# 유전자 알고리즘으로 학습한 베이지안 네트워크에 기초한 질병 모듈 추론

정다예*, 여윤구*, 안재균**, 박상현†
*연세대학교 컴퓨터과학과
**Integrative Biology and Physiology, UCLA
e-mail: rabilish@cs.yonsei.ac.kr

# Inference of Disease Module using Bayesian Network by Genetic Algorithm

Da-Ye Jeong*, Yun-ku Yeu*, Jae-Gyoon Ahn**, Sang-Hyun Park†
*Dept of Computer Science, Yonsei University
**Dept of Integrative Biology and Physiology, UCLA

요 약

사람의 질병은 여러 요인의 복합적인 작용으로 발생하는데 이 중 유전적인 요인에는 유전자 간의 상호작용을 들 수 있다. 마이크로어레이(Microarray) 데이터로부터 유전자의 활성화 및 억제 관계를 밝히려는 다양한 시도는 계속되어왔다. 그러나 마이크로어레이 자체가 갖는 불안정성과 실험조건 수의 제약이 커다란 장애가 되어 왔다. 이에 생물학적 사전 지식을 포함하는 방법들이 제안되었다. 본 논문에서는 질병과 관련된 유전자 간의 상호작용의 집합을 질병 모듈이라 정의하고 이를 유전자 알고리즘으로 학습한 베이지안 네트워크(Bayesian network)로 추론하는 방법을 제안한다.

## 1. 서론

사람의 질병은 많은 환경적인 요인과 유전적인 요인의 복합적인 작용으로 발생한다. 유전적인 요인에는 유전자간의 상호 작용을 들 수 있는데, 여기에는 다양한 조절 관계가 얽혀있다. 유전자가 다른 유전자를 활성화하기도 하고 억제하기도 한다. 또한 유전자 스스로 활성화 또는 억제하기도 한다. 질병과 관련된 유전자 간의 조절 관계를 밝히는 것은 질병을 이해하는 일과 같고 이는 곧 질병에 대처하는 방법을 제시하는 길이다.

그러나 특정 질병과 관련된 유전자 간의 조절 관계에 대한 정보는 유전자에 담겨있지 않다. 유전자 간의 활성화 및 억제 관계는 유전자와 단백질의 중간물질인 mRNA(messenger RNA)를 분석하여 얻을 수 있는데 이 mRNA 발현 값은 마이크로어레이 등을 통해 얻을 수 있다.

마이크로어레이 데이터를 베이지안 네트워크로 분석하여 유전자 간의 억제와 활성화의 관계를 밝히려는 시도는 계속되어왔다[1]. 그러나 마이크로어레이가 갖는 제한적인 샘플 개수와 노이즈가 큰 장애물이 되었다. 이에 마이크로어레이 데이터만 이용하는 것이 아니라 생물학적 사전 지식도 베이지안 네트워크에 적용하는 방법이 제안되었다[2].

질병과 관련된 유전자를 찾는 연구는 상대적으로 많이 이루어져 왔으나, 그에 비해 질병 유전자 간의 연관 관계를 추론하는 연구는 상호 작용의 복잡성으로 인해 어려움을 겪고 있다. 그러나 질병과 생명 활동의 원인을 보다 정확하게 이해하기 위해서는 유전자 간의 연관 관계를 이해하는 것이 필수적이다. 본 논문에서는 질병과 관련 있는 유전자 상호작용의 집합을 질병 모듈이라 정의하고, 이 모듈의 구조를 유전자 알고리즘으로 학습한 베이지안 네트워크로 추론하는 방법을 제안한다.

베이지안 네트워크는 원인-결과 관계를 확률적으로 모델링하고, 구성 요소 간의 부분적인(locally) 상호 작용으로 구성된 프로세스를 나타내는 효과적인 방법(tool)이다[1]. 베이지안 네트워크는 부모 노드의 확률이 조건부 확률로 자식노드에게 영향을 미치기 때문에 네트워크 구조에 따라 노드가 갖는 확률은 변화될 수 있다.

본 논문에서는 유전자 알고리즘으로 네트워크 구조를 학습한다. 유전자 알고리즘은 유전학과 자연 선택의 메커니즘을 기초로 한 탐색 알고리즘이다[3]. 유전자 알고리즘은 변이, 교차를 거쳐 적합도가 높은 개체는 살아남고 그렇지 않은 개체는 도태된다. 본 논문에서 한 개체의 적합도는 생성된 네트워크의 구조가 마이크로어레이 데이터를 얼마나 반영하였는가와 단백질 간의 상호작용 데이터(이하 PPI)를 네트워크 구조에 잘 나타나 있는가로 계산하였다.

