온톨로지를 이용한 단백질 상호작용 네트워크의 개념화

최재훈[0], 박선희
한국전자통신연구원
{jhchoi[0],shp}@etri.re.kr

An Ontology Based Approach for Conceptualizing Protein Interaction Networks

Jae-Hun Choi[0], Seon-Hee Park
Electronic Telecommunication Research Institute(ETRI)

요 약

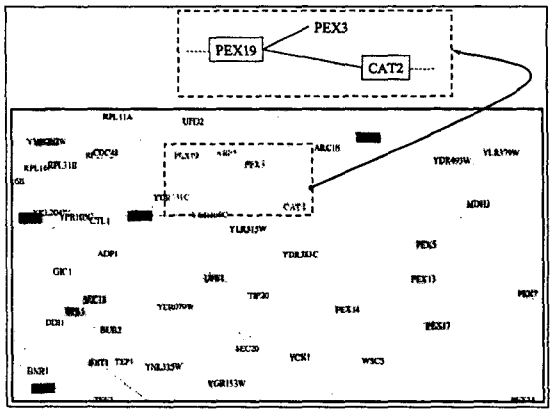
본 논문에서는 생물체의 세포에 존재하는 방대한 단백질들 사이의 복잡한 상화작용 관계 네트워크를 개념화하기 위한 방법을 제안한다. 일반적으로 하나의 단백질은 세포의 특정한 구성요소로서 몇 개의 생물학적 작용에 참여하며 고유의 분자 기능을 수행하게 된다. 즉, 하나의 상호작용 관계 네트워크에 포함된 각각의 단백질들은 구성요소(Cellular Component), 생물학적 작용(Biological Process), 그리고 분자 기능(Molecular Function) 3가지 특징으로 개념화할 수 있다. 또한, 비슷한 특징으로 개념화되는 단백질들은 서로 클러스터링될 수 있기 때문에 단백질 상호작용 네트워크를 일반적인 의미의 개념 네트워크로 표현할 수 있다. 여기서, 단백질 특징을 개념화하기 위해 사용되는 표준 개념과 이 개념들 사이의 관계를 정의하는 유전자 온톨로지(Gene Ontology)가 이용된다.

1. 서론

휴먼 지놈 프로젝트 이후 많은 생물체의 염색체에 대한 기본 서열이 밝혀지고 있다. 이 서열 중에 일부는 후손에게 유전되는 부분인 유전자를 포함하고 있다. 단백질은 이 유전자가 전사 및 번역된 생성물질이며, 세포의 특정한 구성 요소로서 생물학적 역할을 할 수 있는 고유의 분자적 기능을 수행한다.

이 단백질들은 서로 상호작용을 하면서 특정한 역할을 수행하며, 이들을 일반적으로 단백질 상호작용 관계라고 정의한다. 하나의 세포 내에는 많은 단백질들 사이의 상호작용 관계들이 존재하며, 이들은 그래프 형태의 관계 네트워크로 명시될 수 있다. 즉, 단백질은 노드 그리고 단백질 사이의 관계는 링크로 표현될 수 있다[1]. 예를 들어, [그림 1]은 [4]에서 제시한 이스트(yeast) 단백질 2358개에 대한 매우 복잡한 상호작용 관계 네트워크의 일부분을 나타내고 있다. 여기서, 이스트가 가지는 단백질 PEX19, PEX3 그리고 CAT2은 노드로 나타내고 있으며, 단백질 PEX19은 CAT2 그리고 PEX3과 각각 상호작용 관계를 가지고 있기 때문에 이들은 노드들 사이의 링크로 표현하고 있다.

