베이지안망을 이용한 유전자와 약물 간 관계 분석

(Analysis of Gene-Drug Interactions Using Bayesian Networks)

오석준[1], 황규백[2], 장정호[3], 장병탁[4]

요 약

최근의 생물학 연구를 위한 기기의 자동화 및 고속화는 생물학 관련 정보량의 급증을 가져오고 있다. 예를 들어, DNA chip에서 얻어지는 마이크로어레이(microarray)는 수천 종류의 유전자의 발현량을 동시에 측정한다. 이러한 기술들은 생물의 세포나 조직에서 일어나는 일련의 다양한 현상을 전체적으로 조망하는 관점에서 관찰할 수 있는 기회를 제공하고 있으며, 이를 통한 생명공학의 전반적인 발전이 기대되고 있다. 따라서 대량의 생물학 관련 정보의 분석이나 데이터 마이닝이 행해지고 있으며 이를 위한 대표적인 기법들로는 각종 클러스터링(clustering) 및 신경망 계열의 모델 등이 있다. 본 논문에서는 확률그래프모델의 하나인 베이지안망(Bayesian network)을 생물정보분석에 이용한다. 구체적으로 유전자 발현패턴과 약물의 활성패턴 및 암 종류 사이의 확률적 관계를 모델링한다. 이러한 모델은 NCI60 dataset (http://discover.nci.nih.gov)에서 베이지안망을 학습함으로써 구성된다. 분석의 대상이 되는 데이터가 sparse하기 때문에 발생하는 어려움들을 해결하기 위한 기법들이 제시되며 학습된 모델에 대한 검증은 이미 생물학적으로 확인되어 있는 사실과의 비교를 통해 이루어진다. 학습된 베이지안망 모델은 각각의 유전자 간, 혹은 유전자와 처리된 약물 간의 실제 생물학적 관계를 다수 표현하며, 이는 제시되는 방법이 생물학적으로 유의미한 가설을 데이터 분석을 통해 효율적으로 생성하는데 유용하게 활용될 수 있음을 보인다.

1. 서 론

생명 현상의 기본이 되는 세포는 복잡 적응계(complex adaptive system)의 대표적인 예이다. 현재 생물학 분야에서는 관련 기술의 발달로 생명체에 대한 정보가 기하급수적으로 증가하고 있으며, 이를 활용하여 세포 혹은 생물체 내에서 일어나는 다양한 생화학적 반응과 각 유전자의 행동을 동시적으로 관찰하려는 노력이 활발하게 진행되고 있다. 마이크로어레이는 세포나 조직 내 여러 유전자의 동시적 발현 양상을 양적으로 측정할 수 있는 도구로 각광받고 있으며, 이는 high-throughput data analysis가 필요한 기능 유전체학(functional genomics)적 관점에서의 생명 현상 연구 및 질병 진단 등의 분야에서 특히 주목할 만한 것이다.

현재까지 마이크로어레이 데이터의 분석에는 클러스터링 기법이 주로 활용되어 왔으나, 이는 유전자의 발현 양상에 근거한 집단화 분석으로서 각각의 단일 유전자 간의 관계 분석에는 한계를 보인다. 또한, 생화학 반응등 생명 현상의 특성 상 세포 내의 여러 현상들은 계수화될 수 없는 성격을 지니고 있어 이에는 확률 모델이 적합하며, 베이지안망은 유전자의 발현 패턴을 설명할 수 있는 확률 모델을 제시할 수 있는 장점이 있다.

