Bioinformatics and Biosystems 2006, Vol. 1, No. 1, pp. 38-45

대규모 유전자 상호작용 네트워크 추론을 위한 클라이언트-서버 시스템 구조

(Client-Server System Architecture for Inferring Large-Scale Genetic Interaction Networks)

김 영 훈[1], 이 필 현[2], 이 도 헌[1]

[1]한국과학기술원 바이오시스템학과

[2]Queens University, Dept. of Computer Science, Canada

초 록

본 논문은 베이지안 네트워크를 기반으로 대규모 유전자 상호작용 네트워크를 추론하기 위한 클라이언트-서버 시스템 구조를 제시한다. 유전체 수준(genome-wide)의 대규모 유전자 상호작용 네트워크를 베이지안 네트워크 형태로 추론하기 위해서는 병렬 서버를 이용하더라도 통상 수십시간이 소요된다. 따라서, 일반적인 대화형(interactive) 독자(standalone) 시스템 구조보다는 배치형(batch) 분산(distributed) 시스템 구조가 적합하다. 본 논문에서는 그와 같은 상황에 적합한 느슨한 연결의(loosely-coupled) 클라이언트-서버 시스템을 구현할 결과를 기술한다. 유전자 상호작용 네트워크 추론은 크게 두 단계로 나누어진다. 첫째로, 생물주석정보(biological annotation)과 유전자 발현정보(expression data)를 사용하여, 전체 유전자 집단을 서로 중복이 가능한 모듈들로 나누며, 둘째로, 각각의 모듈들에 대해 독립적인 베이지안 학습을 수행하여 추론결과를 얻고, 각 모듈들이 공통으로 포함하는 유전자를 사용하여 각 모듈의 추론결과들을 하나로 통합한다.

키워드: 베이지안 네트워크, 유전자 상호작용 네트워크, 클라이언트-서버

Abstract

We present a client-server system architecture for inferring genetic interaction networks based on Bayesian networks. It is typical to take tens of hours when genome-wide large-scale genetic interaction networks are inferred in the form of Bayesian networks. To deal with this situation, batch-style distributed system architectures are preferable to interactive standalone architectures. Thus, we have implemented a loosely coupled client-server system for network inference and user interface. The network inference consists of two stages. Firstly, the proposed method divides a whole gene set into overlapped modules, based on biological annotations and expression data together. Secondly, it infers Bayesian networks for each module, and integrates the learned subnetworks to a global network through common genes across the modules.

Keywords: Bayesian Network, Genetic Interaction Network, Client-Server

서 론

최근, 마이크로어레이 발현 데이터(microarray expression data)를 통해 유전자 상호작용 네트워크를 추론하는 방법이, 세포활동의 기작을 밝혀내는 데에 성공적으로 사용되고 있다. 이를 위해 불리안 네트워크(Boolean Network)등을 비롯한 여러 방법들이 적용되어 왔고, 그 중 베이지안 네트워크(Bayesian Network)는, 그 이론적 근거와 통계적 안정성을 바탕으로 많은 관심을 받고 있다.

본 논문에서는, 베이지안 네트워크 학습을 사용하여 마이크로어레이 발현 데이터로부터 유전자 간의 관계를 추론한다. 여기서 우리가 추론결과를 통계적으로 신뢰하기 위해서는, 충분히 많은 양의 발현 데이터가 있어야 한다. 그러나, 수백 혹은 수천 개의 유전자를 추론할 때에는, 그에 충분한 데

교신저자 : 이도헌 (Email: doheon@kaist.ac.kr)

본 논문은 과학기술부 국가지정연구실사업(2005-01450)의 지원으로 수행되었음. 연구 및 전산 시설은 정문술 바이오정보전자센터와 IBM SUR 프로그램의 도움을 받았음.

이터를 얻는 것이 현실적으로 어렵거나 거의 불가능하다. 이러한 데이터의 부족은, 결국, 기존 생물학 정보와 일치하지 않는 많은 양성 오류(false positive)를 낳게 된다. 이 문제를 완화하기 위해,기존에 알려진 생물학 지식들과 통계적 기법들을 통합하는, 몇 가지 방법들이 제안되어 왔다.

Friedman 등(2000)은 두 가지 통계적 기법(희귀후보기법 [Sparse Candidates(Friedman 등, 1999)]) 과 모델평균화기법 [model averaging])을 제안 했다. 전자는 탐색범위(search space)를 줄이기 위해, 각 유전자(child gene)에 영향을 미칠 수 있는 상위 유전자(parent gene)의 최대 수를 제한하는 방법이며, 후자는 서로 다른 초기 조건으로부터 여러 개의 네트워크를 만든 후, 이 여러 개의 네트워크로부터 공통된 관계(edge)를 뽑아내는 방법이다. 또한, 다른 연구팀들은 네트워크 구조를 가다듬기 위해, 기존에 알려진 생물학 지식들을 사용했다. Hartemink 등(2002)은 chromatin immuno- precipitation(CHIP) assay 정보를 사용했고, Tamada 등(2003)은 프로모터 서열 모티프 정보(promoter sequence motif information)를 사용했다. 최근에는, 모듈화를 통한 방법이 제안되고 있다. 이를 제안하는 연구팀들은, 군집화(clustering)을 통해 전체유전자 집단을 소집단으로 나누고, 각 집단에 대해 네트워크 학습을 진행하는 방법을 사용하였다.

