유전자 네트워크 모티프 알고리즘을 이용한 인터넷 네트워크 분석

나하선, 김문환, 나상동*
한국전파기지국(주)
*조선대학교 컴퓨터공학부

Analyzing internet networks using an algorithm for detecting network motifs in Genetics

Ha-Sun Na, Moon-Hwan Kim, Snag-Dong Ra*
KRTnet Corporation Network Planning
*Dept. of Computer Engineering, Chosun University
e-mail:(hsna@krtnet.co.kr)

요 약

복잡한 유전학적 네트워크 구조와 조직을 해석하여 인터넷 네트워크에서 상호연결 하는 "네트워크 모티프"를 연구한다. 다양하고 복잡한 유전학적 네트워크 구조의 설계 및 원리를 해석한 네트워크 모티프를 패턴으로 규정되어 있는 것들을 유전학적 네트워크에서 정보 네트워크(www)와의 동적으로 수행할 수 있는지 알고리즘을 통해 해석한다. 유전학적 네트워크에서 모티프들이 네트워크의 보편적인 class를 규정할 수 있는지 이론적으로 접근하여 인터넷 네트워크에 응용 및 해석하고 분석한다.

1. 서 론

세포내의 유생분자(biomolcule)와 선충류(Caenorhabdtis elegans)에서 뉴런 사이의 시냅스 연결만큼이나 다른 요소들을 묘사하지만, 비슷한 모티프(Motifs)들로 정보처리를 수행하는 네트워크에서 발견할 수 있다. 따라서 모티프들은 네트워크 class를 규정할 수 있고 접근에서도 대부분의 네트워크가 기본 구성조각(building block)을 밝힌다. 자연에서 발생한 많고 복잡한 네트워크들은 보편적인 통계수치[1]을 공유하는 것으로 알려져 있으나 이것들은 어떤 node(노드) 사이의 짧은 경로와 고도로 밀집되어 연결된 "작은 세계" 특성의 소유물[1,2,3]들을 포함하고 있다. 또 자연의 여러 네트워크에는 평균적인 node보다 더 많은 연결을 가진 node들이 존재하기 때문에 "scale-free network"[4,5]라는 네트워크에서 k edges일 때 p(k)를 가지는 node의 분류와 힘의 법칙은 $p(k)\sim kY$로 표현하며 Y는 2 와 3 사이에서 소실된다는 연구[6,7]도 있었다. 이러한 총체적인 수치를 넘어서기 위해서는 각 네트워크 분류에 따른 구성 요소들이 필요하기 때문에 네트워크 모티프를 탐색하는 알고리듬으로 반복되는 내부 연결 패턴을 연구한다. 유전자 조절 네트워크에 대한 상세한 적용법의 내용을 삽입해서[8]나타낸다. 이러한 접근법을 모든 종류의 연결성 그래프에 응용하고 넓은 범위의 자연현상에서 네트워크 모티프가 나타남을 확인한다.

2. 네트워크 모티프 알고리즘

유전학에서 세포내의 유생분자(biomolcule)와 선충류(Caenorhabdtis elegans)의 뉴런 사이의 시냅스 연결만큼이나 다른 요소를 묘사하지만[9], 비슷한 모티프들로 정보처리를 수행하는 네트워크에서도 모티프를 네트워크 class로 규정할 수 있다.

유전자 네트워크에서 반복적으로 발견되는 상호연결 패턴인 네트워크 모티프를 검출하기 위한 알고리듬은 다음과 같다.

1. 유전자에서 전사(Transcription) 네트워크를 연결 메트릭스로 표현한다.
2. 단백질제조에 관여하는 유전자의 한 단위 오페론 j 가 전사 인자를 인코드 하여 오페론 i 을 발현시키는 경우 Mij =1 이나 Mij = 0 으로 표현한다.
3. n=3, n=4 인 경우 연결된 그래프에 놓여 있는 n 개의 노드를 선택하여 n * n개에서 M의 서브메트릭스를 스캔 한다.
4. 서브메트릭스들의 개수는 0 이 아닌 i , j 에리먼트를 리버스로 찾을 때 효과적으로 계산된다.

5. 다음 i 열과 j 행에서 0 이 아닌 앨리먼트를 스캔한다.

