재구성된 유전자 네트워크의 섭동적(Perturbational) 토폴로지 변형 분석

이상근[0] 장병탁
서울대학교 컴퓨터공학부
sklee[0]@bi.snu.ac.kr btzhang@cse.snu.ac.kr

Power-law Distributional Perturbation Analysis of the Topology of Reconstructed Genetic Networks

Sang-Keun Lee[0] and Byoung-Tak Zhang
School of Computer Science and Engineering, Seoul National University

요 약

DNA칩 기술로 얻어지는 대규모 섭동데이터(perturbation data)는 생물학적시스템(biological system)의 유전자네트워크(genetic network)를 재구성(reverse-engineering)하는데 있어 유용하다. 그러나 기존의 연구는 유전자 조절 관계의 규명이나 혹은 데이터를 설명하는 최적의 모델을 찾는 방향에만 관심을 두고 있고, 실험적인 한계로 인한 DNA칩 데이터의 오류가 재구성된 네트워크의 구조에 미치는 영향에 대해서는 중요하게 다루고 있지 않다. 본 논문에서는 유전자 네트워크의 멱함수(power-law) 분포 구조를 이용하여, 섭동 데이터의 오류가 재구성된 네트워크의 토폴로지(topology)에 미치는 영향을 분석하였다. 가상의 네트워크에 대한 데이터를 사용하여 실험한 결과, 데이터의 오류 정도에 따른 네트워크 토폴로지의 변형 양상을 관측할 수 있었다.

1. 서 론

생물학적시스템(biological system)의 유전자 네트워크(genetic network)를 재구성하는 작업은 기능게놈학(functional genomics)의 가장 중요한 과제 중 하나이다. 그 세부 과정은 유전자 네트워크의 인과구조(casual relationship)을 규명하고, 각 유전자 산물(gene product)에 대한 반응서열을 결정하는 것이다. 이러한 과정을 수행하는데 있어 필요한 것은 각 유전자의 외부 자극, 혹은 섭동(perturbation)에 대한 반응 데이터이며, DNA칩 기술로 얻을 수 있는 대규모(high-throughput) 섭동데이터는 유전자 네트워크 재구성 연구에 많은 도움을 주고 있다[2].

한편 유전자네트워크 재구성을 위한 기존의 연구는 주로 섭동에 따른 유전자의 발현 양상의 유사성(similarity)에 따라 유전자를 분류하여 유전자 사이의 조절관계(regulatory interaction)를 규명하거나, 계산학적 측면에서 데이터를 설명하는 최적의 네트워크를 찾는 방법론을 개발하는 방향으로 진행되어 왔다[2, 4, 5, 7].

하지만 DNA칩 실험을 통해 얻어지는 섭동 데이터는 그 획득 과정에서 실험기기의 간섭 등으로 인한 여러 가지 오류를 포함할 수 있으며, 이러한 오류는 데이터를 기반으로 재구성된 네트워크의 구조에 중요한 영향을 미칠 수 있다. 기존의 연구에서는 이러한 관점을 중점적으로 다루고 있지 않지만, 실제로 실험을 통하여 네트워크를 재구성하기 위해서는 실험 데이터의 오차 허용 범위와 같은 구체적인 지침이 반드시 필요하다.

본 논문에서는 유전자네트워크의 멱함수(power-law)적인 노드(node)와 링크(link)의 분포 구조를 이용하여 네트워크의 토폴로지(topology)를 분석하는 방법론을 제안하였다. 또 가상으로 구성한 데이터에 이 방법론을 적용하여 데이터의 오류에 따라 재구성된 네트워크의 토폴로지가 얼마나 변형되는가를 측정하였다.

2. 네트워크 재구성

2.1 멱함수(power-law) 분포

복잡네트워크(complex network)의 멱함수 분포란 k 개의 링크를 갖는 노드가 있을 확률 $P(k)$ 가 다음과 같이 주어진다는 것이다[3].