본 논문은 통계 프로그램 R을 이용하여 베이지안 네트워크를 구현하였다. 네트워크를 생성하기 위해 graph 패키지를 사용하였다[6].

---

† 연세대학교 컴퓨터과학과 교수, 교신저자

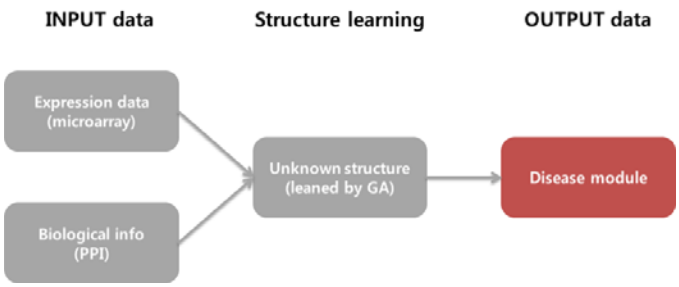
※ 이 논문은 2013년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(2012R1A2A1A01010775).

본 논문에서는 전립선암의 마이크로어레이 데이터와 OPHID(Online Predicted Human Interaction Database)의 PPI 데이터로 실험을 수행하였다[4,5]. KEGG(Kyoto Encyclopedia of Genes and Genomes)에 전립선암의 알려진 대사 경로[7]와 같은 유전자로 노드를 구성한 뒤, 네트워크의 구조를 추론하였을 때, 대사 경로에 나타난 엣지를 포함하는 네트워크를 생성하였다. 이 네트워크들의 유의확률 0.04로 통계적으로 유의미한 네트워크가 생성되었다.

본 논문은 다음과 같은 구성으로 이루어져 있다. 2장에서는 본 논문에서 제안하는 방법론에 대해 설명하고, 3, 4장에서는 베이지안 네트워크 구성 방법에 대해 설명한다. 5장에서는 방법론을 검증하기 위한 실험 환경 및 방법 과 결과를 설명하고 마지막으로 6장에서는 결론 및 추후 연구 방향에 대해서 설명한다.

## 2. 방법론 개요

본 논문에서 사용한 방법론의 개략적인 흐름은 (그림 1)과 같다. 유전자 알고리즘과 PPI 데이터로 구조를 학습하여 베이지안 네트워크를 생성한다.

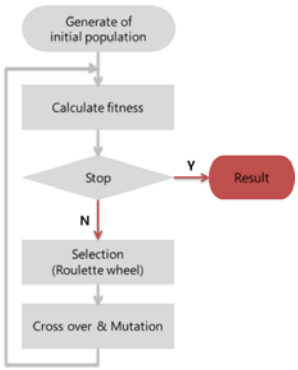



(그림 1) 방법론의 개략적인 흐름

마이크로어레이 데이터는 Friedman et al[1]의 방법을 사용하여 유전자 발현 정도를 +(과다 발현), 0(변화 없음), -(과소 발현)의 3가지 카테고리의 상태로 이산화(discretization)한다.

## 3. 유전자 알고리즘으로 학습한 네트워크

유전자 알고리즘의 개략적인 흐름은 (그림 2)와 같다.




(그림 2) 유전자 알고리즘의 흐름

먼저 최초의 개체군을 무작위로 생성한다. 이때 생성되는 염색체의 형태는 (그림 3)과 같다.

| $G_{2,1}$ | $G_{3,1}$ | $G_{3,2}$ | $G_{4,1}$ | ... | $G_{n,n-4}$ | $G_{n,n-3}$ | $G_{n,n-2}$ | $G_{n,n-1}$ |
|---|---|---|---|---|---|---|---|---|

(그림 3) 유전자 알고리즘의 염색체

염색체는 베이지안 네트워크상에서 노드 간의 관계를 나타낸다. 염색체의 각 유전자는 <-1, 0, 1>의 값 중 하나의 값을 갖는다. 만약 $G_{2,1}$이 -1의 값을 갖는다면 노드1에서 노드2로 가는 엣지가 생성된다. $G_{2,1}$이 1의 값을 갖는다면 노드2에서 노드1로 가는 엣지가 생성되고 0을 갖는다면 두 유전자 사이에 엣지는 없는 것으로 한다.