마이크로어레이 데이터 분석에 베이지안망을 적용할 경우에 발생하는 문제점 중 하나는 data sparseness problem이다. 특정 연구를 위해 제작되는 마이크로어레이의 슬라이드 개수는 실험 상의 시료 준비와 비용 문제로 인하여 일반적으로 수십 개에 불과하다. 반면에 한 장의

1) 서울대학교 바이오정보기술연구센터
2) 서울대학교 컴퓨터공학부
3) 서울대학교 컴퓨터공학부
4) 서울대학교 바이오정보기술연구센터, 서울대학교 컴퓨터공학부

슬라이드에는 수천에서 많게는 수만 개의 시료를 담고 있다. 이것은 마이크로어레이 데이터 분석이 수천에서 수만 개의 속성과 수십 개의 예제를 가지는 sparse data 분석이라는 것을 의미한다. 이는 베이지안망 학습에서 궁극적으로는 수천의 노드를 소화해야 하며, 확률적 추론의 어려움과 학습 결과의 낮은 신뢰도 등의 문제들을 극복해야 함을 의미한다. 따라서 본 논문에서는 이러한 사항들을 고려한, 베이지안망 기반의 마이크로어레이 데이터 분석 기법을 제시한다.

2. 마이크로어레이 분석에의 베이지안망 적용

2.1. 베이지안망

베이지안망(Bayesian networks; Jensen, 1996; Heckerman, 1999)은 다수의 확률 변수에 대한 결합 확률 분포를 변수들 사이의 조건부 독립성에 기반하여 효율적으로 표현하는 확률 그래프 모델이다. 변수 집합 $\mathbf{X} = \{X_1, ..., X_n\}$에 대한 베이지안망은 다음의 2가지 부분으로 구성된다.

(1) X의 변수들 간의 조건부 독립성을 표현하고 있는 망 구조 G
(2) 각 변수의 지역 확률 분포 집합 P

망 구조 G는 DAG(directed acyclic graph) 형태이며 각 노드는 $\mathbf{X}$의 변수들과 일대일 대응이 된다. G에서 간선으로 연결되어 있지 않은 노드들은 서로 조건부 독립 관계에 있으며, 망 구조가 표현하는 조건부 독립성에 의하면 변수 X_i는 자신의 부모 노드 집합 $\mathbf{Pa}_i$의 값이 주어진 경우, 자신의 자손이 아닌 노드들에 대해서는 조건부 독립이 된다. 이에 따른 $\mathbf{X}$의 결합 확률 분포는 수식 (1)과 같이 표현된다.

$$P(\mathbf{X}) = \prod_{i=1}^{n} P(X_i \mid \mathbf{Pa}_i). \qquad (1)$$

각 노드의 지역 확률 분포는 수식(1)의 $\prod$ 안의 각 항에 해당한다. 각 노드의 지역 확률 분포 모델은 연속형 변수의 경우 선형 가우시안 모델(linear Gaussian model), 이산 변수의 경우 다항 분포 모델(multinomial model)이 주로 이용된다 (Heckerman, 1999).

2.2. 베이지안망의 학습

베이지안망은 마이크로어레이로부터 온 수치 자료에 의하여 학습된다. 이때, 각 마이크로어레이 표본을 생성해 낸 확률 분포가 실제 존재한다고 가정한다. 그리고 각 마이크로어레이는 그 확률 분포로부터 추출된 표본이라고 가정한다. 즉, 마이크로어레이 데이터에서 베이지안망을 학습하는 것은 그러한 실제의 확률 분포를 추정하는 것이 된다.

베이지안망의 학습은 대략 두 부분으로 구성되는데, 그 하나는 고정된 베이지안망 구조에서 각 노드의 지역 확률 분포를 학습하는 것이다. 지역 확률 분포의 학습은 몇 가지 가정 아래 간단한 계산으로 해결된다 (Heckerman, 1999). 다른 하나는 망 구조의 학습이다. 구조 학습에는 점수 기반 탐색 기법이 널리 이용된다. 이 기법에서 학습 알고리즘은 주어진 데이터에 가장 적합한 망 구조를 탐색한다. 데이터에 대한 망 구조의 적합성은 점수의 형태로 측정된다. 널리 이용되는 망 구조의 점수로는 Bayesian Dirichlet(BD) 계열의 점수와 minimum description length(MDL) 기반의 점수가 있으며 본 논문의 실험에서는 BD 점수 (Heckerman et al., 1995)를 사용하였다. BD 점수는 지역 확률 분포 모델이 다항 분포 모델인 경우에 사용하며, 학습 데

