Pharmacogenomics를 위한 대규모 베이지안 유전자망 학습

황규백 장병탁

서울대학교 컴퓨터공학부

kbhwang@bi.snu.ac.kr btzhang@cse.snu.ac.kr

Large-Scale Bayesian Genetic Network Learning for Pharmacogenomics

Kyu-Baek Hwang and Byoung-Tak Zhang

School of Computer Science and Engineering, Seoul National University

요 약

Pharmacogenomics는 개인의 유전적 성향과 약물에 대한 반응간의 관계에 대해 연구하는 학문이다. 이를 위해 DNA microarray 데이터를 비롯한 대량의 생물학 데이터가 구축되고 있으며 이러한 대규모 데이터를 분석하기 위해서 기계학습과 데이터 마이닝의 여러 기법들이 이용되고 있다. 본 논문에서는 pharmacogenomics를 위한 생물학 데이터의 효율적인 분석 수단으로 베이지안망(Bayesian network)을 제시한다. 베이지안망은 다수의 변수들간의 확률적 관계를 표현하는 확률그래프모델(probabilistic graphical model)로 유전자 발현과 약물 반응 사이의 확률적 의존 관계를 분석하는데 적합하다. NCI60 cell lines dataset으로부터 학습된 베이지안 유전자망(Bayesian genetic network)이 나타내는 관계는 생물학적 실험을 통해 검증된 실제 관계들을 다수 포함하며, 이는 베이지안 유전자망 분석을 통해 개략적인 유전자-유전자, 약물-약물, 유전자-약물 관계를 효율적으로 파악할 수 있음을 나타낸다.

1. 서론

Pharmacogenomics는 pharmacology와 genomics의 합성어로 개개인의 유전적 성향과 약물에 대한 반응간의 관계를 연구하는 학문이다. 이에 대한 연구를 위해 DNA microarray 데이터, 약물 반응(drug activity) 데이터 등이 대량으로 만들어지고 있으며, 이러한 대량의 생물학 데이터를 효율적으로 분석하는 기법에 대한 연구가 생물정보학(bioinformatics) 분야에서 활발하다. 특히, 기존의 기계학습이나 데이터 마이닝 기법들이 적용되고 있으며 그 중 대표적인 것은 클러스터링으로 [2], [10] 등이 그 예이다. 한편, 유전자망(genetic network)은 유전자와 그 파생물들 사이의 관계(regulatory interaction)를 망 구조(network structure)로 구성하여 분석하려는 시도이다 [12]. 유전자망의 구성에 사용되는 모델로는 여러 가지가 있으며 베이지안망(Bayesian network)에 기반한 유전자망의 구성도 연구되고 있다 [5]. 베이지안망은 다수의 확률 변수(random variable)들간의 의존 관계(dependency relationship)를 DAG(directed acyclic graph) 형태로 표현하는 확률그래프모델(probabilistic graphical model)로 여러 유전자들의 발현간의 의존 관계를 표현하는데 적합하다.

본 논문에서는 베이지안망을 이용한 유전자망의 효율적인 구축 방법 및 pharmacogenomics에의 적용 결과에 대해서 기술한다. 유전자망을 구축하기 위해서 사용된 데이터는 NCI60 cell lines dataset [7]이다. 이 데이터는 9가지 종류의 암 조직에서 기원한 60개의 cell line에 대한 cDNA microarray 데이터, 40개의 molecular marker에 대한 측정치, 그리고 118개의 약물(drug)에 대한 반응정도(activity)에 대한 데이터이다. 이 데이터를 이용하여 다수의 유전자와 약물을 노드로 가지는 베이지안망 형태의 유전자망(genetic network)을 구성했으며 이 베이지안 유전자망(Bayesian genetic network)은 기존에 알려진 유전자-유전자 관계, 약물-약물 관계, 유전자-약물의 관계를 포함하는 여러 가지 관계들을 나타냈다.

2. 베이지안망 (Bayesian Networks)

베이지안망은 DAG(directed acyclic graph) 구조를 가지며 변수들간의 결합확률분포(joint probability distribution)를 효율적으로 표현하는 확률그래프모델이다. 변수 집합 $\mathbf{X} = \{X_1, \ldots, X_n\}$에 대한 베이지안망은 다음의 2가지 부분으로 구성된다.