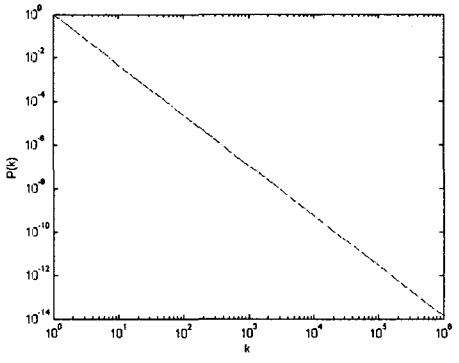


그림 1. k개의 링크를 갖는 노드 수의 분포

즉, 근사적으로 다음과 같은 분포를 갖는 경우이다.

$$P(k) \approx e^{-k} \quad (1)$$

여기서 우리가 관심을 갖는 부분은 k 가 작거나 충분히 큰 경우이다. 즉, 적은 수의 링크를 갖는 노드의 수가 많은 한편, 매우 많은 수의 링크를 갖는 노드, 즉 허브(hub) 노드의 존재를 이 분포로부터 추론할 수 있다. 인터넷이나 대사경로(metabolic pathway)와 같은 많은 복잡네트워크가 멱함수 분포 특성을 보인다는 것이 알려져 있다[3, 6]. 이러한 허브 노드는 다른 많은 노드들과의 연결 관계에 대한 정보를 가지므로, 허브 노드의 토폴로지는 전체 네트워크의 토폴로지를 결정하는데 중요한 역할을 한다.

2.2 섭동 데이터 (Perturbation Data)

우선 섭동데이터로부터 추론 가능한 유전자간의 관계를 직접연결(direct link)관계와 간접연결(indirect link)관계를 구분한다. 직접연결관계는 어떤 유전자가 중간 과정 없이 다른 유전자의 발현에 영향을 주는 경우이고, 간접연결관계는 중간 과정이 있는 경우이다[2]. 예를 들어 유전자 G1의 산물(product)가 유전자 G2의 전사요소(transcriptional factor)인 경우 G1-G2는 직접연결관계에 해당한다. 또 유전자 G2의 산물이 유전자 G3로부터 얻어지는 단백질의 인산화효소(phosphatase)이고, 인산화된 단백질이 유전자 G4의 전사요소로 작용한다면 G2-G4의 관계는 간접연결관계이다. 또, 직접연결관계만의 목록을 인접목록 (adjacency list), 간접연결관계까지를 포함한 목록을 연결목록 (accessibility list)라고 정의한다[2]. 본 논문에서는 완전한 인접목록을 알고 있다는 가정하에 네트워크 토폴로지를 분석한 후, 이 목록에 일정한 오류를 추가하면서 재구성된 토폴로지의 변화 양상을 분석하였다.

2.3 토폴로지 재구성 알고리듬

본 논문에서 제시하는 이 알고리듬은 앞에서 언급한 복잡네트워크의 멱함수 분포 특성을 이용하여 (1) 허브 노드를 찾고, (2) 허브간의 위상구조를 파악하며, (3) 허브에 의한 나머지 노드들의 그룹화 상태를 찾는다. 알고리듬의 개요는 다음과 같다.

Reconstruct(N, adj/acc-list)
- 링크의 분포를 분석한다. (power-law)
- 분석 결과로부터 허브를 찾아낸다.
  → H = set of hubs
- 각 허브에 대해 노드들을 그룹화한다.
  → L = set of new adj/acc-list
  → G = set of group (by H, of N)
- For each G_i in G
  Reconstruct(G_i, new-adj/acc-list)

그 과정을 상세히 살펴보면, 우선 네트워크 전체의 링크 분포를 분석하여 상대적으로 많은 링크를 갖는 노드인 허브를 찾는다. 다음 각 허브와 연결 관계를 갖는 노드들의 인접연결목록을 추출하고, 추출된 각 목록에 대해 허브를 찾는 과정부터 같은 과정을 재귀적으로 반복하는 것이다.

