KEGG 패스웨이 네트워크 동적 구축 및 클러스터링 시스템 개발

Implementing System for Dynamic Constructing and Clustering on KEGG Pathway Network

서 동 민, 이 민 호, 유 석 종
한국과학기술정보연구원 생명의료HPC연구센터

Dongmin Seo, Min-Ho Lee, Seok Jong Yu
Biomedical HPC Research Center, Korea Institute of Science and Technology Information

요약

최근 유전체학, NGS(Next Generation Sequencing) 기술, IT/NT 장비의 발전 등에 따라 방대한 양의 바이오-메디컬 데이터가 생산되고, 이에 따라 빅데이터를 활용한 헬스케어 산업이 급속히 발달하고 있으며, 이와 관련된 빅데이터 기술은 국민의 건강 증대와 건강한 고령 삶을 제공하는 핵심 기술로 급부상하고 있다. 패스웨이는 단백질, 유전자, 세포 등의 생체적 요소 간의 역학관계 혹은 상호작용 등을 네트워크 형식으로 표현한 생물학적 심층지식으로, 바이오-메디컬 빅데이터 분석에 있어서 널리 활용되고 있다. 하지만 패스웨이는 매우 다양한 형태를 갖고 용량이 매우 큰 빅데이터로 이를 분석하는데 많은 시간이 소요된다. 그래서 본 논문에서는 세계적으로 가장 우수하고 방대한 양의 패스웨이를 제공하는 KEGG 패스웨이 데이터베이스로부터 사용자가 관심 갖는 패스웨이만을 자동 수집하고 패스웨이 간 계층구조를 기반으로 네트워크를 구성 후, 해당 패스웨이 네트워크에 대한 클러스터링과 핵심 패스웨이 선정을 통해 패스웨이 간의 역학관계 또는 상호작용을 직관적으로 분석할 수 시스템을 제안했다.

Ⅰ. 서론

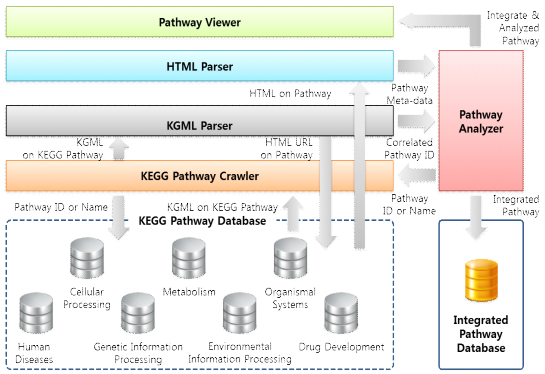
최근에는 유전체학의 발전, 웨어러블 디바이스의 확산, IT/NT의 발전 등에 따라 방대한 양의 바이오-메디컬 데이터가 생산되고, 이에 따라 빅데이터를 활용한 헬스케어 산업이 급속히 발달하고 있으며, 이와 관련된 빅데이터 기술은 국민의 건강 증대와 건강한 고령 삶을 제공하는 핵심 기술로 급부상하고 있다. 일례로, 미국 국립보건원(NIH)는 1,000 Genomes Project를 통해 획득한 200TB의 유전자 정보를 아마존 웹 서비스를 통해 난치병 및 불치병과 관련된 유전자 정보를 분석하는 연구자들에게 제공하고 있다[1][2]. 미국 퇴역 군인국(VA, U.S. Department of Veterans Affairs)은 퇴역 군인들에 대한 DNA 샘플과 전자 의료 기록(EHR, Electronic Health Records)에 대한 분석을 통해 퇴역 군인들에게 맞춤형 의료 서비스를 지원하고 있다[3]. 또한, 캐나다 온타리오 공과대학병원은 신생아의 혈압, 체온, 심전도, 혈중산소 포화도 등의 데이터를 분석하여 미숙아에 대한 패혈증, 폐결핵과 같은 심각한 병원균 감염을 조기 판단하는데 활용하고 있다[1].