위의 알고리듬에서 node 사이의 상호작용이 유도된 edges에 의해 나타나는 네트워크는 그림 1.에서와 같이 A 는 실제 네트워크이고, B 은 임의로 추출된 네트워크로 여기서 "네트워크 모티프"는 임의로 추출된 네트워크에서 A 가 실제 네트워크 B 보다 훨씬 더 자주 반복되는 패턴이다.

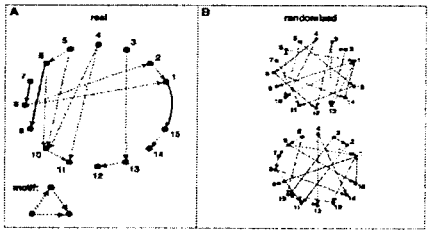
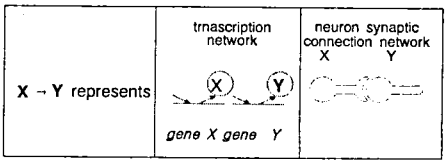


그림 1. 네트워크 모티프 탐색의 개략도.

임의로 추출된 네트워크에서 각 node는 실제 네트워크에 상응하는 node에서와 같은 수가 들어오는 edge와 나가는 edge를 가지기 때문에 점선(빨간색선)은 실행에 옮기기 전에 결함을 예측해 행하는 피드백과정을 제어하는 edges를 나타내며, 실제 네트워크에서는 5배 정도 나타난다. 각 네트워크는 모든 n-node의 서브그래프(subgraph)가 스캔 되면 n=3 과 n=4가 되어 각 서브그래프의 발생 숫자가 기록된다. 각 네트워크는 여러 종류의 n-node 서브그래프를 포함하지만 가장 중요하게 여겨지는 것에 중점을 두기 위해 서 실제 네트워크와 임의로 추출된 네트워크[9,10]을 비교했을 때 선택된 패턴은 그 수에 있어서 실제 네트워크가 임의로 추출된 네트워크 보다 훨씬 더 많았다. 그림 2. A 에서 네트워크 유생분자 비율의 세포 뉴론 X 는 시냅스 적으로 뉴론 Y 에 연결하여 유기체로 되고, X는 오직 Y에 공급 신호 정보위에 전사요소 단백질 X 와 단백질 Y 의 생산율을 조절하는 유전자의 DNA 영역과 정기적으로 묶여진다. 유전자 상호작용에서 네트워크와의 node 사이에서 정확 하고 엄밀한 비교분석을 위해 실제 네트워크와 동일한 하나의 node을 가진 임의의 네트워크를 사용하기 때문에 네트워크에서 각 node

A



B

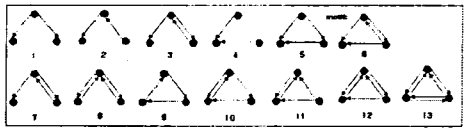


그림 2. (A) 유전자 상호작용에 의해 네트워크 node로 유도된(directed) edge
(B) 13개 3-node 모두 서브그래프 연결

로 실제네트워크에 상응하는 node를 가지는 것과 같은 수가 들어오는 edge와 나가는 edge를 가지게 된다. 그림 2. B 에서도 임의로 추출된 전체적인 효과와 비교할 때 네트워크의 신호 node 특성에 의해 저속 네트워크의 원인이 될 때 나타나는 패턴도 알 수 있다. 여기서 많은 edge를 가진 node가 나타날 때 n-node 서브그래프의 중요도를 계산하기 위해 임의로 추출된 네트워크들은 실제 네트워크에서와 같은 모든 (n-1) - node의 서브그래프에 나타나는 숫자를 보존하기 위해 발생되므로 서브패턴(subpattern)을 가지고 있다. 저속 네트워크 모티프에 나타나는 패턴에서도 edge를 가진 node는 n-node이며 이 node를 서브그래프의 중요도를 계산하기 위해 사용하고, 임의로 추출된 네트워크는 실제 네트워크에서와 같은 (n-1) - node의 서브그래프에 숫자가 발생된다. 이는 중요한 서브패턴에서 네트워크 모티프를 실제 네트워크와 같게 나타내므로 네트워크 확률 p = 0.01이 될 때 컷오프 f 값 보다 낮 은 패턴이 될 수 있다.