특정 생물체에서 단백질들 사이의 상호작용 관계는 일반적으로 이스트 투 하이브리드(yeast two hybrid)라는 생물학적 실험을 통해 추출되고 있다[2]. 이 실험에서 하나의 단백질(bait protein)를 보고 유전자(reporter gene)의 프로모터와 결합할 수 있는 DNA 결합 부위(DNA binding domain)을 갖도록 발현시키고, 다른 단백질(prey protein)을 이 보고 유전자를 발현시킬 수 있는 전사 활성화 부위(transcription activating domain)를 갖도록 발현시킨다. 만약, 두 단백질의 상호작용을 하게 된다면 베이트 단백질의 DNA 결합 부위가 보고 유전자의 프로모터와 결합하게 되면서 동시에 프레이 단백질의 전사 활성화 부위가 보고 유전자를 발현시키게 될 것이다.

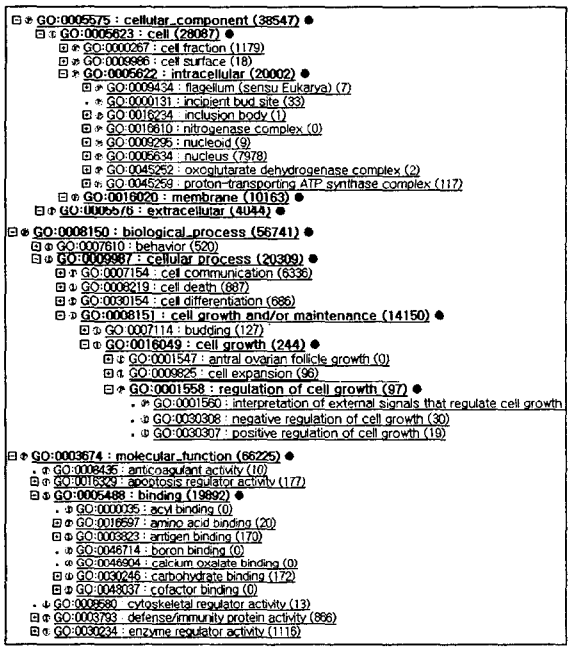


[그림 1] 이스트 단백질들에 대한 상호작용

현재, 이 실험을 통해 구축된 단백질 상호작용 관계 정보는 데이터베이스에 체계적으로 관리되고 있으며, 대표적인 데이터베이스로 PIM(Protein Interaction Map database), BIND(Biological Interaction Network Database), DIP(Database of Interacting Protein), GRID(General Repository for Interaction Datasets) 등이 있다.

단백질 상호작용 네트워크는 단백질의 기능을 유추하는데 매우 중요하게 이용된다 [5]. 즉, 상호작용을 하는 두 단백질들은 일반적으로 서로 연관된 기능을 가지고 있을 것이라고 예측할 수 있다. 이 예측은 신약 물질 개발을 위한 목표 단백질 선정에 이용되고 있다. 그러나, 이 네트워크는 매우 방대하여 사용자가 이 내용을 전체적으로 이해하기는 매우 어렵다[3]. 또한, 네

트워크 노드들을 단백질 이름으로 레이블링하기 때문에 매우 복잡한 형태로 표현되고 있다. 따라서, 네트워크 노드들을 개념으로 레이블링하고 유사한 의미의 노드들을 서로 클러스터링 함으로써 네트워크를 보다 일반적인 의미의 형태로 제공할 수 있는 네트워크 개념화에 대한 요구들이 빈번하게 발생하고 있다. 본 논문에서는 유전자 온톨로지를 이용하여 이 네트워크를 개념화하는 방법을 제안한다.



[그림 2] 온톨로지 표현

2. 단백질 상호작용 네트워크 표현

이장에서는 개념화를 위해 필요한 유전자 온톨로지와 단백질 상호작용에 대한 표현 방법을 설명한다. 본 논문에서 사용하는 유전자 온톨로지는 단백질의 3가지 특징을 기술할 수 있는 제어 용어(Controlled Vocabulary)들에 대한 구조화된 개념 계층 형태로 표현된다. 즉, 유전자 온톨로지는 단백질의 3가지 측면인 세포 구성요소(CC: Cellular Component), 생물학적 작용(BP: Biological Process), 그리고 분자 기능(MF: Molecular Function)에 대한 각각의 3개의 온톨로지로 구성된다. 이 온톨로지를 구성하는 제어 용어들은 생물 종에 독립적인 개념들로 표현된다.

