유전자 네트워크 모티프 알고리즘을 이용한 인터넷 네트워크 분석

나하선, 김문환, 나상동*
한국전기기계(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 상호연결하는 "네트워크 모티프"를 연구한다. 다양한 복잡한 유전학적 네트워크 구조의 설계 및 원리에 해석한 네트워크 모티프를 편리하게 규정할 수 있는 높은 모티프를 통해 해석한다. 유전학적 네트워크에서 모티포드의 네트워크의 보편적 class를 규정할 수 있는지 이론적으로 접근하여 인터넷 네트워크에 응용 및 해석하고 분석한다.

1. 서 론


2. 네트워크 모티프 알고리즘

유전학적 세포내의 유생분자(biomolecule)와 선충류(Caenorhabditis elegans)의 구조 사이의 시냅스 연결만들이나 다른 요소를 묶어지지만[9], 비슷한 모티포드로 정보처리를 수행하는 네트워크에서 모티포드를 네트워크 class로 규정할 수 있다. 유전자 네트워크에서 반복적으로 발견되는 상호 연결 패턴의 네트워크 모티프를 검출하기 위한 알고리즘은 다음과 같다.

1. 유전자에서 전사(Transcription) 네트워크를 연결 매트릭스로 표현한다.
2. 단백질제조에 관여하는 유전자들 간 단위 오해를 j 가 전사 인자들을 인코드 하여 오해된 i 을 발생시키는 경우 Mij = 1 이나 Mij = 0 으로 표현한다.
3. n=3, n=4 인 경우 연결된 그래프에 놓여 있는 개의 노드를 선택하여 n * n개에서 M의 서브매트릭스를 스캔 한다.
4. 서브매트릭스들의 개수는 0 이 아닌 i , j 에러먼트를 림버스로 찾을 때 효과적으로 계산된다.
5. 다음 1열과 2행에서 0 이 아닌 엘리먼트를 삽입한다.

위의 알고리즘에서 node 사이의 상호작용이 유도된 edges에 의해 나타나는 네트워크는 그림 1에서와 같이 A는 실제 네트워크이고, B는 임의로 추출된 네트워크로 여기서 ‘네트워크 모티프’는 임의로 추출된 네트워크에서 A 가 실제 네트워크 B 보다 흔한 더 자주 반복되는 패턴이다.

![그림 1. 네트워크 모티프 탐색의 개념도.](image)

임의로 추출된 네트워크에서 각 node는 실제 네트워크에 상응하는 node와 같은 수가 들어오는 edge와 나가는 edge를 가지기 때문에 점선(별간선)은 실제 올기는 전에 결합을 예측해 볼 때 드러나는 edge를 나타내며, 실제 네트워크에서는 5개 정도 나타난다. 각 네트워크는 모든 n-node의 서브그래프(subgraph)가 스탠 되면 n=3 과 n=4가 되어 각 서브그래프의 발생 숫자가 기록된다. 각 서브그래프는 여러 종류의 n-node 서브그래프를 포함하지만 가장 유의미하게 여겨지는 것에 중점을 두기 위해 실제 네트워크와 임의로 추출된 네트워크[9,10]를 비교했을 때 선택된 패턴은 그 수에 있어서 실제 네트워크가 임의로 추출된 네트워크보다 흔한 더 많았다. 그림 2 A에서 네트워크 유사성 비율의 세포 뉴런 X는 시냅스 적으로 뉴런 Y에 연결하여 유기체로 되고, X는 오직 Y에 공급 사용 정보에 전사소 단위적 2 단계 및 Y의 생산성을 조절하는 유전자의 DNA 영역과 정 기적으로 통한다. 유전자 상호작용에서 네트워크와의 node 사이에서 합당하고 광범위한 비교분석은 위해 실제 네트워크와 동일한 하나의 node를 가진 임의의 네트워크를 사용하기 때문에 네트워크에서 각 node

