등록(Publish)하면 클라이언트는 등록 및 검색 서비스에서 자신이 원하는 서비스를 검색한다. 검색을 통해 자신이 원하는 서비스를 제공하는 제공자를 발견(Discover)하게 되면 서비스 제공자에게 직접 서비스에 대한 요청(Request)을 하게 되고 서비스 제공자는 그에 대해 응답(Response)을 함으로서 서비스의 상호작용이 이뤄지게 된다.

이러한 SOA의 구조는 마이크로어레이 데이터를 공유하기 위한 분산 LIMS에 적용할 수 있다. 하지만 이러한 데이터를 공유하는 것은 데이터의 특성상 유연성 및 편리보다는 데이터에 대한 보안에 더욱 신경을 쓰게 된다. 그러므로 중앙에 관리 계층이 존재해서 임의의 LIMS에 데이터를 공개하는 것보다는 자신이 허용하는 대상에게만 공개하는 것이 더 좋다. 이를 위해 SOA에서의 클라이언트와 서비스 제공자의 역할을 하지만 서로의 존재를 알고 있어서 중앙의 관리 계층이 필요 없는 분산 시스템이 필요하다.

2.2 Friend-to-friend

Friend-to-friend(혹은 F2F P2P)는 P2P의 특정한 형태로 P2P가 서로 잘 모르는 사람끼리 직접 자신의 파일을 공유할 수 있도록 하는 반면 이것은 "친구(friend)" 관계인 사람에게만 직접 연결을 허용해서 파일을 공유할 수 있도록 하는 것이다. 즉, F2F 소프트웨어는 IP 주소나 디지털 서명을 이용해서 자신이 신뢰하는 사람만 자신의 컴퓨터에 직접 연결해서 파일을 교환할 수 있도록 허용한다. 그러면 자신의 친구의 친구는 간접적으로 컴퓨터에 접속해서 파일을 교환할 수도 있다.

이러한 F2F는 모르는 사람이 나의 컴퓨터에 접속해서 중요한 파일을 받아가는 것을 방지할 수 있고 "친구"의 IP 주소를 모두 안다면 방화벽을 사용해서 F2F 포트로 접속하려는 친구의 주소를 제외한 모든 주소를 차단할 수 있다. F2F 소프트웨어는 데이터 링크 레벨에서 암호화하는 방식을 사용하기 때문에 내 컴퓨터에서 어떤 파일이 교환되고 있는지를 비밀 키를 사용해서 제어할 수 있다. 만일 그 파일의 교환을 원하지 않는 경우에는 친구의 공개키를 제거하거나 방화벽을 이용해서 연결을 끊을 수도 있다. 또한 친구에게만 접속을 허용하기 때문에 보안상의 문제도 훨씬 적다. [2] 이러한 개념을 사용하면 특정 데이터에 대해 아무에게나 공개를 하지 않고도 "친구" 관계인 사람에게만 데이터를 공개하도록 할 수 있다.

3. 마이크로어레이 이미지 및 분석 데이터 관리 시스템

분산 LIMS에서 데이터를 교환하기 위해서 standalone으로 동작하는 LIMS가 필요하다. 분산 LIMS는 이러한 LIMS들간에 상호작용하면서 데이터를 교환할 수 있도록 하기 위한 시스템이다. 이러한 시스템을 구성하는 LIMS의 종류가 여러 가지인 경우에는 각각의 LIMS가 사용하는 데이터 포맷이나 데이터베이스 시스템 등이 다르다면 데이터의 교환에 있어서 포맷 변환과 같은 처리를 해주어야 하는 번거로움이 생길 뿐만 아니라 이를 위해 부가적인 시간이 든다는 단점이 있다. 이를 방지하려면 LIMS를 통일하면 된다. 이를 위해서는 실험 데이터를 체계적이고 계층적으로 관리하여 데이터 교환의 목적 뿐만 아니라 데이터 관리의 목적에서도 사용자의 욕구를 충족할 수 있는 LIMS가 필요하다.