(1) **X**의 변수들간의 조건부독립성(conditional independence assertion)을 표현하고 있는 망 구조 B

(2) 각 변수들의 지역확률분포(local probability distribution) 집합 P

망 구조 B는 DAG 형태이며 각 노드는 **X**의 변수들과 일대일대응이 된다. $\mathbf{Pa}_i$는 그래프 B에서 X_i의 부모노드의 집합을 나타낸다. B에서 간선으로 연결되지 않은 노드들은 서로 조건부독립관계에 있으며 망 구조가 나타내는 조건부독립성에 의하면 주어진 구조 B에서 **X**의 결합확률분포는 수식 (1)과 같이 표현된다.

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

각 노드(변수)의 지역확률분포는 수식 (1)의 Π 안의 각 항에 해당한다. 베이지안망 (B, P)가 주어지면 원하는 확률을 추론할 수 있다.

3. 베이지안망에 기반한 유전자망의 구성

3.1 NCI60 Cell Lines Dataset

NCI 60 cell lines dataset은 미국의 NCI(national cancer institute)에서 신약개발과정에 이용하기 위해 구성한 데이터로 각 암세포의 유전자 발현 패턴과 특정 약에 대한 반응과의 상관관계를 밝히기 위해 만들어졌으며 [9], 폐암, 백혈병, 신장암, 피부암 등 9가지 종류의 암세포 60

개를 배양한 cell line에 대한 cDNA microarray 데이터, 40개의 molecular marker에 대한 측정치, 118개의 약물(drug)의 반응정도(activity)에 대한 측정치로 이루어져 있다. cDNA microarray는 1376개의 인간 유전자와 EST(expressed sequence tag)에 대한 측정치이다. 즉, 1534개의 자질(feature)을 가진 60개의 표본(sample)으로 구성된 데이터이다. 모든 자질값들은 평균 0, 표준편차 1로 정규화(normalization)된 실수값이다. 베이지안망 학습을 용이하게 하기 위해 모든 자질값들은 평균값을 기준으로 0(under-expressed)과 1(over-expressed)로 이진화되었다. 또한, 결측치를 가지는 자질들을 제외하고 566개의 유전자, molecular marker와 82개의 약물만을 유전자망 구성에 사용하였다. 각 유전자 발현값(gene expression level)과 약물에 대한 반응(drug activity)들을 노드로 가지는 베이지안망을 통해서 암세포의 약물에 대한 반응과 유전자 발현 패턴과의 관계를 분석할 수 있다.

3.2 베이지안망 학습

베이지안망 학습 알고리즘에는 의존성 분석(dependency analysis)에 기반한 방법 [11]과 최적화(optimization)에 기반한 방법 [3]이 있다. 그러나 대부분의 베이지안망 학습 알고리즘은 지수적으로 증가하는 시간 복잡도 때문에 수백개의 노드를 가지는 베이지안망의 학습에는 적용이 어렵다. 이러한 문제를 해결하기 위해 [4]는 "sparse candidate algorithm"을 제시하였다. 이 알고리즘은 최적화에 기반한 greedy search 알고리즘의 일종으로 불필요한 탐색 공간을 줄이기 위해 각 노드의 후보 부모 집합(candidate parents set)을 이용한다. 본 논문에서는 역시 greedy search에 기반한 알고리즘으로 Markov blanket [8]의 개념을 이용해서 불필요한 탐색 공간을 줄이는 "local to global search algorithm"을 이용했다 [1].

3.2.1 Markov Blanket

n개의 변수의 집합 $\mathbf{X} = \{X_1, \ldots, X_n\}$의 변수 X_i의 Markov blanket [1] $BL(X_i)(\subset \mathbf{X})$는 X_i를 $\mathbf{X} - BL(X_i)$와 확률적으로 조건부독립으로 만드는 변수들의 집합이다. $\mathbf{X}$에 대한 베이지안망의 구조가 알려진 경우 X_i의 Markov blanket은 X_i의 부모노드와 자식노드, 그리고 자식노드의 부모노드이다.

3.2.2 Local to Global Search 알고리즘

"Local to global search" 알고리즘은 표 1과 같다. 알고리즘은 크게 두 단계로 구성된다. "Local Search Step"은 각 노드의 Markov blanket을 구하는 단계이다. 노드 X_i의 Markov blanket, $BL^n(X_i)$는 다음과 같이 구해진다. 우선, X_i의 후보 Markov blanket, CB_i^n을 구한다. CB_i^n의 크기 k는 알고리즘의 파라미터로, 미리 정해주며 보통 7 ~ 15의 값을 사용한다. 현재 알려져 있는 베이지안망의 구조 B_{n-1}을 이용해서 CB_i^n의 원소를 구한다. 그리고 조건부 상호정보량(conditional mutual information)을 이용해서 CB_i^n의 나머지 원소들을 정한다. 그 다음 CB_i^n과 X_i에 대해서 greedy search를 행한다. 이렇게 학습된 지역 구조로부터 $BL^n(X_i)$를 구한다. "Local Search Step"이 끝나면 각 노드의 Markov blanket들을 합쳐서 전체 그래프 H_n을 구성한다. H_n은 cycle을 포함할 수 있기 때문에 베이지안망은 아니다.