이터 D와 망 구조 G에 대한 결합 확률로 다음과 같은 식으로 표현된다.

$$P(G,D) = P(G)\cdot P(D|G)$$
$$= P(G)\cdot\prod_{i=1}^{n}\prod_{j=1}^{q_i}\frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij}+N_{ij})}\cdot\prod_{k=1}^{r_i}\frac{\Gamma(\alpha_{ijk}+N_{ijk})}{\Gamma(\alpha_{ijk})} \quad (2)$$

위의 식에서 $P(G)$는 망 구조 G의 사전 확률 (prior probability)이며 $\Gamma(\cdot)$는 $\Gamma(1) = 1$, $\Gamma(x+1) = x\cdot\Gamma(x)$를 만족하는 감마 함수(gamma function)이다. n은 노드의 개수이며, q_i는 $\mathbf{Pa}_i$가 가질 수 있는 상태의 개수이다. r_i는 X_i가 가질 수 있는 상태의 개수이다. α_{ijk}는 지역 확률 분포 모델의 파라미터 분포(Dirichlet distribution)의 hyper parameter로서 파라미터 분포에 대한 사전 지식에 해당하는 부분이다. N_{ijk}는 충분 통계량으로 데이터 D에서 X_i가 $\mathbf{Pa}_i$의 j번째 상태 하에서 k번째 값을 가지는 경우의 횟수이다. α_{ij}와 N_{ij}는 다음과 같이 계산된다.

$$\alpha_{ij} \equiv \sum_{k=1}^{r_i}\alpha_{ijk} \qquad N_{ij} \equiv \sum_{k=1}^{r_i}N_{ijk} \quad (3)$$

점수 기반 탐색 기법의 핵심이 되는 부분은 탐색 전략이다. 가능한 망 구조의 개수는 노드 개수의 지수승 이상이며 가장 점수가 높은 구조를 찾는 문제는 NP-hard임이 알려져 있다 (Chickering, 1996). 따라서 greedy hill-climbing이 일반적으로 이용된다. 보편적인 greedy search 알고리즘은 다음과 같다.

* Generate the initial graph structure G_0 (usually an empty graph)
* For m = 1, 2, 3, ..., until convergence
  - Among all the possibile local changes (edge addition, edge deletion, and edge reversal) in G_{m-1}, perform the one which leads to the maximum gain in the BD score.
  - The resulting graph is G_m.

알고리즘의 수렴 조건은 Score(G_m) <= Score(G_{m-1})이다.

3. NCI60 dataset

NCI60 dataset (Scherf et al., 2000)은 미국의 국립암연구소(National Cancer Institute, NCI)에서 신약 개발 과정에 이용하기 위하여 만든 마이크로어레이 데이터이다. 이 데이터는 60 명의 암환자의 암 조직에서 추출된 세포를 배양한 60 개의 cell line 표본으로 구성되어 있다. 암의 종류는 결장암, 신장암, 난소암, 유방암, 전립선암, 폐암, 중추신경계암, 백혈병, 피부암 등 9 가지이다. 각 cell line의 유전자 발현 양상을 보기 위하여 cDNA를 이용한 마이크로어레이가 제작되었다. 이 cDNA 마이크로어레이는 9,703 개의 유전자 및 expressed sequence tag(EST)으로 구성되었다. 또한 mRNA 패턴 이외의 특성을 보기 위하여 40 개의 molecular target(단백질)의 양이 각 cell line마다 측정되었다. 유전자 발현과는 별도로 1,400 종의 약물이 각 cell line에 미치는 영향력도 측정되었다. 약물의 활력은 약물을 처리한 cell line의 48 시간 후의 세포 단백질 총량 측정에 기반한 성장 억제 정도이며 sulphorodamine B assay로 측정되었다 (Scherf et al., 2000).