본 논문에서 우리는, 생물주석정보(biological annotation information) 및 유전자 발현정보(gene expression information)라는 두 가지 상호보완적인 정보원을 통하여, 모듈화 된 유전자 네트워크를 추론하는 새로운 방법을 제안한다. 먼저, 특정한 실험조건에서만 특이하게 반응하는 유전자를 씨앗 유전자(seed gene)로 간주하고, 이 유전자들을 선택한다. 둘째로, 생물주석과 발현 데이터를 근거로 씨앗 유전자와 밀접한 유전자들을 선택하여, 중복이 가능한(overlapped) 모듈들을 만든다. 그 후, 베이지안 네트워크 학습을 각 모듈들에 대해 실시하고, 추론결과를 매개 유전자를 통해서 통합한다. 본 방법의 개요는 그림 1에서 볼 수 있다. 본 방법은 다음과 같은 가정을 바탕으로 하고 있다: '세포 조직은 지역적으로 상호작용하는, 생물학적인 모듈들로 구성되어 있으며, 대부분의 유전자들은 생물학적으로 기능이 다른 모듈 보다는, 같은 기능을 가진 모듈의 유전자들과 관련이 높을 것이다.' 본 방법은 분할해결법(divide-and-conquer)을 사용함으로써, 서브네트워크의 독립적인 형성을 가능하게 할 뿐만 아니라, 유전자 수에 대한 실험 데이터의 비율을 개선하여 학습의 성능을 향상시켜 준다. 또한 본 방법은, 서로 상호작용을 하는 유전자들이 서로 같은 생물주석을 공유하거나, 비슷한 발현 패턴을 나타낼 것이라는 가정을 포함하고 있다. 최근 Tong 등(2004)의 보고에 의하면, 유전자 상호관계는 기존에 알려진 기능적 관계와 많은 경우 일치하며, 유전자 상호작용의 12%이상이 서로 동일한 생물주석(Gene Ontology annotation)을 가진 유전자들로 구성되어 있고(우연의 경우보다 12배), 27%이상이 서로 유사한 생물주석을 가진 유전자 사이에서 일어난다는 것을 나타내고 있다.

본 방법은 모듈화를 통해, 고차원(high dimension)의 데이터로 부터 발생되는 잘못된 추론을 줄이고자 하는 면에서, 다른 모듈화 접근법들과 기본적으로 일치한다. 그러나, 본 방법은 몇 가지 특별한 면을 가지고 있다. 먼저 본 방법은, 유전자들이 다중의 세포작용이나 기능을 수행하기에, 유전자들을 단순 군집화하는 대신, 중복이 가능한 모듈들로 나누는 접근법을 사용했다. 이러한 중복된 유전자들을 '매개 유전자'라고 부르며, 이들은 독립적으로 학습된 여러 서브네트워크들을 하나의 전체 네트워크로 통합하는 역할을 한다. 둘째로, 각 모듈에서의 상세한 유전자 상호작용 관계가 먼저 추론되고, 이것이 매개 유전자를 통해 큰 전체의 그림으로 통합될 수 있기에, 본 방법을 통해 우리는, 각 모듈 내부의 상세한 유전자 상호작용 관계를 얻는 것과 동시에, 모듈 간의 유전체규모(genome-wide)의 유전자 상호작용 관계를 얻을 수 있다. 셋째로, 마이크로어레이(microarray) 발현 데이터에 비해서, 더욱 신뢰성 있고 노이즈가 적은 생물주석정보를 사용하여 발현 데이터 부족을 극복한다. 또한 생물주석은 오직 모듈생성 단계에서만 사용되고 네트워크 추론단계에서는 사용되지 않는다. 따라서 생물주석의 사용이 네트워크 학습자체를 방해하지 않으면서, 잘못된 추론의 가능성을 줄여주게 된다. 마지막으로, 본 방법은 두 가지 독립적인 정보(생물주석과 발현 데이터)의 합집합(union-set)으로 모듈들을 생성한다. 따라서 기존 생물학 지식(생물주석)이 두 유전자의 관계를 말해주지 못할 지라도, 둘의 발현 데이터가 유사하다면 같은 모듈이 될

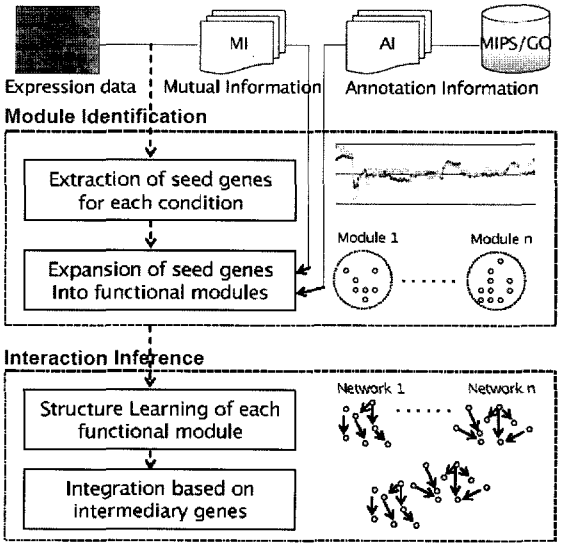

그림 1. 모듈화 된 네트워크 학습의 개요. 이것은 크게 두 부분으로 이루어져 있다. 1) 모듈 구축 부분은 전체 유전자 집단을 중복이 가능한 유전자 모듈들로 나누며, 2) 상호작용 추론 부분은 베이지안 네트워크 기술을 사용하여 유전자 간의 관계를 추론하고 이를 통합한다.