본 논문에서 소개하는 시스템의 이름은 SMILE(Small and solid MIcroarray Lims for Experimentors)로 웹 기반의 LIMS이다. Linux에 웹 서버를 설치하여 웹 서비스를 제공하며 Windows와 Linux 사용자 모두 익스플로러나 넷스케이프와 같은 프로그램을 사용하여 이를 사용할 수 있다. 본 LIMS에 저장된 데이터는 Project, Experiment, Work 단위로 계층적인 관리가 이루어진다. 이러한 계층은 실험의 진행 상황에 따라 나눠지는데 Project는 연구의 단위가 되며, Experiment는 Project 내에서 실험 대상에 따라 분류되는 단위이다. Work는 Experiment에서 정한 실험 대상에 대해 실험 조건에 따라 분류되는 단위이다. 예를 들어, Project가 식물의 생장에 대한 하나의 프로젝트라고 하면 Experiment는 애기장대와 같은 실험 대상을 설정하고 Work에서는 물의 양과 같은 실험 조건을 선택한다. Work에서는 물의 양이 10cc, 20cc와 같이 변함에 따라 실험한 각 이미지를 등록할 수 있다.

본 시스템에서 제공되는 기능은 다음과 같다.

① 데이터 저장 및 관리 : 마이크로 어레이 실험에 대한 정보와 스캔 이미지, 해당 이미지의 분석 파일 등을 저장하고, 이를 수정, 삭제할 수 있는 기능을 제공한다.
② 검색 : Project, Experiment, Work에 저장된 실험 정보에서 키워드 검색을 제공한다.

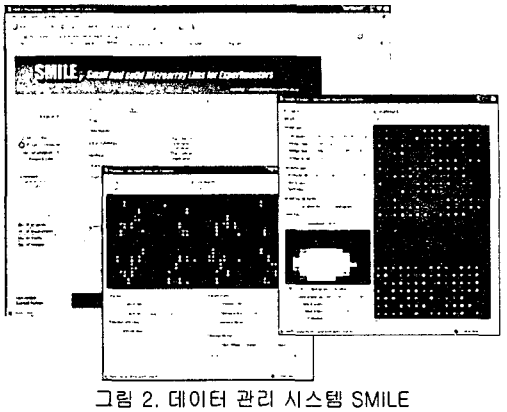

그림 2. 데이터 관리 시스템 SMILE

③ 백업 및 복구 : Project 단위별로 백업을 해 둘 수 있으며 데이터가 손실되었을 경우, 복구 기능을 사용하여 원래 형태로 복구할 수 있다.
④ 메타 파일 생성 : 여러 개의 분석 파일로부터 사용자가 원하는 정보만을 골라서 새로운 분석 파일을 생성하여 한 눈에 비교할 수 있는 기능을 제공한다.
⑤ 외부 데이터베이스와의 연결 : 유전자 이름으로 Entrez, GenBank, PubMed, GeneCard 등의 외부 데이터베이스에서 검색을 할 수 있도록 하는 기능을 제공한다.
⑥ 이미지 보기 : 등록된 이미지를 보여주고, 분석 파일에서 제공된 정보를 토대로 이미지에서 스팟 클릭시 스팟 이미지와 그 정보를 보여준다. 또한 선택한 스팟의 픽셀 강도 값에 따른 3D histogram을 보여주고 전체 스팟의 강도에 따른 scatter plot과 histogram을 보여준다. 이러한 스팟은 여러 개를 선택할 수도 있다.
⑦ HeatMap : 반복 실험한 여러 분석 파일로부터 실험이 얼마나 잘 되었는지 색으로 보여주고 스팟 선택시 그 스팟 정보를 보여주는 기능을 제공한다.
⑧ 분석 파일 정규화 : 등록된 이미지 분석 파일 각각에 대해서 싱글 정규화를 제공한다. 또한 반복 실험한 여러 분석 파일로부터 멀티 정규화를 제공한다.
⑨ 메모장과 프로젝트 스케줄러 : 각 Project 단위별로 속한 연구원과 그 관리자 사이에 연구의 진행상황 보고 및 관리를 위한 메모장과 프로젝트 스케줄러를 제공한다.
⑩ 유전자 조절 네트워크 분석 시스템 : 분석 파일의 일부 데이터를 이용하여 유전자 조절 네트워크를 구성할 수 있는 시스템을 Java Web Start를 이용해서 제공한다.