표 1. Local to global search 알고리즘

Input:
- A data set D.
- An initial Bayesian network structure B_0.
- A decomposable scoring metric,

$Score(B,D) = \sum_i Score(X_i \mid Pa_i^B(X_i), D)$.

Output: A Bayesian network structure B.

Loop for $n = 1, 2, \ldots$, until convergence.

-Local Search Step:

*Based on D and B_{n-1}, select for each variable X_i, a set $CB_i^n(|CB_i^n| \le k)$ of candidate Markov blanket of X_i.

*For each set $\{X_i, CB_i^n\}$, learn its local structure and determine the Markov blanket of X_i, $BL^n(X_i)$, from this local structure.

*Merge all the local network structures $G(\{X_i, BL^n(X_i)\}, E_i)$ into a global network structure H_n (usually cyclic).

-Global Search Step:

*Find the Bayesian network structure $B_n \subset H_n$, which maximizes $Score(B_n, D)$ and retains all non-cyclic edges in H_n.

"Global Search Step"에서는 H_n에 포함되면서 점수가 높은 베이지안망 구조 B_n을 greedy search로 탐색한다. 이때, cycle을 이루지 않는 H_n의 간선은 모두 유지한다. 알고리즘의 수렴 조건은 $Score(B_{n-1}, D) \ge Score(B_n, D)$이다. 또한 "Global Search Step"에서 H_n의 non-cyclic edge는 모두 유지함으로써 greedy search의 불필요한 탐색 공간을 크게 줄인다.

4. 실험 결과

실험에서는 3.2절의 베이지안망 학습 방법을 이용해서 648개의 노드(유전자 및 molecular marker 566개 + 약물 82개)를 가지는 베이지안 유전자망을 학습했다. greedy search의 $Score(B, D)$로는 BD (Bayesian Dirichlet) metric [6]을 이용했으며 Dirichlet prior인 α_{ijk}값으로는 uninformative prior인 1.0이 설정되었다. 베이지안망 학습에 걸린 시간은 펜티엄III 1GHz 컴퓨터에서 약 8분 정도였으며 학습된 베이지안망의 최종 점수는 -356.18이었다. 학습된 베이지안망이 나타내는 관계에는 생물학적으로 이미 검증된 사실들이 포함되어 있으며 그 예는 다음과 같다.

4.1 약물 반응간의 관계

실험 데이터에 포함된 약물 중 "aphidicolin-glycinate", "floxuridine", "cyclocytidine", "cytarabine"은 모두 pyramidine analogue이며 [9]에서의 클러스터링에서는 하나의 클러스

[1] 어떤 변수에 대한 Markov blanket 은 다수가 존재할 수 있으며 [8]에서는 이러한 Markov blanket 중 최소의 크기를 가지는 것을 Markov boundary 라고 정의하고 있다. 본 논문에 나오는 Markov blanket 은 모두 Markov boundary 에 해당한다.

터에 속했다. 학습된 베이지안망이 표현하는 이들간의 관계는 그림 1과 같다.

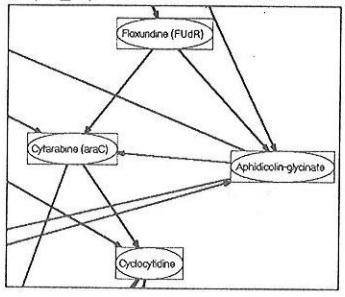

그림 1. Pyramidine analogue 약물군간의 관계

그림 1의 베이지안망에서 이 약물들의 반응은 서로 밀접한 관련이 있다는 사실을 알 수 있다. 실제로 이들은 같은 암세포에 대해서 비슷한 반응을 보인다는 사실이 알려져 있다 [9].

4.2 약물 반응과 유전자 발현간의 관계

신약 개발 과정에 있어서 중요한 부분의 하나는 약물의 반응과 유전자 발현간의 관계를 밝히는 일이며 NCI 60 cell lines dataset은 이러한 분석을 위해 만들어 졌다. 실험적으로 검증된 관계 중 하나는 약물 "L-asparagine"에 대한 반응과 유전자 "ASNS"의 발현간의 관계이다. 이 관계는 학습된 베이지안망에서 그림 2와 같이 나타난다.