수 있으며, 이의 역도 또한 동일하게 된다. 이처럼 교집합(joint-set)대신 합집합 방법을 사용함으로써 정보의 제한을 극복하고, 아직 밝혀지지는 않았지만, 상호작용 가능성이 높은 유전자들의 관계을 놓치지 않게 된다.

우리는 본 방법론을 효모(Yeast, S. cerevisiae) 데이터에 적용하여 분석하였다. 분석 결과, 기존의 생물학 지식과 일치하는 상호관계를 잘 찾아내었을 뿐 아니라, 아직 알려지지 않은 유전자에 대한 새로운 가설을 제안할 수 있었다.

방법 및 알고리즘

1. 모듈 구축

본 방법은 두 부분으로 구성되며, 그 처음은 모듈 구축 부분이다. 씨앗유전자로부터 시작하여, 생물주석 및 발현정보가 밀접한 유전자들이 같은 그룹을 형성함으로써 모듈 구축이 이루어진다.

1.1 씨앗유전자의 선택

본 방법에서 씨앗 유전자는, 특정 실험조건에서 나머지 다른 실험조건보다 그 발현수치가 유의하게 크거나 적은 유전자라고 정의된다. Gasch 등(2000)의 논문에서 얻은 효모 실험 데이터는 16가지 실험조건에서 연속적으로 얻어진, 총 173개의 실험으로 이루어져있다. 한 실험조건 c에서 유전자 i의 특이성(Distinctiveness) D는 샤미어(Sharmir)의 측도(measure)(Sharmir, 2002)를 바탕으로 하며, 다음과 같이 정의된다.

$$D(gene_i, condition_c) = \frac{|\mu_{ci} - \mu_{\neg ci}|}{\sqrt{\frac{\sigma^2_{ci}}{n_{ci}} + \frac{\sigma^2_{\neg ci}}{n_{\neg ci}}}}$$

μ_{ci}는 실험조건 c에 속하는 실험에서 유전자 i의 평균 발현량이며, $\mu_{\neg ci}$는 c에 속하지 않는 실험에서의 유전자 i의 평균 발현량이다. 또한 σ_{ci}와 $\sigma_{\neg ci}$는 전자와 후자에 대응되는 표준편차를 의미한다. 직관적으로, μ_{ci}와 $\mu_{\neg ci}$사이의 큰 차는 유전자 i가 다른 실험조건보다 실험조건 c에서 특이적인 발현 패턴을 나타내는 것으로 볼 수 있다. 이 식을 통해, 특이성(Distinctiveness) D가 임계값을 넘는 유전자들이 씨앗 유전자로 추출된다. 여기에서 우리는 한 실험데이터 안에 있는 모든 유전자들의 D값 분포를 근거로, 상대적인 임계값을 계산하였고, $\mu_D + 3\times\sigma_D$를 그 값으로 사용하였다. 이 값은 전체 유전자의 약 5%정도로서, 씨앗 유전자의 개수를 제한하기 위해 실험적으로 선택되었다.

1.2 생물 기능 주석의 이용

씨앗 유전자와, 같은 세포 활동에 속해 있는 유전자들을 선택하기 위해서, 생물학 주석정보인 MIPS(Mewes 등, 1997) 및 Gene Ontology(The Gene Ontology Consortium, 2001)가 사용되었다. 생물 주석 정보의 적절한 사용을 위해서, 그 몇 가지 특성들에 대해 이해하는 것이 필요하다. 첫째, 생물 주

