# 유전자 조절 네트워크 분석을 위한 통합 시스템 개발

이경신°, 조환규, 박선희
부산대학교 컴퓨터 공학과, ALGORIGENE Laboratory
{kslee°, adagio}@pearl.cs.pusan.ac.kr, shp@etri.re.kr

# Development of an Integrated System for Genetic Regulatory Network Analysis

Kyung-Shin Lee°, Hwan-Gue Cho, Seon-Hee Park
Dept. of Computer Engineering, Pusan National University
Bioinformatics team, Electronics and Telecommunications Research Institute

**요 약**

마이크로 어레이 기술로 인해서 유전자의 발현 데이터를 대량으로 얻을 수 있게 되었다. 따라서 실험 조건에 따른 유전자 발현 양상을 한눈에 볼 수 있게 되었고, 이를 기반으로 유전자간의 조절 관계를 예측할 수 있게 되었다. 또한 실험 이미지와 분석 파일들이 많아짐에 따라서 이러한 데이터를 효율적으로 관리하고, 저장하는 시스템이 필요하게 되었다. 이 두 가지 시스템을 통합함으로써 유전자 조절 네트워크 분석에 필요한 발현 데이터를 체계적으로 관리하고 손쉽게 얻을 수 있을 뿐만 아니라 분석 결과 또한 효율적으로 관리할 수 있다. 본 논문에서는 유전자 네트워크 분석 시스템과 마이크로 이미지 및 분석 데이터 관리 시스템을 통합한 시스템을 소개하고 각 시스템에서 제공하는 기능과 통합 시스템의 특징에 대해서 소개한다.

## 1. 서 론

유전자의 기능을 알아내는 것은 생물학의 최종 목표라고 할 수 있다. 유전자들이 발현되기 위해서는 그 유전자가 가지고 있는 기본 염기서열도 중요하지만, 그 유전자 자체만이 아닌 다른 여러 가지 환경적인 요소가 많이 작용한다. 그 요소들 중 하나가 그 유전자의 발현에 영향을 미치는 다른 유전자들과의 관계이다. 유전자는 상호 복합적으로 작용을 하기 때문에 유전자들의 조절 관계를 분석하는 것은 유전자의 기능을 예측하는데 필수적이다.

마이크로 어레이 기술이 발달함에 따라서 대량의 발현 데이터를 얻을 수 있게 되었다. 관심 있는 유전자들의 time series 데이터를 이용하여 각 유전자들의 발현 양상을 쉽게 관찰할 수 있게 되었고, 이를 기반으로 하여 유전자들 간의 조절 관계를 분석할 수 있게 되었다.

또한 데이터가 많아짐에 따라서 마이크로 어레이 실험 결과를 효율적으로 관리할 수 있는 시스템이 필요하게 되었고, 이러한 역할을 하는 것이 LIMS(Laboratory Information Management System)이다. LIMS에서는 마이크로 어레이 이미지 데이터뿐만 아니라 이미지의 분석 데이터 등을 포함하여 저장, 관리함으로써 실험자간의 데이터를 공유할 수 있고, 불필요한 중복 실험을 방지할 수 있는 장점이 있다.

본 논문에서는 유전자 조절 분석시스템과 실험 정보 및 데이터 관리 시스템을 통합한 시스템을 소개하고자 한다. 두 시스템을 통합함으로써 유전자 조절 분석에 사용되는 발현 데이터를 효율적으로 관리 이용할 수 있을 뿐만 아니라, 유전자 네트워크 분석 결과까지 함께 관리할 수 있고, 유전자 네트워크 분석을 위해 필요한 발현 데이터를 손쉽게 구할 수 있는 장점이 있다.

## 2. 유전자 조절 네트워크 분석 시스템

유전자는 하나 이상의 activator, inhibitor를 가지고 있다. Activator는 유전자의 발현을 위한 신호로써 activator가 없으면 낮은 값의 발현 상태만을 나타낸다. Inhibitor는 유전자의 발현을 억제하는 기능을 한다. 예를 들어 유전자 a가 유전자 b를 조절한다고 할 때, 유전자 a가 발현 한 뒤 b가 발현하게 되면, a가 b의 activator라고 할 수 있다. 이와 반대로 유전자 a가 발현을 한 뒤, 유전자 b가 발현 상태에서 발현하지 않은 상태로 변하게 되면 a는 b의 inhibitor라고 예상할 수 있다[1]. 이와 같은 방법으로 발현 양상을 분석하여 조절 관계를 예측할 수 있다.

본 논문에서 소개하는 유전자 조절 네트워크 시스템의 전체적인 구조는 그림1과 같다.