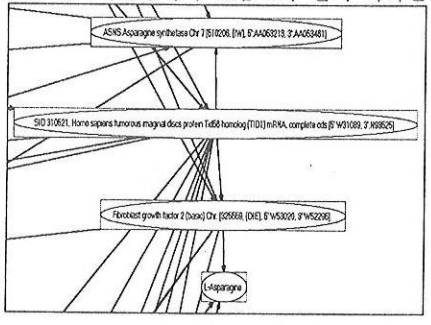

그림 2. 유전자 "ASNS"의 발현과 약물 "L-asparagine"에 대한 반응 사이의 관계

그림 2에서 "ASNS"와 "L-asparagine" 사이의 관계는 두 개의 유전자 "Homo sapiens tumorous imaginal discs protein Tid56 homolog (TID1) mRNA, complete cds"와 "Fibroblast growth factor 2 (basic) Chr."을 통해서 이루어진다. 실제로 서로 관련이 있는 유전자와 약물이 학습된 베이지안망 구조에서 가깝게 위치함을 알 수 있다. 물론, 그림 2는 아직 생물학적으로 검증된 관계는 아니며 cDNA microarray 데이터와 약물반응 데이터로부터 학습된 베이지안망에서 얻어진 결과로 검증을 위해서는 생물학 실험이 필요하다.

5. 결론

본 논문에서는 NCI 60 cell lines dataset을 이용해서 약물 반응과 유전자 발현간의 관계를 분석할 수 있는 베이지안 유전자망을 구성하였다. 이를 위해서 "local to global search" 알고리즘을 이용, 648개의 노드를 가지는 베이지안망을 구성하였다. 학습된 베이지안망은 유전자-유전자 관계, 유전자-약물 관계, 약물-약물 관계를 모두 나타내며 생물학적 실험을 통해 실제로 검증된 관계들도 다수 포함한다. 이는 cDNA microarray 데이터와 다른 실험 데이터를 함께 이용해서 약물에 대한 반응과 유전자 발현과의 관계를 분석할 수 있는 가능성을 보인다고 할 수 있다. 물론 학습된 베이지안망이 나타내는 관계들은 cDNA microarray 데이터를 비롯한 생물학 데이터에 내재되어 있는 잡음과 에러, greedy search 알고리즘의 한계 때문에 모두 정확하다고는 볼 수는 없다. 따라서 실제 생물학적 관계의 검증을 위해서는 보다 정밀한 실험이 따로 필요할 것이다. 하지만, 베이지안망의 학습에 걸리는 시간은 아주 적으며 다음 단계의 생물학적 실험 방향에 대한 일종의 안내자 역할을 할 수 있을 것으로 기대한다. 향후, 좀 더 정밀한 베이지안망을 효율적으로 구성하기 위한 데이터 전처리와 학습 알고리즘에 대한 연구가 필요하다.

감사의 글

이 논문은 교육부 BK21 사업에 의하여 지원되었음.

참고 문헌

[1] 황규백, 장병탁, 대규모 베이지안망 구조 학습 알고리즘, *'01 한국뇌학회 학술대회 논문집*, pp. 100-101, 2001.

[2] Eisen, M.B. et al., Cluster analysis and display of genome-wide expression patterns, *PNAS USA*, vol. 95, pp. 14863-14868, 1998.

[3] Friedman, N. and Goldszmidt, M., Learning Bayesian networks with local structure, *Learning in Graphical Models*, pp. 421-459, MIT Press, 1999.

[4] Friedman, N. et al., Learning Bayesian network structure from massive datasets: the "sparse candidate" algorithm, In *Proc. of UAI'99*, pp. 206-215, 1999.

[5] Friedman, N. et al., Using Bayesian networks to analyze expression data, In *Proc. of RECOMB'00*, pp. 127-135, 2000.

[6] Heckerman, D. et al., Learning Bayesian networks: the combination of knowledge and statistical data, *Machine Learning*, vol. 20, pp. 197-244, 1995.

[7] http://discover.nci.nih.gov/nature2000/

[8] Pearl, J., *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Morgan Kaufmann, 1988.

[9] Sherf, U. et al., A gene expression database for the molecular pharmacology of cancer, *Nature genetics*, vol. 24, pp. 236-244, 2000.

[10] Spellman, P.T. et al., Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization, *Molecular Biology of the Cell*, vol. 9, pp. 3273-3297, 1998.

[11] Spirtes, P. et al., *Causation, Prediction, and Search*, 2nd ed., MIT Press, 2000.

[12] Szallasi, Z., Genetic network analysis in light of massively parallel biological data acquisition, In *Proc. of PSB'99*, pp. 5-16, 1999.