실험에서 베이지안망 분석은 암의 종류에 따른 발현 양상의 변화를 강하게 보이는 1,376 개의 유전자 및 EST와 실제 치료에 이용되고 있는 항암제 118 개에 대하여 행해졌으며, 유전자와 약물 중 결측치를 4 개 이상 가지는 것과 이름이 없는 EST 등은 속성에서 제외하여, 결과적으로 890 개의 속성 (805 개의 유전자, 84 개의 약물, 암의 종류)을 가지는 60 개의 표본으로 구성된 데이터를 분석에 이용하였다.

4. 분석 결과

4.1. 데이터 차원의 축소

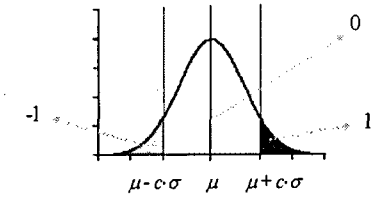
수백 개의 노드로 이루어진 베이지안망에서의 확률적 추론은 거의 불가능하다. 따라서 본 논문에서는 이를 위하여 마이크로어레이 데이터 속성의 개수를 줄여서 분석하는 방법을 택하였다. 이러한 데이터 차원 축소 방법으로는 개개의 유전자나 약물 대신 대표값을 속성으로 이용하는 방법과 전체 속성 가운데 관심있는 부분만을 선택하는 방법이 있으며, 두 가지 방법 모두 데이터의 속성에 대한 클러스터링을 통하여 이루어진다.

유전자 및 약물의 대표값을 생성하기 위하여 마이크로어레이 상에서의 각각의 데이터를 발현 및 활력 패턴의 유사도에 따라 클러스터링하였다. 그 후, 각 클러스터의 유전자 및 약물들의 평균 발현 값을 그 클러스터에 속하는 속성들의 대표값으로 설정하였다.

실험에서 데이터 차원 축소를 위한 클러스터링 방법으로는 Graepel (1998)의 soft topographic vector quantization (STVQ)을 이용하였다. STVQ는 self-organizing map (SOM)과 같은 topographic map 형태의 클러스터 구조를 제공하는 soft clustering 기법의 하나이며 EM 알고리즘 (Dempster et al., 1977)에 기반한 강건한 학습 알고리즘을 가지고 있다.

4.2. 속성값의 이산화

데이터에서 유전자의 발현 정도와 약물의 활력은 실수값이며 평균 0, 표준 편차 1로 표준화되어 있다. 이 속성값들은 베이지안망의 지역 확률 분포 모델로 다항 분포 모델을 이용하기 위하여 이산화되었으며, low (-1), normal (0), high (1)의 3 구간으로 나누었다. 이산화를 위한 경계값은 각 속성의 평균과 표준 편차에 따라 결정하였다. 구체적으로 $\mu - c \cdot \sigma$와 $\mu + c \cdot \sigma$가 경계값으로 이용되었다. 여기서 μ는 속성의 평균값이며 σ는 표준 편차이다. c는 경계값의 위치를 결정하기 위한 상수로서 0.43, 0.50, 0.60을 적용하였다. 그림 1은 각 속성값이 표준 정규 분포를 따른다는 가정 하에 low (-1), normal (0), high (1)의 c값에 따른 분포이다.



c	Distribution Ratio		
	-1	0	1
0.43	33.3%	33.3%	33.3%
0.50	30.8%	38.3%	30.8%
0.60	27.4%	45.1%	27.4%