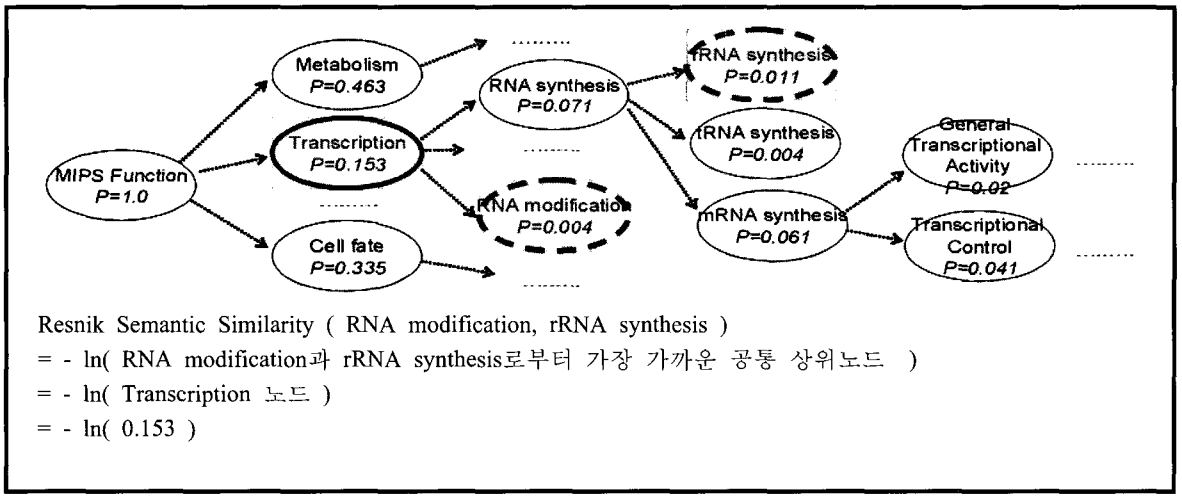

그림 2. MIPS 생물주석으로부터 만들어진 시맨틱 트리(Semantic Tree). 이 트리의 각 노드는 각각의 주석들과 대응되며, 각 노드는 정보량 P를 가지는 데 이것은 각 노드 혹은 그 하위노드들이 얼마나 많은 유전자들에 대한 주석을 제공하는가를 퍼센트 단위로 나타낸다. 두 주석의 유사성은 레스닉 측도(Resnik Measure)에 의하여 계산된다. 만일 RNA modification과 rRNA synthesis 유사성을 계산한다면 둘의 가장 가까운 공통 상위노드인 Transcription의 정보량 P를 주석정보(AI: Annotation Information) 계산식에서 사용하게 된다.

석은 계층구조를 가진다. 따라서 서로 다른 두 주석이 공통의 부모노드를 가질 수 있다. 둘째, 한 유전자가 복수 개의 주석을 가지는 것이 가능하다. 따라서 두 유전자가 같은 주석을 가지는 지 여부만을 고려하는 것이 아니라, 얼마나 많은 주석을 서로 공유 하는가가 또한 고려되어야 한다. 셋째, 생물 주석들은 서로 다른 정도의 '특정성'을 가진다. 예를 들어 Translational elongation의 기능을 갖는 GO:0006414는 309개의 효모 유전자에 대한 주석을 제공한다. 그에 반해 regulation of translational elongation의 기능을 갖는 GO:0006448은 오직 3개의 유전자에 대한 주석만을 제공한다. 따라서 생물 주석의 특징에 의해, 유전자 간의 유사 정도는 서로 공유하는 주석의 수 뿐 만 아니라 주석들의 특정성 정도에 달려 있다. Lord 등(2003)은 시맨틱 트리(semantic tree)를 통해 계층성과 특정성에 기초하여 두 생물 주석의 유사성을 계산할 수 있음을 보여주었다. 본 논문에서는 두 유전자의 유사도를 계산하기 위해 이 개념을 채택하였고, 두 유전자의 유사성의 측도로 주석정보(AI: Annotation Inform- ation)점수를 정의하였다.

첫째, 우리는 생물 주석으로부터 시맨틱 트리(semantic tree)를 만들었다. 이 트리의 각 노드는 각각의 주석들과 대응된다. 또 각 노드는 '정보량' P를 가지는 데, 이것은 자신을 포함한 그 하위노드들이 얼마나 많은 유전자들에 대한 주석을 제공하는가를 퍼센트 단위로 나타낸다. 두 주석 f_i, f_j의 유사성 점수 S는 레스닉 측도(Resnik Measure)(Resnik, 1999)에 의하여 계산되며 그 식은 다음과 같다.

$$S(f_i, f_j) = -\log(P(f_i, f_j))$$

$P(f_i, f_j)$: f_i, f_j에서 가장 가까운 공통 상위노드의 정보량

그림 2는 유사성 점수 계산의 예를 나타내며 자세한 알고리즘은 참고자료(Lord 등, 2003)에서 찾을 수 있다. 두 유전자 g_i, g_j의 주석정보(AI:Annotation Information) 점수는 이 주석들의 유사성 점수 S를 기초로 정의된다.

$$AI(g_i, g_j) = \sum_{f_k \in (AT(g_i) \cap AT(g_j))} S(f_k, f_k) + \max_{(f_i \in (AT(g_i) \cap AT^c(g_j))) \wedge (f_j \in (AT^c(g_i) \cap AT(g_j)))} S(f_i, f_j)$$

$AT(g_i)$: 유전자 i가 가지는 주석들의 집합

$AT(g_j)$: 유전자 j가 가지는 주석들의 집합

$AT(g_i)$와 $AT(g_j)$에 속한 주석들은 두 분류로 나뉠 수 있다. 하나는 두 집합이 공통으로 가지는 주석들이며, 다른 하나는 그렇지 않은 주석들이다. 만약 두 유전자가 같은 주석을 공유하면 그 주석들의 유사성 점수는 모두 더해진다. 이것은 두 유전자가 서로 여러 주석을 공유한다면 그렇지 않은 유전자 쌍 보다 더 유사한 것으로 생각할 것이라는 가정을 근거로 한다. 또한, 한 쪽 세트에만 속한 주석들은, 이들의 모든 가능한 짝의 조합 가운데 최대 유사도 S만을 AI 점수에 추가한다. 이를 통해 일부 특이하게 많은 주석을 가지는 유전자로 인한, AI 점수의 잘못된 증가를 막을 수 있다.