대표적인 마이크로어레이 LIMS인 BASE[3], Argus[4]와 비교하면 이미지 분석 기능이 부족하지만 BASE와 Argus에서는 제공하지 않는 백업과 복구 기능, 메타 파일 처리 기능, 유전자 조절 네트워크 분석 기능이 제공된다는 차이점이 있다.

4. 분산 LIMS의 데이터 교환 과정

마이크로어레이 실험에서 얻어진 데이터는 부가가치가 높기 때문에 아무에게나 공개되는 것을 꺼려진다. 이러한 연구에서 여러 연구실 간에 공동 연구를 하는 경우 데이터를 교환하기 위해서는 신뢰할 수 있는 대상에게만 데이터를 교환하는 것이 필요하다. 또한 마이크로어레이 실험에 사용되는 하드웨어와 소프트웨어는 실

험실마다 종류가 달라서 포맷이 맞지 않으면 공유를 하더라도 데이터에 대한 분석이 어려워지는 단점이 있다. 이를 위해 신뢰할 수 있는 대상을 "친구"로 등록하여 이러한 "친구"를 맺은 LIMS끼리 데이터를 주고받도록 하고 마이크로어레이 데이터에 대해서 통일된 포맷을 이용하도록 하는 분산 LIMS를 구현하였다. 이러한 분산 LIMS의 구조는 그림 3과 같다. 그림 3에서 보듯이 마이크로어레이 데이터 공유를 위한 분산 LIMS는 데이터를 제공하는 Service Provider와 데이터를 제공받는 Client로 구성된다. 이러한 역할을 하는 LIMS가 여러 개가 모여서 여러 실험실 간에 데이터 공유가 이뤄지게 된다. 이들 각각은 하나의 실험실에서 사용되는 LIMS이며 데이터 공유를 위한 통일된 포맷을 제공하기 위해 자신의 실험실에서 얻어진 데이터를 업로드할 때 통일된 포맷으로 변환되어 업로드된다.

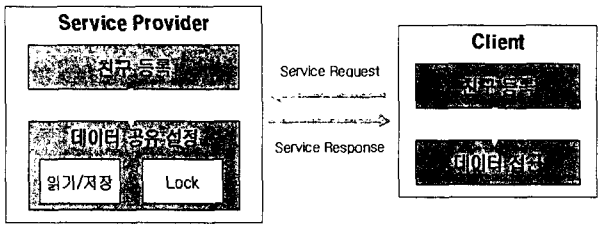

그림 3. 분산 LIMS의구조도

각 LIMS에서 다른 LIMS로의 접근은 자신의 LIMS 내부에서 이루어지며 자신에게 공개된 데이터에 대해서 접근하는 방식이 자신에게 저장된 데이터에 대해서 접근하는 방식과 같도록 transparent하게 구현되었다. 접근하는 방식은 같지만 데이터에 대해서 허용되는 권한은 Service Provider에서 지정한 허용되는 범위 안에서만 데이터 접근이 가능하도록 구현되었다.

분산 LIMS에서 기본 LIMS로 사용되는 SMILE의 기본적인 데이터 접근 권한은 public, protected, private, 개인 사용자로 구분되고 사용자 등급은 시스템 관리자, 프로젝트 관리자, 실험자, 손님으로 구분된다. Public은 모든 사용자 등급에게 데이터를 공개하는 것이고 protected는 시스템 관리자, 프로젝트 관리자, 실험자에게 데이터를 공개하는 것이다. Private은 하나의 프로젝트를 같이 수행하는 사람들에게만 데이터를 공개하는 것이고 개인 사용자 권한은 그 데이터를 등록한 사용자만 데이터에 접근이 허용되는 것이다. 이러한 기본적인 데이터 접근 권한에 덧붙여 "친구"가 데이터를 접근하려고 할 때의 설정은 처음에 설정을 하지 않았을 경우 기본적으로 public인 데이터에 대해서는 접근이 가능하다. 하지만 그 이외의 데이터에 대해서는 그 데이터를 등록한 사람이 특정 "친구"가 접근 가능하도록 설정할 수 있다. 이러한 "친구"LIMS 내부에는 어느 특정 사용자나 사용자 집단에게만 공개할 것인지를 설정할 수 있다.