생성된 개체의 적합도는 베이지안 네트워크의 리프 노드(leaf node)의 조건부 확률의 값이 가장 높은 상태와 테스트 셋의 상태를 비교하여 얼마나 일치하는지에 대한 비율과 네트워크에서 나타난 PPI의 정보를 가진 엣지의 비율로 얻어진다. 적합도가 높은 염색체는 룰렛 휠(roulette-wheel) 알고리즘에 의해 더 많이 선택된다. 선택된 개체들은 변이와 교차를 거쳐 다시 적합도를 계산한다. 적합도의 계산 방법은 4장에서 더 자세히 설명한다.

유전자 알고리즘은 정해둔 세대만큼 수행하며 이에 도달하면 종료한다. 여러 번의 세대를 거쳐 적합도가 높은 염색체는 살아남고 그렇지 않은 염색체는 도태된다.

유전자 알고리즘으로 인해 노드 간 엣지가 임의로 생성되기 때문에 사이클이 생성될 수 있다. 베이지안 네트워크는 비순환 방향성 그래프이기 때문에 노드 사이에 순서를 매겨 순서가 먼저인 노드에서 나중인 노드로 엣지를 생성하여 사이클을 방지하였다.

또한, 엣지의 개수가 임의로 생성되어 밀집도가 높은 네트워크가 생성될 수 있다. 이는 노드의 개수의 2~4배 사이의 정도로 엣지의 개수를 제한하여 비교적 밀집도가 낮은 네트워크를 생성하였다.

## 4. 네트워크 적합도

데이터 D에 대한 네트워크 G의 사후 확률 $P(G|D)$는 네트워크가 주어졌을 때 데이터의 가능성(likelihood) $P(D|G)$과 사전 지식에 기초한 네트워크 구조의 확률, 즉 네트워크의 사전 확률 $P(G)$의 곱에 비례한다. 데이터 D에 대해 네트워크의 구조의 사후 확률이 높은 네트워크를 가장 적합한 네트워크로 선택한다.

$$P(G|D) \propto P(D|G) * P(G)$$

(수식 1) 네트워크의 사후확률

본 논문에서는 네트워크가 주어졌을 때 데이터의 가능성, 즉 $P(D|G)$는 리프 노드와 테스트 셋의 유전자 발현 상태를 비교하여 구하였다. 또한 네트워크의 구조의 사전

확률 $P(G)$는 생성된 네트워크의 전체 엣지와 PPI의 사전 지식이 있는 엣지의 비율로 구하였다.

개체의 적합도는 테스트 셋의 상태와 일치하는지를 비교하기 때문에 테스트 셋에 대한 오버피팅이 생길 수 있다. 또한, 마이크로어레이는 유전자에 비해 샘플이 적기 때문에 교차 검증법을 사용한다. 본 논문에서는 leave-one-out 교차 검증법을 사용하였다.

## 5. 실험 및 결과

질병 모듈을 추론하기 위해서 전립선암과 관련된 마이크로어레이 데이터를 수집하였다[4]. 수집한 데이터는 전립선암 환자의 샘플 18개, 정상인의 전립선 샘플 21개로 구성되어 있고 모든 샘플은 45220개의 유전자 발현 값을 포함하고 있다. 본 논문에서는 그 중 전립선암 환자의 샘플 18개를 사용하여 실험하였다.

네트워크의 사전 확률을 계산하기 위해 OPHID의 PPI 데이터 중 인간의 데이터만 사용하였다[5].

마이크로어레이 데이터는 노이즈를 많이 포함하고 있고 알고리즘의 복잡성 때문에 해당 질병과 관련 있는 유전자들로 특징 선택(feature selection)을 수행하였다. 본 논문에서는 KEGG에 전립선암의 대사 경로에 나타난 41개의 유전자를 선택하여 노드로 구성하였다[7].

실험에 사용한 유전자 알고리즘 관련 파라미터들은 <표 1>과 같다.

<표 1> 실험의 사용한 유전자 알고리즘 파라미터

| 파라미터 | 값 |
|---|---|
| Generation | 100 |
| Population size | 40 |
| Mutation rate | 0.5 |
| Crossover rate | 0.5 |

n개의 추론된 엣지에서 k개의 대사경로에서 나타난 엣지, 즉 정답 엣지를 지닌 네트워크의 유의 확률(p-value)은 같은 수의 엣지를 임의로 선택하는 무작위 네트워크에서 최소 k개의 정답 엣지가 포함될 확률 분포를 모델링하여 추정할 수 있다. 노드의 개수가 고정되어 있다고 가정하고, N을 네트워크 내에서 생성 가능한 엣지의 최대 개수라 할 때, M개의 정답 엣지 집합과 N-M개의 오답 엣지 집합이 존재한다고 가정한다. 초기하분포에 따라 m개의 정답 엣지와 n-m개의 정답 엣지를 찾을 확률은 수식 (2)와 같다. N은 노드의 개수가 g개였을 때 ${}_gC_2$가 된다. 본 논문에서는 노드에 순서를 매겨 사이클을 제거하였으므로 생성된 네트워크의 엣지의 방향을 고려하지 않고 정답 엣지의 개수를 계산하였다.