1.3 마이크로어레이 발현 데이터의 이용

생물 주석을 통해 선택되지는 않았지만, 씨앗 유전자와 동일한 세포 활동에 참여하는 유전자를 찾아내기 위해, 마이크로어레이 발현 데이터로부터 유전자 간의 상호정보(MI: Mutual Information(Kohane 등, 2003))을 구하는 방법이 사용되었다. 상호정보(MI)는, 한 무작위 변수가 얼마나 많은 정보를 상대에게 전달하는가를 나타낸다. 따라서, 두 유전자 발현 정보의 상호정보 점수는, 두 유전자가 마이크로어레이 발현 분포를 따라 얼마나 많은 상관도가 있는가를 나타낸다. 극단적인 예로, 만약 두 유전자의 발현 분포가 완전히 서로 독립적이라면, 상호정보 점수는 0이 될 것이다. 두 유전자 g_i, g_j의 상호정보(MI) 점수는 다음과 같이 정의된다.

$$MI(g_i, g_j) = \sum_{x_i} \sum_{x_j} p(x_i, x_j) \log \frac{p(x_i, x_j)}{p(x_i)p(x_j)}$$

x_i : 유전자 g_i의 이산화 된 발현값

x_j : 유전자 g_j의 이산화 된 발현값

1.4 씨앗 유전자를 모듈로 확장

먼저 주석정보(AI: Annotation Information)와 상호정보(MI: Mutual Information)을 기초로 씨앗 유전자와 밀접한 유전자들을 선택하며, 이들을 모듈에 포함시키는 과정을 통해 씨앗 유전자들이 각각의 모듈로 성장한다. 기본적으로, 한 씨앗 유전자는 한 모듈로 자라가기 위한 시작점이다. 그러나 여러 씨앗유전자들이 AI 및 MI의 기준값 이상으로 서로 가까이 있을 때, 이들은 하나의 모듈로 합쳐지게 되며, 이를 통해 거의 비슷한 구성유전자를 가지는 복수 개의 모듈이 생기는 것을 방지한다.

2. 네트워크 학습

2.1 각 모듈들에 해당하는 서브네트워크의 학습

각 모듈들에 대한 베이지안 학습을 수행하기 위해, 힐클라이밍(hill climbing), 희귀후보(sparse candidates), 모델평균화(model averaging)기법을 사용한 베이지안 네트워크 학습 기술을 적용되었다. 무작위로 주어지는 최초의 네트워크를 시작으로, 힐클라이밍 방법에 의해 지역 최적값(local optimum)를 찾아내고, 다시 무작위 시작과 함께 학습을 반복함으로써 주어진 데이터에 대한 최고점수를 가지는 네트워크 구조를 찾아낸다. 본 방법은 네트워크 구조에 대한 평가 함수로 MDL(Minimum Description Length) 점수를 사용하며, N개 생성된 후보 네트워크 중에서 출현빈도가 높은 관계(edge)들을

```
mark all edges in subnetworks as non-connected;
network_index=-1;

do {
    org_edge = the first non-connected edge;
    mark an org_edge as connected;
    network_index++;
    put an org_edge in a global_network[network_index];

    do {
      for( each subnetwork ) {
        for( each non-connected edge ) {
          if( edge is connected to edges
                in a global_network[network_index]) {
            mark it as connected;
            put it in a global_network[network_index];
          }
        }
    } while(no more edge is connected);

} while( no edge is non-connected );
```

그림 3. 매개 유전자를 통한 통합 알고리즘

선택하여 최종 네트워크를 구성한다. N개의 후보 네트워크 안에서 edgei의 신뢰도 점수는 다음과 같이 정의된다.

$$\text{Confidence}\ (edge_i) = \frac{\sum_{n_k \in S \wedge edge_i \in n_k} Score(n_k)}{\sum_{n_j \in S} Score(n_j)}$$

S = N개의 추론된 네트워크의 집합

신뢰도가 0.75 이상인 관계(edge)가 최종 네트워크의 기본 골격을 이루게 되고, 관계가 0.5에서 0.75사이의 신뢰도를 가지면서 그것의 한 끝이 기존 기본 골격에 이미 존재할 경우, 이것은 기본 골격에 추가된다. 각 모듈로부터 학습된 최종 네트워크를 서브네트워크라고 한다.

2.2 매개유전자를 통한 서브네트워크의 통합

서로 공통 유전자를 공유하고 있는 서브네트워크들을 합침에 의해서 서브네트워크의 통합이 이루어진다. 우리는 이러한 유전자들을 매개 유전자라고 부르며 이들은 서로 다른 세포 활동들을 연결하는 것을 의미함과 동시에 서브네트워크사이의 중계자 역할을 한다. 통합 알고리즘은 그림 3과 같다.