3. 조절된 네트워크의 재구성

본 연구에서 생화학적 유전조절(transcriptional gene regulation)과 생태학 및 신경생물학적 뉴론 연결 연산법 등을 정보공학과 WWW (World Wide Web)회로에 적용하는 네트워크 모티프를 표1에서와 같이 전사 네트워크 세포에서 유전형질이 발현되도록 조절하는 생화학적 네트워크와 같이 된다.

이것은 WWW 에 유도된 그래프에서 그림 1. A 와 같이 node들은 유전자를 나타내고 있으며, edge들은 전사 단백질에서 들어오는 신호를 유전자에서 전사요소에 의해 조절하면서 유전자를 향해 유도된다. 또 유전체에 상응하는 두개의 가장 특성 있게 전사하면서 조절하는 네트워크 모티프를 분석한 결과 진한 핵 생물[11]과 박테리아[12,13]들에서 셀과 셀 사이에 연결하는 네트워크를 상응할 수 있다. 이와 같이 전사는 동일한 모티프로 네트워크가 연결되면 명령을 반복적으로 사용 할 때 3-node 모티프

와 "bi - fan"(BIO이행)[14]이라고 불리는 4-node 모티프로 나타난다.

아래 그림 3. A부터 E까지는 임의의 변수로 각각 직접 종속물이 되고, B는 베이스의 네트워크 구조에서 지정된 생산물 형태이며, C는 P(C I A, B) 생산물 형태에서 조건적 분포보인것이며, D는 똑 같은 다섯 개의 변수에 대한 마르코프 네트워크이다. E는 마르코프 네트워크 구조를 야기한 부분집합의 생산물 형태 나타낸 것으로 분배 생산형태를 생산물 특성으로 나타낸 것을 압축해서 추론적으로 학습한다.

유전자 베이스의 네트워크에서 임의의 변수 $X = \{X_1, \ldots X_n\}$은 이음매 분배 조건 확률의 생산물로 베이스의 네트워크에 각 변수 X_i와 조건적 확률 P $(X_i \mid U_i)$와 관련되어 있으므로 $U_i \subseteq X$는

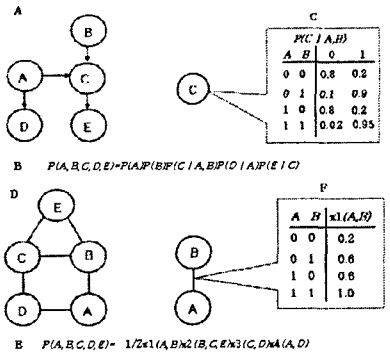


그림 3. 5개 이원(binary)을 임의의 변수에 대한 베이스의 네트워크.

X_i 의 패런트(Parent)로 불리는 변수이다. 패런트값은 X_i 값의 선택에 직접 영향이 있으므로 그 결과의 생산물은 식(1)과 같다.

$$P(X_1, \ldots X_n) = \Pi P(X_i \mid U_i) \qquad (1)$$

그림 3.에서 지시된 그래프에 의해 X_i 는 패런트로부터 X_i 까지 edge를 놓은 것은 그림 A 에서 C 까지 이므로 그래프가 acyclic 이라면 생산물 해체가 될 때 확률 분포가 된다. 유전자 베이스의 네트워크는 몇 개의 생물학 영역에서 자연스럽게 나타날 수도 있지만 계통분석에서 두개의 생물학적 패런트 유전형이 주어질 때 유전자형의 이음매 분배는 유전형 조건부 확률이 된다. 각 변수와 관련된 조건부 확률은 통계적 역행 모델을 사용하기 때문에 $P(X_i \mid U_i)$되고, U_i에 대해 X_i의 선형 역행 모델이 된다. 또 패런트의 값이 주어질 때 이산변량 X_i의 확률을 나타내므로 조건적 확률에서 특정 파라미터의 표현을 선택하는 영역에서 네트워크에 의해서 연결된다. D 에서 F 는 마르코프 네트워크(Markov Network)의 퍼펜셜(potential)의 생산물로서 이음매 분배를 나타내고 있으므로 각 퍼텐셜은 작은 변수 사이에서 상호작용을 찾는 변수들에서 이음매 값을 지정하는 값 P는 다음과 같은 식으로 전의한다.