<table>
<thead>
<tr>
<th>X - Y represents</th>
<th>transcription network</th>
<th>neuron synaptic connection network</th>
</tr>
</thead>
<tbody>
<tr>
<td>gene X gene Y</td>
<td>X</td>
<td>Y</td>
</tr>
</tbody>
</table>

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

서브그래프에 상응하는 node를 가지는 것과 같은 수가 들어오는 edge와 나가는 edge를 가지게 된다. 그림 2 B에서도 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 네트워크와 같은 모든 (n-1) - node의 서브그래프에 나타나는 숫자를 보존하기 위해 발생되므로 서브패턴(subpattern)을 가지고 있다. 저속 네트워크 모티포에서 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 네트워크모티프를 실제 네트워크와 같이 나타내므로 네트워크 확률 p=0.01이 될 때 것으로 f 값 보다 낮은 패턴이 될 수 있다.

3. 조절된 네트워크의 재구성

와 “bi-fan”(BIO이행[14])이라고 불리는 4-node 모양으로 나타난다.

이제 그림 3. A부터 E까지는 임의의 변수로 각각 직접 종속들이 되고, B는 베이스의 네트워크 구조에 지정된 생산물 형태이며, C는 P(C I A, B) 생산물 형태에서 조건적 분포표현형이며, D는 독립한 다섯 개의 변수에 대한 마르코프 네트워크이다. E는 마르코프 네트워크 구조를 가진 기본적 질병의 생산물을 형성 나타낸 것으로 문제 생산형태를 생성물 복잡성으로 나타낸 것을 압축하여 추론적으로 학습한다.

유전자 베이식의 네트워크에서 임의의 변수 \( X = \{X_1,...,X_n\} \)은 이에 대해 분해 조건 확률의 생산물로 베이식의 네트워크에 각 변수 \( X_i \)와 조건적 확률 \( P(X_i | U_i) \)이 관련되어 있으므로 \( U_i \subseteq X \)는 유전형 조건부 확률이 된다. 각 변수와 관련된 조건부 확률은 통계적 역행 모델을 사용하기 때문에 \( P(X_i | U_i) \)이고, \( U_i \)에 대해 \( X_i \)의 선형 역행 모델이 된다. 또 피렌트의 값이 주어질 때 이산형량 \( X_i \)의 확률을 나타내므로 조건적 확률에서 특정 피렌트의 표현을 선택하는 영역에서 네트워크에 의해 연결된다. D에서 F는 마르코프 네트워크의 피렌트(potential)의 생산물로서 이에 대해 분해를 나타내고 있으므로 각 피렌트들은 각 변수 사이에서 상호작용을 갖는 변수들 사이에 이용할 값을 지정하는 값 \( P \)은 다음과 같은 식으로 정의한다.

\[
P(t_{X_i}, ..., t_{X_n}) = \prod_{i=1}^{n} P(t_{X_i} | C_i)
\]

위 식에서 \( t_{C_i} \)일 때 변수 \( C_i \subseteq X \)에 대한 \( j \) 번째 피렌트의 확률은 \( Z \)은 전체 확률 분포를 \( 1 \)로 놓고襞 표준지로 본다. 유전자 마르코프의 네트워크는 임의의 변수로 원소 기전을 나타내며 원소 사이의 피렌트들은 두개의 원소 사이의 확률로 각각 할당 값을 결정한다.

표 1은 유전자피렌트(11)의 선정단위(17), 정보처리 도메인(WWW)(7)로 구성된 것으로 확률 분포에서 모티프들은 각각 네트워크와 다양한 생활적 기능을 수행하는 이와 유전시스템에서 많이 나타난 네트워크이다. 여기서 임의로 추출된 네트워크에서 나타나는 평균 숫자는 10의 표준편차 이상 빈도로 나타낸 것이다. 그림 1, 2, 3과 같이 13개에서 가능한 3-node 서브그래프의 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 네트워크의 임의의 네트워크에서 발생단자가 나타난다. 임의의 1000개 네트워크와 비교한 결과 WWW에서 100개가 되므로 모든 모티프 확률 P 값은 P<0.01이 된다. 통계적 수치로 Z score = (N real-N rand) / SD로 나와 있다. node의 완전히 다른 세트에서 최소한 U=4 회 발생하는 모티프로서 네트워크는 다음과 같다. 박테리아 Ecoli와 요소 S (serovia)의 최소 5개 시험으로 연결된 뉴론을 포함하여 C. eileatus는 뉴론 사이의 시험률 연결 생태학에서의 상호 영향 관계를 원활 생물 및 환경 생물, 조류, 여우 등 무척추동물은 ISCS589 벤치마크 네트워크에서 분석된 경우 연속생리 최적화 및 몇 개의 도메인에서 월 드레자의 사이의 인터넷 하이퍼링크 WWW에서.
유전자 네트워크 모티프 알고리즘을 이용한 인터넷 네트워크 분석