시스템 구조

본 방법을 사용하여, 마이크로어레이 데이터로부터 유전자간 상호작용 네트워크를 추론하는 시스템이 구현되었다. 접근의 편리성을 고려하여 본 시스템은 웹 기반 서비스를 채택하였다. 또한 효과적이고 신속한 작업처리를 위해서 슈퍼컴퓨터(P690)를 계산장치로 사용하였고, 이를 위해 삼단구조(three-tier)의 시스템을 채택하였다. 이 구조는 그림 4와 같다. 그림에서의 화살표와 그 표지는 데이터의 흐름 및 전송 프로토콜을 나타낸다. 알고리즘의 특성상 단 시간에 추론작업을 마치기 어렵기 때문에, 본 시스템은, 사용자가 작업을 요청한 후 작업 중간에 진행상황을 확인할 수 있고 작업 종료 후 그

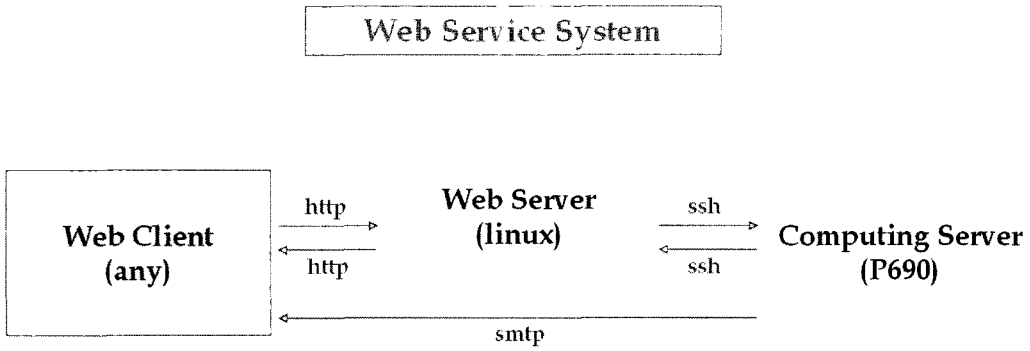

그림 4. MONET 웹 서비스 시스템 구조

결과를 메일을 통해서 제공받을 수 있도록 설계되었다.

1. 웹 클라이언트(Web Client)

사용자는 추론하기 원하는 마이크로어레이 데이터를 웹을 통해 입력하며, 필요한 파라메터를 선택한다. 작업 요청과 함께 사용자는 작업 id를 부여받게 된다. 이 작업 id를 통해 사용자는 해당하는 작업의 현재 진행 상태를 확인할 수 있으며, 그 작업을 취소할 수도 있고, 작업 완료시 그 추론결과를 볼 수 있게 된다.

2. 웹 서버(Web Server)

웹 서버는 웹 클라이언트로부터 요청된 작업을 넘겨받아서 계산서버(computing server)를 구동하여 추론작업을 수행시키는 역할을 한다. 먼저, 웹 클라이언트로부터 받은 데이터에 대한 적합성 여부를 확인하고, 이것을 파라메터와 함께 계산서버로 전달하여 추론작업을 수행시킨다. 웹 클라이언트의 요청에 따라 해당 작업의 진행상태를 확인 및 보고하며, 예상 소요시간을 계산하여 함께 제공한다. 또한 계산서버의 가용자원(프로세서) 상황을 확인 및 보고하며, 작업 수행 기록을 데이터베이스를 통해 유지한다.

3. 계산 서버(Computing Server)

계산 서버는 실제 추론작업이 수행되는 곳이며, 신속한 작업 처리를 위해 본 시스템은 슈퍼컴퓨터를 사용하였다. 또한 많은 시간이 소요되는 베이지안 네트워크 학습에서의 속도향상을 위해서 쓰레드(thread)를 사용한 병렬처리 기법이 적용되었다. 요청된 작업은 가용한 프로세서에 할당하여 수행하고, 작업 종료시 이메일을 사용자에게 발송하여 결과를 통보한다.

실험 결과

상기의 방법을 통해 얻어진 네트워크는 그림 5와 같다. 최종 네트워크에 대한 분석결과, 현재 그 기능이 알려져 있지 않은 유전자에 대한 증거 및 가설들을 얻어낼 수 있었다. 유전자 YBL010C의 경우, 현재 그 기능이 알려져 있지 않다. 그러나.이 유전자의 한쪽과 ECM10, PIM1, PUP1들이 연결되어 있는 것을 볼 때, YBL010C는 열충격(heat shock) 조건하에서, 단백질의 안정화 및 분해에 관여하는 것으로 생각된다. 추가적인 증거로서, Minddendorf 등(2004)의 실험에서 이 유전자는, USV1 낙아웃(knockout - 열충격(heat shock)과 삼투(osmolarity))조건에서 이 반응에 관여하는 것으로 나타났다. 또한 이 유전자로부터 생성된 단백질은 온도감지 성장결함(temperature sensitive growth defect)의 억제자인 SPP382p와 반응하는 것으로 알려져 있다. 또한, YBL055C 유전자 역시 그 기능이 알려져 있지 않다. 이 유전자의 한쪽과 연결된 NUP157 유전자는 핵 전달 활동에 관여한다. 또 이 유전자는 뉴클레오타이드(nucleotide)교환에 관여하는 PRP20(Bader 등, 2003)와 물리적 반응을 가진다. 이에 따라, YBL055C 유전자가 전달자의 기능 또한 가지고 있다는 사실을 추론할 수 있다. 기능이 알려지지 않은 또 다른 유전자로, YHR207C 유전자가 있다. 이 유전자는 다중 약물 전달체(multi-drug transpoter)인 FLR1 유전자와 연결되어 있으며, 핵세포질(nucleocytoplasmic) 전달체인 RIO2와 물리적인 관계를 가진다. 이를 통해 이 유전자가 약물 반응에 관여한다는 사실을 간접적으로 추론할 수 있다.