$$P(X_1, \ldots, X_n) = \frac{1}{Z}\prod_j \pi_j[C_j] \qquad (2)$$

위 식에서 $\pi_i[C_i]$일 때 변수 $C_i \subseteq X$ 에 대한 j 번째 퍼텐셜이 되며 Z는 전체 확률 질량을 1 로 놓을 때 표준치로 본다. 유전자 마르코프의 네트워크는 임의의 변수로 원소 기원을 나타내며 원소 사이의 퍼텐셜은 두개의 원소 사이의 확률로 각각 할당 값을 결정 한다.

표 1.은 유전자규칙(11)과 신경단위(17), 정보처리 도매인(WWW)(7)로 구성된 것으로 확률 분포에서 모티프들은 각각 네트워크와 다양한 생화학적 기능을 수행하는 이체 유전시스템에서 많이 나타난 네트워크이다. 여기서 임의로 추출된 네트워크에 나타나는 평균 숫자는 10의 표준편차 이상 빈도로 나타낸 것이다. 그림 1. B 와 같이 13개에서 가능한 3-node 서프그래프와 188개의 다른 4-node 서브그래프 중에서 네트워크 모티프의 확률이 된다. 다른 3-node와 4-node 서브그래프들이 네트워크를 통해 반복해서 나타나므로 각 네트워크의 node와 edge가 나타날 때도 각 모티프에 대한 실제 네트워크와 임의의 네트워크에서 발생숫자가 나타난다. 임의의 1000개 네트워크 와 비교한 결과 WWW에서 100개가 되므로 모든 모티프 확률 P 값은 P<0.01이 된다. 통계적인 수치로 Z score = (N real-N rand) / SD 로 나와 있다. node의 완전히 다른 세트에서 최소한 U=4회 발생하는 모티프이므로 네트워크는 다음과 같다. 박테리아 E.coli와 효모 S (serevisiae)의 최소 5개 시냅스로 연결된 뉴론을 포함하여 C. elegans는 뉴론 사이의 시냅스 연결 생태학에서의 상호 영양 관계를 원양 생물 및 해저 생물, 조류, 어류 등 무척추동물은 ISCAS89 벤치마크 세트에서 분석된 전기 연속논리 회로와 한 개의 도매인에서 웹 페이지 사이의 인터넷 하이퍼링크 WWW에서

3-node 모티프가 나타날 때 e가 10의 거듭제곱으로 하면 1.45 e 5 = 1.45×10^6 곱으로 된다.

유전학 네트워크에서 연산을 생태학에 적용할 때 node는 각 종의 그룹을 나타내고, edge 은 node로부터 생태학을 나타내는 node로 향한다.표 1.

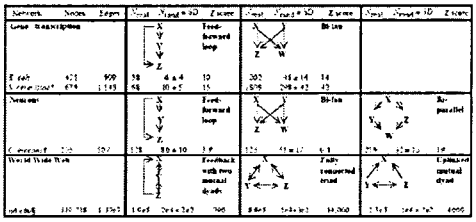


표 1. 생물학적 네트워크와 정보공학적 네트워크에서 네트워크 모티프 추출

에서 3개의 생태학 중 다섯 개가 한 개의 3-node 모티프를 공유하고, 3개 전체가 한 개의 4-node 모티프 확률로 공유한다. 3-node 모티프와는 대조적으로 3-node 앞으로 되돌아오는 loop에서는 생태학에서 많이 나타나지 않는다. 이것은 두 층으로 분리할 시 직접적으로 상호작용에 의해 반대할 수도 있고, 선택 된다. 표 에서 4-node 모티프는 bi-fan과 bi-병렬을 포함한 모티프 확률로 되돌아오는 loop와 bi-fan등 두 개로 전사 유전조절 네트워크가 된다. 모티프 간의 유사성은 두 종류 네트워크의 설계 시 제약을 받는 유사성을 가지고 있다. 두개의 네트워크는 감각 기관소자 신경단위와 생화학적 신호로 조절된 전사요소에서 실행하는 구조적 유전자로 정보를 전달하는 기능이다. 두 네트워크에서 공통적으로 되돌아오는 loop 모티프는 정보처리에서 기능적 역할을 수행하고, 이 회로에서 가능한 한 하나의 입력 시그널이 지속될 때에만 출력을 활성화시키며 입력이 사라지면 신속한 비활성화를 수행한다. 또 신경이 되돌아오는 loop에서 많은 입력과 출력 node는 감각기관 소자로서, 변화하고 시끄러운 환경에서 발생하므로 순간적인 입력의 변동에 대처하기 위해 정보처리가 필요로 한다.