$$P_x = \frac{\binom{M}{m}\binom{N-M}{n-m}}{\binom{N}{n}}$$

(수식 2) m개의 정답 엣지와 n-m개의 오답 엣지를 찾을 확률

따라서 네트워크의 유의 확률은 (수식 3)과 같다.

$$P-value = \sum_{m=0}^{k} \frac{\binom{M}{m}\binom{N-M}{n-m}}{\binom{N}{n}}$$

(수식 3) 네트워크의 유의 확률

대사 경로에 관여된 유전자 41개로 구성된 노드로 엣지의 최대 개수에 제한을 두어 실험하였을 때 결과는 <표 2>와 같았다. 실험은 각기 3번씩 수행하여 생성된 네트워크의 평균적인 유의 확률을 계산하였다. 찾아야 하는 정답 엣지의 수 M은 36개이다.

<표 2> 엣지 개수를 다르게 하였을 때 유의 확률

| 노드 개수 | 엣지의 최대개수 | 생성된 엣지 수 | 정답 엣지 수 | 유의 확률 |
|---|---|---|---|---|
| 41 | 100 | 48, 54, 41 | 3, 7, 4 | 0.06 |
| 41 | 200 | 116, 120, 108 | 8, 13, 8 | 0.03 |
| 41 | 400 | 215, 204, 204 | 14, 14, 13 | 0.03 |

위 실험에서 평균 유의 확률 0.04로 유의수준 0.05에서 유의미한 것으로 나타났다.

## 6. 결론 및 추후 연구 방향

질병 모듈을 찾는 것은 질병에 대한 이해도를 높이는 방법 중 하나이다. 그러나 대부분 질병과 관련된 유전자 간의 상호 작용은 생물학적인 실험을 통해 밝혀지는 것이 보통이다.

본 논문에서는 유전자 알고리즘으로 베이지안 네트워크를 학습하여 질병 모듈의 구조를 추론하는 방법론을 제안하였다. PPI와 마이크로어레이만으로 네트워크의 구조를 추론하였으며, 유전자간의 상호 작용에 대하여 통계적으로 유의미한 결과를 얻을 수 있음을 확인하였다.

본 논문은 알고리즘의 복잡성 때문에 노드를 전립선암의 대사 경로의 유전자들로만 노드를 구성하여 실험하였다. 노드를 다양하게 구성하여 대사 경로의 상호 작용 뿐만 아니라 알려지지 않은 유전자 간의 상호 작용도 추론할 수 있도록 개선할 예정이다. 네트워크 구조의 사전 확률을 단순히 네트워크에 나타난 엣지와 PPI 데이터가 존재하는 엣지의 비율로만 구하는 것이 아니라 Pubmed같은

문헌 데이터나 알려진 유전자 간의 조절 관계 데이터도 접목시킬 예정이다.

## 참고문헌

[1] Friedman, Nir, et al. "Using Bayesian networks to analyze expression data." Journal of computational biology 7.3-4 (2000): 601-620

[2] Imoto, Seiya, et al. "Combining microarrays and biological knowledge for estimating gene networks via Bayesian networks." Journal of Bioinformatics and Computational Biology 2.01 (2004): 77-98.

[3] Larrañaga, Pedro, et al. "Structure learning of Bayesian networks by genetic algorithms: A performance analysis of control parameters." Pattern Analysis and Machine Intelligence, IEEE Transactions on 18.9 (1996): 912-926.

[4] Aryee MJ, Liu W, Engelmann JC, Nuhn P et al. DNA methylation alterations exhibit intraindividual stability and interindividual heterogeneity in prostate cancer metastases. Sci Transl Med 2013 Jan 23;5(169):169ra10. PMID: 23345608

[5] Brown, Kevin R., and Igor Jurisica. "Online predicted human interaction database." Bioinformatics 21.9 (2005): 2076-2082.

[6] R. Gentleman, Elizabeth Whalen, W. Huber and S. Falcon (2006). graph: A package to handle graph data structures. R package version 1.38.3.

[7] Kanehisa, M. and Goto, S.; KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. 28, 27-30 (2000).