우리는 본 방법을 전체집단 접근법(a whole-set approach)및 발현정보를 통한 군집화 접근법(expression- based clustering approach)과 비교분석하였고, 그 결과를 표 1에 정리하였다.

표를 살펴보면 Case II를 통해 볼 때, 전체집단 접근법이 가장 낮은 성능을 보인다. 이에 비해 본 방법은 전체집단 방법보다 일치하는 관계(consistent edge: 두 유전자의 GO term이 서로 동일함)의 비율이 더 높을 뿐 아니라 추론된 관계(edge)의 수도 훨씬 많다. 또한 군집화 접근법은, 전체집단 방법에 비해서 좋은 성능을 나타내었지만, 본 방법론에 비해서는 좋지 않았다. 표를 통해 보는 바와 같이, 본 방법론의 성능향상은 32%에서 116%에 이르렀다. 또한, 본 방법론을 통해서 추론된 관계(edge)의 개수 자체가 군집화 방법에서의 개

표 1. 제안된 방법론과 전체집단 접근법 및 발현정보를 통한 군집화 접근법과의 비교 Case I : 최고 성능의 경우, Case II : 최저 성능의 경우 N/A : 과다한 유전자 수로 인해 추론 실패. Case I : (x_{MI}, x_{AI}) = (3,4). Case II : (x_{MI}, x_{AI}) = (5,3)

	Method	Module No.	Final Edge No.	Consistent Edge No.	Inconsistent Edge No.
Case I	The proposed method	76	328	71(21.6%)	257(69.8%)
(1612	Whole-set-based learning	1	N/A	N/A	N/A
genes)	SOM-cluster-based learning	60	94	10(10.6%)	73(77.7%)
	k-means-cluster-based learning	60	113	10(10.0%)	87(77.0%)
Case II	The proposed method	61	480	81(16.9%)	325(67.7%)
(1290	Whole-set-based learning	1	92	2(2.2%)	77(83.7%)
genes)	SOM-cluster-based learning	60	180	23(12.8%)	128(71.1%)
	k-means-cluster-based learning	60	148	19(12.8%)	107(72.3%)

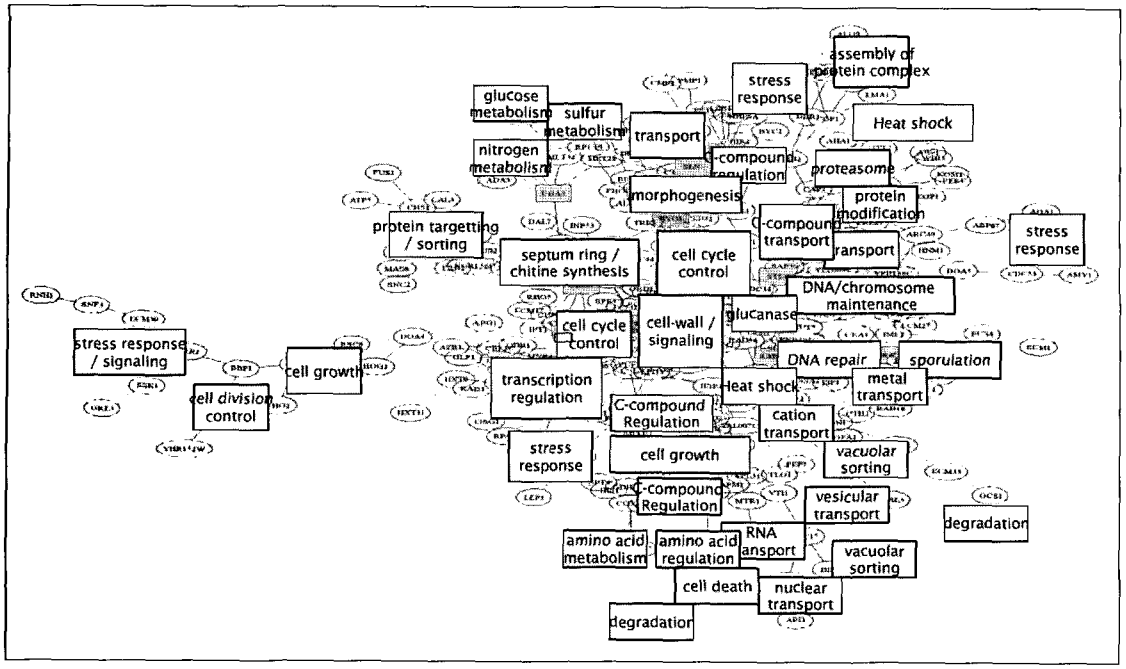

그림 5. 매개 유전자를 통해 서브네트워크들을 통합하여 얻어진, 최종 유전자 상호작용 네트워크. 각 부분에 대해 생물 주석이 표시되어 있다.

수보다 훨씬 많다는 사실을 통해서, 단순 군집화로 인해 발생되는 정보의 손실을 많이 줄일 수 있었다는 것을 확인할 수 있었다.