그림 1. 이산화 경계값 결정 상수 c에 따른 각 속성 값의 분포도

4.3. 실험 결과

약물 L-asparaginase 주변의 확률적 관계를 분석하기 위해 12 개의 유전자와 4 개의 약물을 선택하였으며, 이는 유전자와 약물을 함께 클러스터링한 결과로부터 얻은 것이다. 먼저, 약물의 반응은 ATP-binding cassette, sub-family B, member 1 (ABCB1) 유전자와 같은 약물 저항 유전자에 의하여 비활성화되는 경우가 있다는 점에 착안하여 (Scherf et al., 2000), 각 약물은 60 개 시료에서의 성장 억제 값에 대한 음(negative)의 값의 벡터로 표현하였다. 이후, topographic map 형태의 클러스터링 결과에서 L-asparaginase가 속한 클러스터와 그 인접 클러스터들을 분석 대상으로 삼았다. 그림2는 이 데이터에서 학습된 17 개의 노드를 가지는 베이지안망의 일부분을 나타낸다 그림에서 암의 종류와 L-asparaginase의 활력 간의 직접적인 의

존 관계를 파악할 수 있으며 L-asparaginase와 ASNS 간의 확률적 관계도 알 수 있다. 또한 ASNS는 pyrroline-5-carboxylate reductase (P5CR) 유전자에 직접적으로 의존함도 알 수 있다.

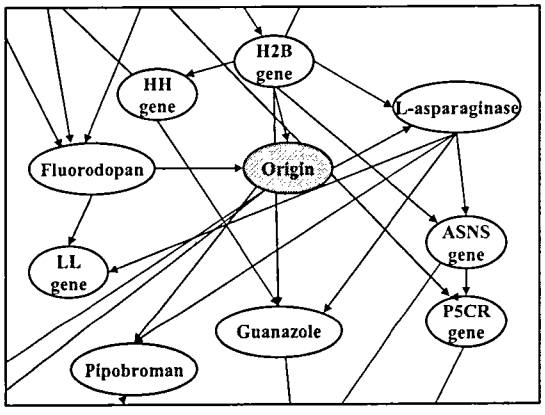


그림 2. 17 개의 노드를 가지는 베이지안망의 일부

표 1과 표 2는 이 베이지안망에서의 확률적 추론의 몇 가지 결과를 보여 준다. 표3의 조건부 확률은 ASNS와 L-asparaginase 간의 negative correlation를 의미한다. 게다가 암의 종류가 백혈병(leukemia)일 때에는 더욱 강한 negative correlation을 보인다. 이러한 사실 역시 생물학적으로 알려진 사실과 일치한다 (Scherf et al., 2000).

표 1. 그림3의 17 개의 노드를 가지는 베이지안망에서의 $P(L\text{-}asparaginase|ASNS)$에 대한 확률적 추론 결과. 괄호 안의 숫자는 cancer type이 leukemia인 경우를 나타낸다.

	L-asparaginase = low	*L-asparaginase* = normal	*L-asparaginase* = high
ASNS = low	0.19857 (0.17536)	0.27471 (0.22838)	0.52672 (0.59626)
ASNS = normal	0.31110 (0.27128)	0.49795 (0.53790)	0.19095 (0.19081)
ASNS = high	0.42159 (0.38500)	0.36279 (0.42437)	0.21561 (0.19063)

표 2. 그림3의 17 개의 노드를 가지는 베이지안망에서의 $P(L\text{-}asparaginase|P5CR)$에 대한 확률적 추론 결과

	L-asparaginase = low	*L-asparaginase* = normal	*L-asparaginase* = high
P5CR = low	0.27510	0.35226	0.37263
P5CR = normal	0.31621	0.41072	0.27307
P5CR = high	0.33837	0.39664	0.26499

표 3에서는 P5CR과 L-asparaginase 간의 negative correlation도 확인할 수 있다. 실제로, P5CR은 alanine과 aspartate metabolism에 관여하고 있는 유전자이다. 또한, ASNS는 arginine과 proline metabolism에 관여하고 있다. 이 두 metabolism은 metabolic pathway map (http://www.genome.ad.jp/kegg) 상에서 서로 상당히 인접한 위치에 있다. 따라서, L-asparaginase와의 negative correlation에 있어서 P5CR과 ASNS가 유사한 양상을 보이는 것은 생물학적으로도 의미를 가질 가능성이 크다.