패스웨이(Pathway)는 기술문헌에 출현한 다양한 전문용어와 그들 간의 의미적 상관관계를 네트워크 형식으로 표현한 자료구조로서, 생명공학 관점에서는 단백질, 유전자, 세포 등의 생체적 요소 간의 역학관계 혹은 상호작용 등을 세밀하게 기술한 생물학적 심층지식이다. 하지만 패스웨이는 매우 다양한 형태를 갖고 용량이 매우 큰 빅데이터로 이를 분석하는데 많은 시간이 소요되며, 현재까지도 다양한 패스웨이를 다차원 분석할 수 있는 시스템은 전무하다. 그래서 본 논문에서는 세계적으로 가장 우수하고 방대한 양의 Pathway를 제공하는 KEGG (Kyoto Ency clopedia of Genes and Genomes)[4]의 패스웨이 데이터베이스에 대한 다차원 분석 및 가시화 기능을 제공하는 시스템을 제안했다.

Ⅱ. 제안하는 패스웨이 다차원 분석 시스템

제안하는 패스웨이 다차원 분석 시스템은 그림 1과 같이 KEGG Pathway Crawler, KGML Parser, HTML Parser, Pathway Analyzer, Integrated Pathway Database 그리고 Pathway Viewer로 구성되어 있다. KEGG Pathway Crawler는 입력받은 Pathway ID 또는 Name에 대한 패스웨이 KGML(KEGG Markup Language) 파일을 KEGG 패스웨이 데이터베이스로부터 수집한다. KGML은 패스웨이를 구성하는 단백질, 유전자, 세포와 생체적 요소 간의 역학관계 혹은 상호작용을 XML로 표현한 언어이다. KGML Parser는 KEGG Pathway Crawler로부터 수집된 KGML에서 역학관계 또는 상호작용하는 패스웨이들에 대한 모든 링크를 획득한다. KGML에서 type="map"을 속성으로 갖는 entry 요소는 패스웨이를 나타내고, 해당 패스웨이에 대한 메타 정보를 나타내는 HTML 문서의 URL은 link 속성을 통해 획득할 수 있다. 패스웨이에 대한 HTML 문서는 해당 패스

웨이에 대한 ID, Name, Class, KGM URL, Description과 해당 패스웨이와 관련된 Disease, Drug, Gene 그리고 이와 같은 정보를 획득한 Reference Authors, Title, Journal 등의 패스웨이 관련 메타 정보를 제공한다. HTML Parser는 KGML Parser로부터 수집된 HTML URL로부터 HTML 문서를 수집하고, 수집한 문서로부터 패스웨이와 관련된 메타 정보와 해당 패스웨이에 대한 KGML을 수집할 수 있는 URL을 수집한다. Integrated Pathway Database는 HTML Parser로부터 수집된 패스웨이에 대한 메타 정보와 수집된 패스웨이들에 대한 계층(Parent-Child) 구조 정보를 저장한다.



▶▶ 그림 1. 제안하는 패스웨이 다차원 분석 시스템 구성도

Pathway Analyzer는 수집된 패스웨이들에 대한 클러스터링 기능과 클러스터 그룹 내 핵심 패스웨이 선정 기능을 제공한다. 클러스터링 기능은 패스웨이 간의 역학 관계 또는 상호작용이 다른 패스웨이들에 비해 밀접한 패스웨이들만을 동일 그룹으로 편성하는 것으로 대용량 패스웨이를 효율적으로 분류하는데 활용할 수 있다. Pathway Analyzer가 지원하는 클러스터링 기법은 식 1과 같은 Modularity를 기반으로 한 대용량 그래프 클러스터링 기법 [5]을 패스웨이에 맞게 적용했다. 또한, 핵심 패스웨이 선정 기능은 한 클러스터 그룹 내에서 중요도가 높은 패스웨이를 선정하는 것으로 선정 기준은 그림 2와 같다.