3-node 모티프가 나타날 때 $e$가 10의 격등계급으로 하면 $1.45 \times 10^2$으로 된다.

유전학 네트워크에서 연산을 생태학에 적용할 때 node는 각 종의 그룹을 나타내고, edge는 node로부터 생태학을 나타내는 node로 한다. 표 1.

<table>
<thead>
<tr>
<th>노드</th>
<th>에דורальные 네트워크와 정보공학의 네트워크에서 모티프 추출</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

에서 3개의 생태학 중 다섯 개가 한 개의 3-node 모티프를 공유하고, 3개 전체가 한 개의 4-node 모티프 핵심으로 공유한다. 3-node 모티프는 대조적으로 3-node 앞으로 되돌아오는 loop는 생태학에서 많이 나타나지 않는다. 이것은 두 종으로 분리할 시 직접적으로 상호작용에 의해 반대할 수도 있고, 선택 된다. 표 4에 4-node 모티프는 bi-fan과 bi-bridge 패턴을 포함한 모티프 핵심으로 되돌아오는 loop와 bi-fan 등 두 개로 전사 유전조건 네트워크가 된다. 모티프 간의 유사성은 두 종류 네트워크의 설계 시 제약을 받는 유사성을 가지고 있다. 두개의 네트워크는 경사 기판조차 시각화의 생체학적 신호로 조절된 전사요소에서 설계하는 구조적 유전자의 정보를 전달하는 기능이다. 두 네트워크에서 공통적으로 되돌아오는 loop 모티프는 정보처리에서 가능성을 활용하고, 이 화학에서 가능한 한 하나의 입력 시그널이 지속될 때마다 출력을 영향시키며 입력이 사라지면 신속한 비향상화를 수행한다. 또한 신경 모티포이 loop는 많은 입력과 출력 node는 감각기관 소자로서, 변화하고 시각적 환경에서 발생하므로 순환적인 입력의 변동에 대처하기 위해 정보처리가 필요로 한다.

유전학적 네트워크와 전자정보학적 네트워크 회로에서 node들은 논리 게이트와 flip-flop으로 나타나기 때문에 노드들은 유도된 edges에 의해 상호작용으로 네트워크 모티프는 회로의 기능에 따라 class로 분류되는 모티포는 feedforward loop, bi-fan, bi-parallel motif로 공유하며, 이것은 유전 정보처리 네트워크와 뉴런 정보 처리 네트워크에서도 유사하게 공유된다. 정보처리 도메인에서 WWW의 페이지 사이지시된 아이피
크 네트워크에서 다른 네트워크의 모티포가 나타날 수 있다. 생태학에서 공유되는 네트워크 모티포는 유전 조절 네트워크와 WWW에서 발견되는 모티포와는 상호 연결되지 않는다. 생태학과 일치된 네트워크 중 오직 한 개만이 뉴런 네트워크에서 나타났으며, 다른 모티포 세트가 다른 기능을 정보처리 네트워크 회로에서 발견된다. 이는 모티포들이 각각 특정 종류의 같은 구조를 가진 네트워크 class를 정의할 수 있음을 시사하는 것으로 모티프는 각 네트워크 구조를 반영시키는 정차를 반영한다. 예를 들어, 생태학의 바닥에서 목대가기 뿐만 아니라 유지의 효과를 이용시키기로 전개되는 반면, 유전자 조절과 뉴런 네트워크에서도 정보처리를 위해 전개되었고, 정보처리는 유지의 효과를 허용하는 것과 상당히 다른 구조로 전개된다.