데이터에 대해 "친구"가 행할 수 있는 권한은 "읽기"와 "저장"이 있다. "읽기"의 경우 데이터에 대해서 읽기만 가능하도록 하는 것이고 "저장"의 경우 데이터를 읽을 수 있을 뿐만 아니라 "친구"의 LIMS에 접근하지 않아도 볼 수 있도록 자신의 LIMS에 저장도 허용하는 기능이다. "저장"을 허용할 경우 consistency와 security 측면에서 문제점이 생긴다. Consistency의 경우, 데이터에 접근시마다 원본 데이터가 업데이트 되었는지를 항상 체크해서 데이터를 업데이트해야하고 자신의 LIMS에 저장한 데이터가 업데이트 되었다면 원본 데이터가 업데이트 되었다하더라도 업데이트를 하지 말도록 해야 한다. Security의 경우, 서비스 제공자가 자신의 "친구"인 클라이언트에게만 데이터를 공유하고 싶었는데 클라이언트가 이를 자신의 LIMS에 저장하면 그 클라이언트의 "친구"가 이 데이터를 받아갈 수 있는 문제이다. 이를 방지하기 위해 원본 데이터를 가진 서비스 제공자가 이 데이터에 대해 lock을 걸도록 함으로서 "친구"의 "친구"가 이 데이터에 접근하지 못하도록 할 수 있다. 또한 클라이언트에게 저장하도록 허용한 데이터에 대해 lock을 걸지 않는다면 "친구"의 "친구"가 이 데이터에 대해 "저장" 권한을 갖더라도 원본 데이터에 대한 업데이트는 "친구"를 통해서 받을 수 있도록 제한한다.

5. 결론

지금까지 마이크로어레이 데이터 저장을 위한 LIMS와 그를 기반으로 하는 분산 LIMS에 대해 소개하였다. 분산 LIMS에서 각각의 LIMS가 동일한 시스템이면 서로 간의 데이터 공유를 할 때 데이터 포맷 변환 등의 번거로운 작업이 없어진다. 이를 위한 LIMS로 데이터의 계층적 관리와 이미지 보기, 정규화 등이 가능한 SMILE을 소개하였다. 공동 연구를 할 때 데이터를 공유하기 위해서는 LIMS에 저장된 데이터를 바로 공유하는 것이 편리하고 이러한 데이터 간에는 보안이 필요하다. 이를 위해 신뢰할 수 있는 대상 간에 데이터를 공유할 수 있도록 "친구" 관계를 맺은 것끼리만 특정 데이터에 대해 권한을 부여하여 사용할 수 있도록 하는 분산 LIMS를 소개하였다. 이러한 분산 LIMS는 대부분의 분산 시스템이 갖고 있는 문제점인 consistency에 대해서 "친구"에게서 받은 데이터를 따로 관리하여 업데이트시 바로 업데이트한다든지 여기서 변환되었다면 원본 데이터의 업데이트 여부만 알려주도록 한다. 또한 "친구"의 "친구"에게 데이터가 공개될 수 있는 것을 막음으로서 훨씬 더 자동화되고 보안된 방식으로 데이터 공유가 가능하리라 기대된다.

본 논문에서 소개한 LIMS인 SMILE은 아래의 주소에서 사용할 수 있다.

시스템 홈페이지 : http://neobio.cs.pusan.ac.kr:8080/smile
관련 웹페이지 : http://pearl.cs.pusan.ac.kr/~smile

6. 참고 문헌

[1]SOA(Service-Oriented Architecture) : 서비스 지향구조. http://terms.co.kr/SOA.htm

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

[3]Lao H. Saal, Carl Troein, *et al.*, "Bioarray software environment(base): a platform for comprehensive management and analysis of microarray data. *Genome biology*, 2002.

[4]Jason Comander, Griffin M. Weber, *et al.*, "Argus-a new database system for web-based analysis of multiple microarray data sets.", *Genome Research*, 11(9):1603-1610, 2001