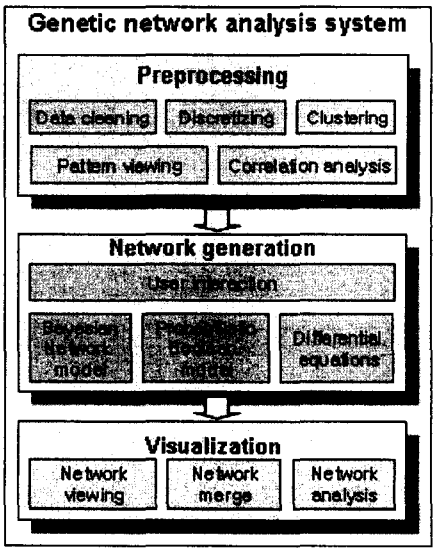


그림 1. 유전자 조절 네트워크 분석 시스템의 구조도

각 단계별로 오프라인으로 연결되어, 사용자가 반복적으로 여러 데이터를 분석하기 편리하며, 유전자 네트워크 분석을 위해서 세 가지 방법론을 제공하는 장점이 있다. 또한 가시화 단계에서는 여러 개의 결과를 합병해서 볼 수 있어 좀 더 정확도가 높은 결과를 도출해 낼 수 있다.

### 2.1. 전처리 단계

전처리 단계는 본격적으로 네트워크를 분석하기 전에 필수적을 필요한 단계로서 입력으로 들어온 데이터를 cleaning하는 기능을 제공한다. 분석하기 위한 유전자의 수가 너무 적거나 너무 많은 경우에는 받아들이지 않으며, missing value를 보정하는 기능을 제공한다. 또한 사용자가 원하는 유전자만으로 새로운 입력 데이터를 생성할 수 있는 기능을 제공함으로써 실험 목적에 따라 손쉽게 입력 파일을 생성할 수 있다. 그리고 발현 데이터를 분석하기 위해서 발현이 된 상태(over expressed), 발현이 억제된 상태(under expressed)로 분류하는 이진화와 발현이 된 상태, 변화가 없는 상태, 발현이 억제된 상태로 나누는 삼진화로 나타내는 이산화 기능이 있다.

Clustering과정에서는 $K$-Means, SOM(Self-Organizing Map), Hierarchical, Graph based clustering, 네 가지 방법을 제공하며, 선택한 유전자의 발현 패턴을 보여주는 *pattern viewing*을 통해서 발현 패턴이 유사하거나 상반되는 유전자를 알아 볼 수 있는 기능을 제공한다. 또한 correlation coefficient를 기반으로 분석하는 기능도 제공한다.

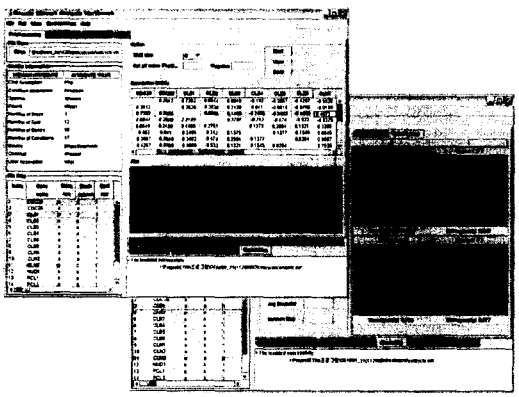

그림2. 전처리 단계 snapshot

### 2.2. 네트워크 생성 단계

네트워크 분석을 위해서는 세 가지 분석 모델, Bayesian *network model*[2], *Probabilistic boolean network model*[3], Differential equations[4]를 이용한다. 이 세 가지 모델은 유전자 조절 네트워크 분석을 위해서 이용되는 대표적인 방법이다.

첫 번째 모델인 Bayesian network model은 직접적으로 의존성이 있거나 연관성이 있는 데이터들의 관계를 확률을 기반으로 나타내는 모델이다. 특히, 부분적으로 상호연관성을 가지는 구조에 적합하여 유전자 조절 네트워크 분석에 적합한 모델로 여겨지고 있다[2].

두 번째 모델인 Probabilistic boolean network model는 Boolean network model을 개선한 방법이다. Boolean network model은 생물이 가지고 있는 불확실성(uncertainty)을 표현할 수 없기 때문에 이를 보완하기 위해서 확률 모델을 추가한 것으로 베이즈 추론과 유사한 모델이다[3].

마지막 모델인 Differential equations에서는 시간에 따른 변화량을 함수로 모델링하여 모델에 가장 적합한 매개 변수를 찾아서 유전자들 간의 조절 관계를 표현하는 방법으로 연속적인 모델을 사용하는 장점이 있다[4].