5. 결 론

본 논문에서는 클러스터링등 기존의 마이크로어레이 분석 방법이 갖고 있는 한계를 극복하기 위한 여러 가지 시도의 한 예로서 베이지안망을 기반으로 한 분석 방법을 제시하고자 하였다. 특히 각각의 유전자 그리고 약물 간의 의존도를 추론하기 위하여 기존의 베이지안망을 이용한 마이크로어레이 데이터의 분석 (Friedman et al., 2000)과는 달리 확률적 추론을 적용하여 각 유전자 간 혹은 유전자와 약물 간의 관계를 정량적으로 분석하는 방법을 제시하고자 하였다. 이를 위하여 데이터값의 이산화와 데이터차원축소를 이용하였으며 실제 마이크로어레이 데이터에 적용하여 생물학적으로 유의미한 다수의 사실을 밝혀낼 수 있었다. 특히, ASNS 유전자와 P5CR 유전자의 L-asparaginase 약물에 대한 저항성이 밀접한 관련이 있다는 실험 결과는 실제 생물학 관련 연구 논문을 참조한 결과 그 가능성이 매우 높음을 알 수 있었다. 이 사실은 제시된 기법이 마이크로어레이를 이용한 생명 공학 연구에 있어서 high-throughput data 분석을 통한 가설 생성기의 역할을 할 수 있음을 보여주었다고 하겠다. 또한, 마이크로어레이 데이터를 분석하기 위하여 확률적 추론을 도입함으로써 분석 대상이 갖고 있는 생물학적 특성에 좀 더 가까운 방법을 고안해 낼 수 있는 가능성을 보였다.

앞으로 이러한 분석법에 추가되어야 사항은 데이터 차원의 축소를 행하지 않고서도 확률적 추론을 통한 정량적 분석이 가능하게 하기 위하여 대규모 베이지안망에서도 효율적인 확률적 추론을 수행할 수 있는 알고리즘이다. 또한 클러스터링 결과와 베이지안망의 구조와의 관계에 대한 보다 정밀한 분석을 통하여 적절한 데이터 차원 축소의 정도를 결정할 수 있도록 해야 하겠다. 끝으로 sparse data에서 신뢰도 있는 결과를 얻기 위하여 evolutionary Markov chain Monte Carlo (eMCMC; Zhang and Cho, 2001) 등을 기반으로 한 베이지안망 구조 학습법의 개선이 이루어져야 하겠다.

참 고 문 헌

Chickering, D. M. (1996). "Learning Bayesian networks is NP-complete," Fisher, D. and Lenz, H. -J. (eds.), *Learning from Data: Artificial Intelligence and Statistics V*, Springer-Verlag, NY, 121-130.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). "Maximum likelihood from incomplete data via the EM algorithm (with discussion)," *Journal of Royal Statistical Society B*, **39**, 1-38.

Friedman, N., Linial, M., Nachman, I., and Pe'er D. (2000). "Using Bayesian networks to analyze expression data," In *Proceedings of the 4th Annual International Conference on Computational Molecular Biology (RECOMB'00)*, 127-135.

Graepel, T. (1998). "Self-organizing maps: generalization and new optimization techniques," *Neurocomputing*, **21**, 173-190.

Heckerman, D. (1999). "A tutorial on learning with Bayesian networks", Jordan, M. I. (edt.), *Learning in Graphical Models*, MIT Press, MA, 301-354.

Heckerman, D., Geiger, D., and Chickering, D. M. (1995). "Learning Bayesian networks: the combination of knowledge and statistical data," *Machine Learning*, **20**, 197-243.

Jensen, F. V. (1996), *An Introduction to Bayesian Networks*, University College London Press.

Scherf, U. et al. (2000). "A gene expression database for the molecular pharmacology of cancer," *Nature Genetics* **24**, 236-244.

Zhang, B. -T. and Cho, D. -Y. (2001). "System identification using evolutionary Markov chain Monte Carlo," *Journal of Systems Architecture* **47**, 589-599.