유전학적 네트워크와 전자정보학적 네트워크 회로에서 node들은 논리 게이트와 flip-flop로 나타나기 때문에 노드들은 유도된 edges에 의해 상호연결되므로 네트워크 모티프는 회로의 기능에 따라 class로 분리되는 모티프는 forward logic chip과 feedforward loop, bi-fan, bi-parallel motif로 공유하며, 이것은 유전 정보처리 네트워크와 뉴론 정보처리 네트워크에서도 유사하게 공유된다. 정보처리 도메인에서 WWW의 페이지 사이 지시된 하이퍼링크 네트워크에서 다른 세트의 모티프도 나타날 수 있다. 생태학에서 공유되는 네트워크 모티프는 유전조절 네트워크나 WWW에서 발견되는 모티프와는 결코 연결되지 않는다. 생태학과 일치된 네트워크 중 오직 한 개만이 뉴론 네트워크에서 나타났으며, 다른 모티프 세트가 다른 기능을 정보처리 네트워크회로에서 발견된다. 이는 모티프들이 각각 특정 종류의 기본 구조를 가진 넓은 네트워크 class를 정의할 수 있음을 시사하는 것으로 네트워크 모티프는 각 네트워크 종류를 발생시키는 절차를 반영한다. 예를 들어, 생태학은 바닥에서 꼭대기까지 에너지의 흐름을 이동시키도록 전개되는 반면, 유전자 조절과 뉴론 네트워크에서도 정보처리를 위해 전개되었고, 정보처리는 에너지 흐름을 허용하는 것과 상당히 다른 구조로 전개된다.

그림 4.와 같이 유전자 네트워크에서 다양한 크기의 하부 네트워크들을 고려함으로써 네트워크 크기 기능으로서 모티프의 통계학적 의미를 규정하였고 하부 네트워크에서 모티프에 집중한다. 대조적으로, 하부네트워크에서 임의로 추출된 구조에 상응하는 하부그래프에 집중하는 것은 크기를 급격히 감소시키므로 통계학적 물리현상을 분석함에 있어서, 실제 네트워크에서 각 모티프의 발생 숫자는 광대한 변수 시스템의 크기와 함께 선형으로 변한다. 이러한 변수는 임의로 추출된 네트워크에서는 변수의 존재는 진화되고 설계된 시스템의질을 통합하는 것일지 모르지만 그림 4와 같이 임의로 네트워크 사이즈 S를 가지는 Concentration C의 감소는 C~1/S인 Erdos - Reniy 로 표현할 수 있으며 실제 네트워크의 node와 edge의 수만을 보존하는 임의의 그래프이다. 또 표 1.에서 와같이 네트워크가 커질수록 모티프의 중요성도 커진다는 것을 다른 사이즈의 네트워크와 비교하여 나타난다. 또한 네트워크 모티프 탐지 연산 알고리듬에서 다소 작은 네트워크인 100 edge가 순서에서도 효과적이다. 이는 작은 네트워크에서도 3 내지 4 -node 서브그래프가 대량으로 발생하기 때문에 확률적 접근법은 데이터 에러에도 민감하지 않았다. 네트워크에 무작위로 edge를 20%정도 추가하고 제거하고 재배열하더라도 중요한 네트워크 모티프는 변하지 않았다.