유전자의 수가 많을수록 계산양이 많아지므로 계산양이 적은 경우에는 optimal한 결과를 도출하며, 그렇지 않은 경우에는 GA(Genetic Algorithm)를 이용하여 suboptimal한 결과를 도출해 낸다[5].

신뢰성 있는 분석 결과를 얻기 위해서 위 세 가지 분석 모델을 이용하여 분석하기 전과 후 단계에 user interaction 모듈을 이용하여 사용자가 이미 알고 있는 유전자들 간의 조절 관계를 추가, 삭제할 수 있을 뿐만 아니라 하나의 유전자를 조절하는 유전자들의 수를 유전자 별로 조절할 수 있게 하였다. 즉, 유전자 조절 네트워크를 그래프로 생각하였을 때, 유전자가 그래프의 노드를 나타내면 노드의 in-degree의 수를 사용자가 조절할 수 있도록 하였다. 그러므로 사용자가 입력한 조절 관계는 변하지 않으면서 사용자의 입력에 따른 in-degree의 수를 만족하는 가장 optimal한 결과를 찾아낸다.

### 2.3. 가시화 단계

생성된 유전자 네트워크를 사용자에게 한눈에 볼 수 있는 기능을 제공한다. 네트워크 편집 기능과 유전자, 조절 관계(edge)의 속성에 대한 정보를 제공하며, 그래프 이론을 기반으로 하여 유전자 네트워크의 특성을 분석한다.

또 다른 기능으로는 같은 입력 파일을 이용하여 분석한 두 개 이상의 다른 결과 파일을 합병(merge)하여 볼 수 있다. Time series 데이터를 이용할 경우, 동일한 유전자의 집합에 대하여 각각 다른 조건에 대한 실험 결과를 이용하여 분석한 결과를 합병하면, 결과에서 중복으로 나타난 유전자간의 조절 관계는 더욱 신뢰성 있는 것으로 생각할 수 있다. 또한 실험 조건보다 유전자의 조절 관계에 의해서 발현 형태가 변화한 유전자를 찾아낼 수 있다.

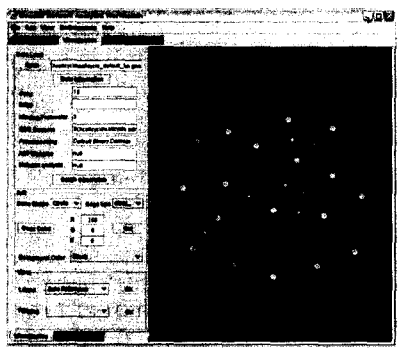

그림 3. 가시화 단계 snapshot

## 3. 이미지 및 분석 데이터 관리 시스템

생물학적으로 의미 있는 결과를 도출해 내기 위해서는 실험 데이터의 축적이 필요하다. 이를 체계적이고, 계층적으로 저장 관리하도록 하는 기능을 LIMS에서 하며, 본 논문에서 소개하는 시스

템은 웹 기반 시스템으로써 Linux, Windows에서 사용할 수 있으며, 쉽게 설치할 수 있는 장점이 있다. 데이터의 관리는 Project, Experiment, Work 단위로 계층적으로 데이터를 관리한다. Project는 연구의 단위가 되며, Experiment는 실험 조건에 따라 분류가 되며, Work는 실험 조건의 변화량에 따라 분류되는 단위이다.

제공하는 기능으로는 다음과 같은 것이 있다.

① 데이터 저장 관리 기능 : 마이크로 어레이 이미지와 실험 정보 그리고 유전자 네트워크를 포함한 해당 이미지의 분석 파일등을 저장하고, 정보를 수정, 삭제할 수 있는 기능이다.
② 검색 기능 : 실험 정보나 유전자의 이름을 이용하여 저장 단위별로 검색할 수 있다.
③ 백업 및 복구 기능 : Project 단위별로 백업을 해 둘 수 있으며 데이터가 손실되었을 경우, 복구 기능을 사용하여 원래 형태로 복구할 수 있다.
④ 메타 파일 처리 기능 : 여러 개의 분석 파일로부터 사용자가 원하는 정보만을 골라서 새로운 분석 파일을 생성할 수 있는 기능이다.
⑤ 외부 데이터베이스와의 연결 : Entrez, GenBank, PubGene 등 외부 데이터베이스와 연결기능을 제공함으로써 유전자에 대한 정보를 제공한다.
⑥ 가시화 기능 : 동록된 이미지를 보여주고, 분석 파일로부터 발현 패턴, 히스토그램 등을 보여준다.

대표적인 LIMS인 BASE[6], Argus[7]와 비교하면 이미지 분석 기능과 정규화 기능이 부족하지만 BASE와 Argus에서는 제공하지 않는 백업과 복구 기능, 메타 파일 처리 기능, 유전자 조절 네트워크 분석 기능이 제공된다는 차이점이 있다.