결 론

본 논문에서는, 유전자 수에 비해 상대적으로 부족한 실험 데이터의 문제를 경감시키기 위해, 생물 주석 정보와 발현 데이터정보를 동시에 사용하여, 유전자를 모듈화 시켜 학습을 수행하는 새로운 방법을 제안하였다. 이 방법을 통해 첫째로는 발현 데이터만을 사용하는 군집화 방법에 비해 더 좋은 성능(consistence)을 얻을 수 있다. 또한 단순한 분할대신 중복이 가능한 모듈화 방법을 사용하여, 단순 분할로 인해 생길 수 있는 정보의 손실을 줄일 수 있고 동시에, 다중의 기능을 가진 유전자를 좀 더 실제에 가깝게 표현할 수 있다. 둘째로, 모듈화를 통해 실험데이터 수와 유전자 수의 비율을 개선하여 잘못된 추론의 가능성을 경감시켰으며. 또한 여러 모듈의 공통 유전자들이 모듈들을 서로 연결하는 매개자의 역할을 하기에, 각 세포활동을 나타내는 모듈들 간의 관계(intra-relationship)를 추론할 수 있고, 더 나아가 세포 활동에 대한 유전체규모(genome-wide)의 모습을 추론할 수 있는 특징을 가진다.

향후, 본 논문에서 사용된 생물 주석 정보와 발현 데이터 이외에도, 전사인자 결합 정보 혹은 유전자 퍼터베이션(perturbation)등의 정보를 본 프레임웍에 통합하여 활용하는 연구를 진행할 계획이다.

참고문헌

[1] Akutsu,T. et al. (2000) Algorithms for inferring qualitative models of biological networks. In *Proc. of Pacific Symposium on Biocomputing*, pages 290 301.

[2] Bader,G. et al. (2003) Bind: the biomolecular interaction network database. Nucleic Acids Research,31(1):248-250.

[3] Fashing,M. et al. (2002) A clustering algorithm explicitly designed to produce priors for bayesian network discovery from whole-genome expression level data. Spring.

[4] Friedman,N. et al. (1999) Learning bayesian network structure from massive datasets: The "sparse candidate" algorithm. In *Proc. of Fifteenth Conference on Uncertainty in Artificial Intelligence*. Fifteenth Conference on Uncertainty in Artificial Intelligence.

[5] Friedman,N. et al. (2000) Using bayesian networks to

analyze expression data. *Journal of Computational Biology*, 7:601 620.

[6] Gasch et al. (2000) Genomic expression program in the response of yeast cells to environmental changes. *Molecular Biology of Cell*, 11:4241 4257.

[7] Hallinan,J. (2004) Gene duplication and hierarchical modularity in intracellular interaction networks. *BioSystems*, 74:51 62.

[8] Hartemink,A. et al. (2002) Combining location and expression data for principled discovery of genetic regulatory network models. In *Proc. of Pacific Symposium on Biocomputing*.

[9] Lam,W. et al. (1994) Learning bayesian belief networks: an approach based on the mdl principle. *Computational Intelligence*, 10:269 293.

[10] Liang,S. et al. (1998) Reveal, a general reverse engineering algorithm for inference of genetic network architectures. In *Proc. of Pacific Symposium on Biocomputing*, pages 18 29.

[11] Lord P. et al. (2003) Investigating semantic similarity measures across the gene ontology: the relationship between sequence and annotation. *Bioinformatics*, 19(10):1275 1283.

[12] Michael,J. et al. (1998) SGD: Saccharomyces genome database. *Nucleic Acid Research*, 26(1):73 79.

[13] Neapolitan,R. (2004) *Learning Bayesian Networks*. Prentice Hall

[14] Peer,D. et al. (2001) Inferring subnetworks from perturbed expression profiles. *Bioinformatics*, 17(S1):S215 S224.

[15] Resnik,P. (1999) Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *Journal of Arti-ficial Intelligence Research*, 11:95 130.

[16] Segal,E. et al. (2003) Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nature Genetics*, 34(2):166 176.

[17] Segal,E. et al. (2003) Discovering molecular pathways from protein interaction and gene expression data. *Bioinformatics*, 19(Suppl.1):i264 i272.

[18] Segal,E. et al. (2003) Genome-wide discovery of transcriptional modules from dna sequence and gene expression. *Bioinformatics*, 19(Suppl.1):i273 i282.

[19] Shamir,R. (2002) Lecture note: Analysis of gene expression data. Tel. Aviv. University.

[20] Mewes,H. et al. (1997) MIPS: A database for protein sequences, homology data and yeast genome information. *Nucleic Acid Research*, 25:28 30.

[21] Middendorf,M. et al. (2004) Predicting genetic regulatory response using classification. *Bioinformatics*, 20:i232 i240.

[22] Tamada,Y. et al. (2003) Estimating gene networks from gene expression data by combining bayesian network model with promoter element detection. *Bioinformatics*, 19(2):ii227 ii236.

[23] The Gene Ontology Consortium. (2001) Creating the gene ontology resource: design and implementation. *Genome Research*, 11(8):1425 1433.

[24] Tong,A.H.Y. et al. (2004) Global mapping of the yeast genetic interaction network. *Science*, 303:808 813.

[25] Yoo,C. et al. (2002) Discovery of causal relationships in a gene regulation pathway from a mixture of experimental and observational dna microarray data. In Proc. of *Pacific Symposium on Biocomputing*, pages 498 - 509.