$$Q(C) := \sum_{C \in C} \left(\frac{f(C,C)}{f(V,V)} - \frac{\deg(C)^2}{\deg(V)^2} \right) \qquad \text{(식 1)}$$

```
Qv = (Num of nodes connected with a node / Num of all nodes in a cluster) × 100
Qc = User threshold
ArrayList Results

CN = Num of all clusters
while (CN)
    while (V ≤ Num of all nodes in a cluster)
        // VC = Num of connections on other clusters
        if (Qv ≥ Qc && VC == CN) Results.add(V)
        else if (Qv ≥ Qc) Results.add(V)
        else if (VC == CN) Results.add(V)
        V++
    END while
    if (Results.size() > 0) return Results
    CN--;
END while
```

▶▶ 그림 2. 클러스터 그룹 내 핵심 패스웨이 선정 기준

Ⅲ. 성능 평가

표1은 알츠하이머(has05010) 패스웨이를 기준으로 deg(3)에 포함되는 모든 패스웨이를 수집 후, 클러스터 그룹 수를 4로 설정하고 임계값 변경에 따른 클러스터링과 핵심 패스웨이가 선정된 결과를 보여준다. $\theta_U = 100$는 핵심 패스웨이로 선정되기 위해서는 동일 클러스터 그룹 내 모든 패스웨이들과 연결되는 것을, $\theta_U = 80$는 동일 클러스터 그룹 내 80% 이상의 다른 패스웨이들과 연결되는 것을 의미한다. 그래서 $\theta_U = 100$으로 설정되었을 때 핵심 패스웨이가 선정되지 않았던 클러스터 그룹 G3에서도 Calcium signaling pathway가 핵심 패스웨이로 선정되었고 클러스터 그룹 G0에서는 $\theta_U = 80$을 만족하면서 다른 클러스터들과 가장 많은 연결을 갖는 Apoptosis가 핵심 패스웨이로 선정되었다. 이와 같이, 사용자는 수집된 패스웨이들에 대한 클러스터링과 핵심 패스웨이 선정을 통해, 패스웨이 간의 역학관계 또는 상호작용을 직관적으로 분석할 수 있게 되었다.

표 1. θ에 따른 핵심 패스웨이 선정 결과

	$\theta_U = 100$	$\theta_U = 80$
G0	MAPK signaling pathway	Apoptosis
G1	NF-kappa B signaling pathway	NF-kappa B signaling pathway
G2	PI3K-Akt signaling pathway	PI3K-Akt signaling pathway
G3	-	Calcium signaling pathway

Ⅳ. 결론 및 향후연구

본 논문에서 개발한 시스템은 KEGG 네트워크 빅데이터를 기반을 둔 다양한 질병의 시범 연구 환경을 마련하는데 초석이 될 것이라 기대한다. 향후 연구로는 대용량 패스웨이 네트워크 내에서 사용자가 관심을 갖는 서브 네트워크 및 유사 네트워크를 효율적으로 탐색하는 기능을 추가할 계획이다. 또한, 오픈 소스 기반의 바이오 네트워크 분석 시스템이 있다면 그들과의 성능 평가를 통해 본 논문에서 개발한 시스템의 우수성을 보다 명확하게 입증할 계획이다.

■ 참 고 문 헌 ■

[1] 백인수, 박지혜, "데이터 시대: 데이터 분석의 중요성", IT&Future Strategy, 제9호, pp.12, 2013.
[2] http://1000genomes.org
[3] http://www.va.gov/health/
[4] http://www.genome.jp/kegg/pathway.html
[5] L. P. Cordella, P. Foggia, C. Sansone, M. Vento, "An Improved Algorithm for Matching Large Graphs", 3rd IAPR-TC15 Workshop on Graph-based Representations in Pattern Recognition, Cuen, pp. 149-159, 2001.