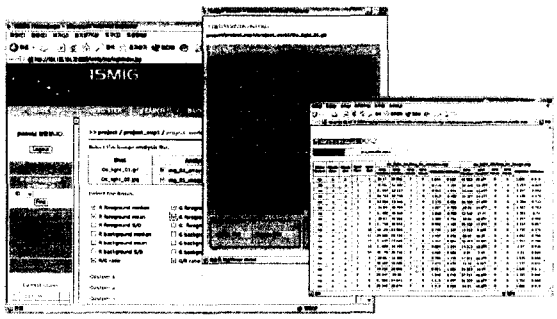


그림 4. 데이터 관리 시스템 snapshot

## 4. 유전자 네트워크 분석 시스템과 관리 시스템의 연결

데이터 관리 시스템은 웹 기반의 시스템으로써 동시에 많은 사람이 사용할 수 있다. 그러나 유전자 네트워크 분석은 많은 계산양을 필요로 하기 때문에 동시에 여러 명의 사용자가 유전자 네트워크 분석을 위해서 수행하게 될 경우, 오버헤드가 아주 크게 된다. 이러한 문제를 해결하기 위하여 JAVA web start를 이용하여 분석에 필요한 자원을 서버가 아닌 클라이언트에서 사용할 수 있도록 하였고, 유전자 조절 네트워크 분석 시스템은 데이터 관리 시스템의 서버에서 관리를 해줌으로써 사용자는 서버로부터 시스템을 수행하는 것과 같은 효과를 얻을 수 있다.

데이터의 전체 흐름도는 데이터 관리 시스템에서 저장하고 있는 이미지 분석 파일들을 메타 파일 처리 과정을 통해서 손쉽게 유전자 네트워크 분석에 필요한 입력 파일을 생성한 후, 이 데이터를 이용하여 유전자들의 조절 관계를 분석하고, 그 결과 파일을 해당 분석 파일이 위치한 곳에 저장하는 방식으로 진행된다.

## 5. 결론

지금까지 유전자 조절 네트워크 분석 시스템과 데이터 관리 시스템을 통합한 시스템을 소개하였다.

유전자 조절 네트워크 분석 시스템의 특징은 한 개의 분석 모델이 아닌 세 가지 분석 모델을 이용한 분석 방법을 제공하다는 것이며, 사용자가 직접 정보를 입력하고, 수정, 삭제할 수 있다는 특징이 있다. 마지막으로 여러 개의 결과를 합병하여 여러 결과에서 동일하게 나타나는 조절 관계를 분석함으로써 정확도 높은 분석을 할 수 있도록 제공해 준다.

데이터 관리 시스템은 웹 기반의 시스템으로써 다른 LIMS에서는 제공하지 않는 백업과 복구기능, 메타 파일처리 기능이 있으며, 유전자 조절 네트워크 분석 시스템과 연동할 수 있는 장점이 있다. 두 시스템을 결합함으로써 상호 보완적으로 장점을 보강할 수 있는 시스템을 제공할 수 있게 되었다.

본 논문에서 소개하는 통합 시스템은 아래 와 같은 주소에서 사용할 수 있다.

시스템 홈페이지 : http://164.125.164.72:8080/ismig
관련 웹 페이지 : http://pearl.cs.pusan.ac.kr/~genaw
http://garnet.cs.pusan.ac.kr/~ismig/

## 4. 참고 문헌

[1]Ting Chen, Vladimir Filkov, Steven S. Skiena, "Identifying Gene Regulatory Networks from Experimental Data", *RECOMB*, 94-103p, 1999.
[2]Nir Friedman, Michal Linial, *et al.*, "Using Bayesian Networks to Analyze Expression Data", *Journal of Computational Biology* 7: 601-620p, 2000.
[3]Ilya Shmulevich, Edward R. Dougherty, *et al*, "Probabilistic Boolean networks: a rule-based uncertainty model for gene regulatory network", *BIOINFORMATICS*, Vol.18 no.2 261-274p, 2002.
[4]Ting Chen, Hongyu L. He, George M. Church, "Modeling Gene Expression with Differential Equations", *Pacific Symposium of Biocomputing*, 1999.
[5]Hitoshi Iba, atsushi Mimura, "Inference of a gene regulatory network by means of interactive evolutionary computing", *Information Sciences*, 225-236p, 2002.
[6]Lao H. Saal, Carl Troein, *et al.*, "Bioarray software environment(base): a platform for comprehensive management and analysis of microarray data. *Genome biology*, 2002.
[7] 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