3. 실 험

실험은 멱함수 분포를 갖는 가상적인 유전자 네트워크를 대상으로 하였다. 우선, 인터넷 네크워크 생성 소프트웨어인 BRITE[1]를 사용하여 300개의 노드와 600개의 링크를 갖는 멱함수 분포 네크워크를 생성하였다. 이 네크워크에 대해 오류가 없는 섭동데이터를 생성하고, 2.2의 알고리듬을 사용하여 허브 토폴로지를 재구성하였다. 다음으로 오류를 포함하는 섭동데이터를 생성하고, 2.2의 알고리듬을 사용하여 토폴로지를 재구성한 후 원래의 토폴로지와의 거리를 측정하였다.

3.1 실험 데이터 오류

섭동데이터에서 발생 가능한 오류는 다음의 3가지 경우로 정의하였다.

(1) False Adjacent: 인접노드목록에 다른 노드가 끼어드는 경우
(2) Spurious: 인접 노드의 정보가 바뀌는 경우
(3) Missing: 인접 노드의 정보가 누락되는 경우

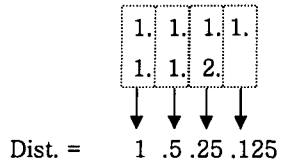
(1)은 네트워크 재구성에 사용한 잘못된 사전지식 혹은 실험에서 얻어진 시계열(time-series) 데이터의 오류를 나타낸 것이며, (2), (3)은 그 밖의 실험기기의 오차로 인해 발생할 수 있는 오류를 나타낸 것이다.

3.2 네트워크 토폴로지 유사도

재구성된 네트워크 토폴로지간의 거리(distance) 를 측정하기 위해 각 허브마다 인터넷(internet)의 IP주소(IP address)와 유사한 체계의 이름을 부여하였다. 부여된 이름의 예를 들면 다음과 같다.

1. (hub id = 37)
 1.1 (hub id = 4)
  1.1.1 (hub id = 128)
  1.1.2 (hub id = 12)
 1.2 (hub id = 97)
...

다음으로 서로 다른 두 네트워크에서 각 허브의 이름(허브 거리) 혹은 허브가 아닌 노드가 속한 허브의 이름(그룹화 거리)을 비교하여 거리(distance)를 측정하였다. 즉, 이름의 각 자리(position)을 비교하여 맞지 않을 경우 다음과 같이 자리에 따라 거리를 부여하여 합산하였다. 앞쪽에 있는 번호일수록 큰 허브를 지칭하게 되므로 1의 거리를 부여했고, 뒤로 갈수록 1/2씩의 거리를 갖도록 하였다.



3.3 실험 결과

3.1에서 정의한 3가지 상황 각각에 대하여 오류율을 변화시키면서 생성된 네트워크 토폴로지의 거리를 측정하였다.

그림 2에서 볼 수 있듯이, 오류에 따른 허브 구조의 변형은 오류율 0~10%에서 급격히 증가하지만, 세 가지 오류 발생 경우에 있어 비슷한 경향을 보이고 있음을 알 수 있다. 즉 노드의 인접 목록에 직접적인 영향을 미치는 세가지 오류 발생 경우 모두 허브 토폴로지에 직접적인 영향을 미치고 있음을 추론할 수 있다.