그림과 같이 유전자 네트워크에서 다양한 크기의 핵부 네트워크를 고려함으로써 네트워크 크기가 가능함으로써 모티프의 동적학적 의미를 규정하고 하부 네트워크에서 모티포에 집중한다. 대조적으로, 하부네트워크에서 입력으로 주어진 구조에 상응하는 하부그램에 집중하는 것은 크기를 급격히 감소시키므로 통계학적 물리현상을 분석함에 있어서, 실제 네트워크에서 각 모티포의 발생 수는 광범한 변수 시스템의 크기에 의해 선형으로 변한다. 이러한 변수는 임의로 추출된 네트워크에서는 변수의 존재는 전달되고 실제 시스템의 통계는 모티포가 많은 것치 모양과 비슷하다. 이러한 모티포는 모티포가 전자 정보학의 노드와 edge의 수만을 보존하는 입력의 그래프이다. 또 표 1에서 왼쪽 네트워크가 커질수록 모티포의 중요성도 커진다는 것을 다른 사이즈의 네트워크와 비교하여 나타낸다. 또한 네트워크 모티포가 전자 정보학의 노드와 edge의 수만을 보존하는 입력의 그래프이다. 이는 작은 네트워크에서도 정보처리 네트워크의 모티포는 변하지 않아도 된다. 정보처리 네트워크의 모티포는 기본 계산 회로에서 특정 가능은 가정치 모로나 이것들을 네트워크 전체 시 작지치는 특별한 계약으로 인해 발생된 구조다. 네트워크와 네트워크 이중형의 class 평가보고 네트워크의 동적수행에 대한 동작을 얻기 위해 네트워크 모티포를 탐색하여 분석하고,
네트워크 모티프의 접근법은 edges, nodes가 다양한 "색상"을 지닌 네트워크에도 포함한다.

4. 결 론
유전자 세포 간 네트워크를 상호 연결하는 구조를 해석하여 정보 네트워크에 접근하기 위해 유전자 네트워크에서 네트워크 모티프를 경유하여 제한된 알고리즘을 통하여 유전자 처리 설정 X, Y의 유기체에서 실제 네트워크 node와 edge를 가진다는 것을 연구하였다.

표에서와 같이 유전자 네트워크와 기술적 네트워크에서 네트워크 모티프를 증명하고 각각기관 소자 신경단위와 생물화학 신호로 조절한 전사요소에서 실험하는 구조적 유전자 정보를 전달하는 기능으로 보았다.

유전자 정보처리 네트워크와 뉴론 정보처리 네트워크에서 인터넷, WWW 사이트 같은 주요하게 지시된 하이퍼링크 네트워크에서 네트워크 모티프를 나타내었다.

통계학적 물리현상에서 실제 네트워크 모티프의 발생, 전도가 흔한 변수였으며 이 변수들로부터 알고리즘을 통해 추출된 네트워크에서 모티프가 흔히 나타나고 있다. 실제 네트워크 모티프의 발생, 전도가 흔한 변수였기 때문에 전사에서 2차로 추출된 네트워크에서 모티프가 쌍대화하는 procurement와 결과적으로 유전체의 신호작용이 많은 데이터를 서로 처리하기 위해 다양한 데이터의 양을 정보처리가 될 수 있으므로 동적 수행에서 통찰력을 얻기 위해 네트워크 모티프를 탐색하고 네트워크 전체에서 주어진 모티프를 분석하였다. 나아가서 네트워크 모티프를 앞으로 충족시키기 위해서 새로운 실험분석, 실험디자인, 전세계 급조간 유기체, 그리고 사회 전체 수준에서 시스템이도물적으로 응용에 겪어 연구해야 한다.

참 고 문헌
[9] Methods are available as supporting material on Science Online.
[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 ~ 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.