정보처리 네트워크에 있어서 모티프는 기본 계산 회로에서 특정 기능을 가질지 모르나 이것들은 네트워크 전개 시 부딪치는 특별한 제약으로 인해 발생된 구조다. 네트워크와 네트워크 이체동형의 class를 정의하고 네트워크의 동적수행에 대한 통찰력을 얻기 위해 네트워크 모티프를 탐색하여 분석하고,

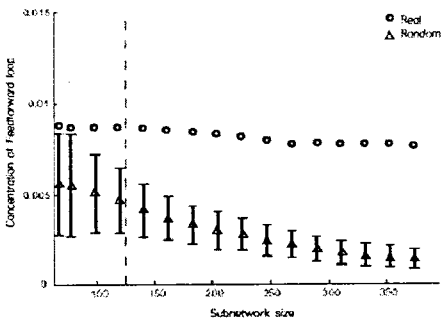


그림 4. 전체적인 네트워크에서 임의로 추출된 하부네트워크 크기.

네트워크 모티프의 접근법은 edges, nodes가 다양한 "색상"을 지닌 네트워크에도 포함한다.

4. 결 론

유전자 세포 간 네트워크를 상호 연결하는 구조를 해석하여 정보 네트워크에 접근하기위해 유전자 네트워크에서 네트워크 모티프를 검출하여 제한된 알고리듬을 통하여 유생분자비율 세포 X, Y의 유기체에서 실제 네트워크 node와 edge를 가진다는 것을 연구하였다.

표에서와 같이 유전자 네트워크와 기술적 네트워크에서 네트워크 모티프를 증명하고 감각기관 소자 신경단위와 생물학적 신호로 조절된 전사요소에서 실행하는 구조적 유전자로 정보를 전달하는 기능으로 보았다.

유전자 정보처리 네트워크와 뉴론 정보처리 네트워크 도매인 WWW페이지 사이를 유사하게 지시된 하이퍼링크 네트워크에서 네트워크 모티프를 나타냈다.

통계학적 물리현상에서 실제 네트워크 모티프의 발생 숫자가 광대한 변수였으며 이 변수를 임의로 알고리듬을 통해 추출된 네트워크에서 모티프가 광대하지 않았다. 또 실제 네트워크 모티프의 발생숫자가 광대한 변수였기 때문에 변수에서 임의로 추출된 네트워크에서 모티프가 광대하지는 않았다.

결과적으로 유전체 네트워크와 상호작용하여 많은 데이터를 서로 처리하기 위해 다양한 데이터의 양을 정보처리가 될 수 있으므로 동적 수행에서 통찰력을 얻기 위해 네트워크 모티프를 탐색하고 네트워크 전개에서 주어진 모티프를 분석했다. 나아가서 네트워크 모티프를 앞으로 충족시키기 위해서 새로운 실험분석, 실험디자인, 전체적인 유기체, 그리고 사회전체 수준에서 시스템이 포괄적으로 응용되게 연구해야 한다.

참 고 문 헌

[1].B, Bollobas, Random Graphs (Academic, London, 1985).

[2].A. -L Barabasi, R. Albert, Science286, 509 (1999).

[3].H. Jeong, B. Tombor, R. Albert, Z. N. Oltval, A. A. Barabasi, Nature 407, 651 (2000).

[4].R. F. Cancho, R. V. Sole, Proc. R. Soc. London Ser. B 268, 2261 (2001).

[5].B. Hubemasn, L. Adamic, Nature 401, 131 (1999).

[6].P. Holland, S. Leinhardt, in Sociological Methodology, D. Heise, Ed. (Jossey-Bass, San Francisco, 1975), pp. 1-45.

[7]. N. Guelzim, S Bottani, p. Bourgine, F. Kepes. Nature Genet. 31. 60 (200).

[8]. S. Maslov, K. Sneppen, Science 296, 910 (2002).

[9]. Methods are available as supporting material on Science Online.

[10]. M. C. Costanzo et al., Nueleic Acids Res. 29. 75 (2001).

[11]. R. Williams, N. Matinez, Nature 404, 180 (2000).

[12]. J. White, E. Southgate, J. Thomson, S. Brenner.Philos. Trans. R. Soc. London Ser. B 314. 1 (1986).

[13].In Erdos-Renyi randomized networks with a fixed connectivity (2). the concentration of a subgraph with n nodes and k edges scales with network size as $C \sim S^{n-k-1}$(thus, C ~ 1/5 for the feedforward loop of Fig. 3 where n = k = 3). The Sole exception in Table 1 in which C should not vanish at large S is the three-chain pattern in food webs where n = 3 and k = 2.