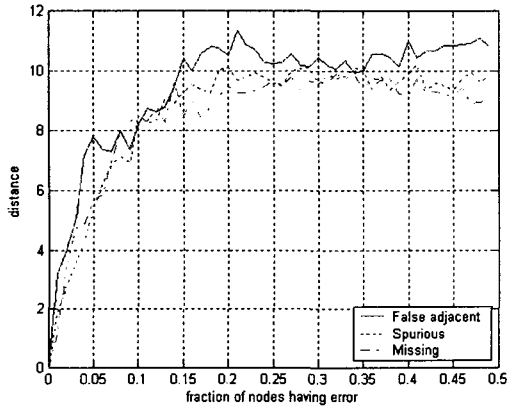


그림 2. 오류에 따른 허브 토폴로지의 거리 변화

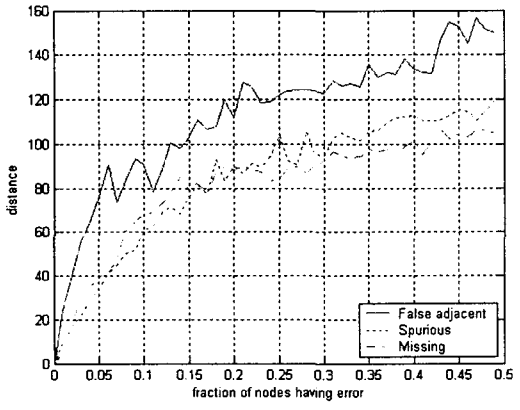


그림 3. 오류에 따른 노드 그룹 거리 변화

한편 그림 3를 보면 오류에 따른 노드의 그룹 거리는 False adjacent의 경우가 가장 큰 영향을 미치고 있음을 알 수 있다. 이것은 다른 두 경우에 비해 False adjacent 오류로 인한 허브 구조의 변형이 노드들의 그룹화에 더 많은 영향을 주고 있기 때문이라 생각된다. 또한 전반적으로 오류율 0~15% 구간에서 급격한 변형을 이루고 있지만, 그림 2과 비교해 볼 때 허브 토폴로지의 변형보다는 완만한 변화 양상을 보이고 있다. 이것은 랜덤 에러에 대한 노드 그룹의 평균적인 변화 양상에 비해 허브 노드가 영향을 받기 쉬운 때문인 것으로 추정되지만, 이 점에 대해서는 추가적인 연구가 필요하다.

4. 결 론

본 논문에서는 유전자 네트워크와 같은 복잡 네트워크에서 허브 토폴로지를 재구성하는 알고리듬을 제안하였다. 이 알고리듬을 이용하여 섭동 데이터로부터 유전자네트워크를 재구성하는 경우에 있어 데이터의 오류가 네트워크 토폴로지에 어떠한 영향을 미치는가를 실험적으로 분석하였고, 그 결과 오류의 정도에 따른 토폴로지의 변형 양상과 토폴로지가 급격히 변화하는 오류 구간을 발견하였다. 향후 이 결과는 유전자네트워크를 재구성하기 위하여 실제 실험을 행하는 경우에 있어 실험의 오차범위를 결정하는데 중요한 기준이 될 것으로 기대된다.

감사의 글

이 논문은 과학기술부의 NRL 및 Systems Biology 사업에 의하여 지원되었음.

참고문헌

[1] Alberto, M., Anukool, L., Ibrahim, M., and John, B., BRITE: An approach to universal topology generation, *Proceedings of MASCOTS '01*, 2001.

[2] Andreas, W., How to reconstruct a large genetic network from n gene perturbations in fewer than n2 easy steps, *Bioinformatics,* vol. 17, pp. 1183-1197, 2001.

[3] Barabsi, A., Albert, R., Jeong, H., and Bianconi, G., Power-law distribution of the World Wide Web, *Science*, vol. 287, p. 2115a, 2000.

[4] DeRisi, J.L., Iyer, V.R., and Brown, P.O., Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, vol. 278, pp. 680-686, 1997.

[5] Eisen, M.B., Spellman, P.T., Brown, P.O., and Botstein, D., Cluster analysis and display of genome-wide expression patterns. *PNAS*, vol. 95, pp. 14863-14868, 1998.

[6] Jeong, H., Tombor, B., Albert, R., Oltvai, Z., and Barabasi, A., The large-scale organization of metabolic networks, *Nature,* vol. 407, p. 651, 2000.

[7] Tavazoie, S., Huges, J.D., Campbell, M.J., Cho, R.J., and Church, G.M., Systematic determination of genetic network architecture. *Nature Genetics.*, vol. 22, pp. 281-285, 1999.