하나의 단백질에 대해 MF 온톨로지는 분자 수준의 행위에 대한 개념들의 관계를 나타낸다. CC는 단지 세포 내에서 이 단백질이 참여하는 구성 요소들에 대한 개념들의 관계를 표현하고 있다. BP 온톨로지는 하나 이상의 분자 수준의 행위들이 결합된 처리 과정을 나타내는 개념들의 관계를 나타낸다. 예를 들어, [그림 2]는 이 온톨로지의 일부를 나타내고 있다. 이 온톨로지들은 part-of와 instance-of를 통해 개념 관계를 표현하지만, 본 논문에서는 이 두 관계의 포괄적인 의미로 하위 개념 관계로 간주하여 사용한다. 즉, 하나의 개념은 자신보다 보다 구체적인 의미의 개념들을 하위 개념으로 가질 수 있다. 예를 들어, CC온톨로지 개념 'cell'은 하위 개념으로 '*intracellular*'와 '*membrane*' 등을 가지고 있으며, MF온톨로지 개념 'binding'은 하위 개념으로 'acyl binding'과 'cofactor binding' 등을 가지고 있다. 또한, BP온톨로지 개념 'cellular process'는 하위 개념으로 'cell communication', 'cell growth' 등을 가진다.

하나의 단백질 상호작용 네트워크에 포함된 단백질들은 하나 이상의 온톨로지 용어들로 개념화될 수 있다. 예를 들어, 'cytochrome c'는 MF 용어 'electron transporter activity', BP 용어 'oxidative phosphorylation' 과 '*induction of cell death*' 그리고 CC 용어 'mitochondrial matrix'과 'mitochondrial inner membrane'으로 개념화될 수 있다. 따라서, 단백질 상호작용 네트워크는 [그림 3] 그리고 [그림 4]와 같이 2개의 관계형 테이블로 표현될 수 있다. MF_SET, BP_SET 그리고 CC_SET은 3개의 온톨로지 개념들에 대한 집합이고 PID와 BID는 단백질 ID에 대한 외부 키(foreign key)이다. 또한, TYPE은 해당 DRECTION에 대한 역할을 나타낸다.

ID	NAME	MF_SET	BP_SET	CC_SET
100	*cytochrome c*	electron *transporter* activity	oxidative phosphorylation, induction of cell death	mitochondrial matrix, mitochondrial inner membrane
101	...	...	...	...

[그림 3] 단백질 표현

PID	*BID*	*DRECTION*	*TYPE*
100	101	1	activated
101	102	0	Inhibited
...	...	...	...

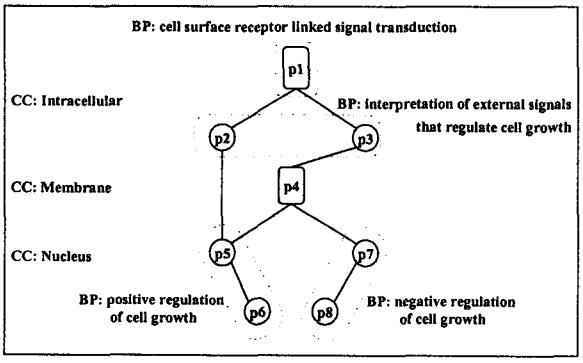
[그림 4] 상호작용 네트워크 표현

3. 네트워크 개념화

이장에서는 하나의 상호작용 네트워크를 온톨로지를 통해 개념화하는 과정을 설명한다. 즉, 네트워크가 복잡할 경우 사용자가 전체적인 관점에서 그 내용을 이해하기란 매우 어렵다. 따라서, 개념화는 네트워크 노드들을 개념으로 표시하고 유사한 의미의 노드들을 서로 클러스터링하여 일반적인 의미의 네트워크로 표현하게 된다. 다음은 이 네트워크를 개념화하기 위한 3단계 과정을 예를 통해 설명한다. 이 예에서는 설명의 편의를 위해 CC와 BP온톨로지만을 고려하였다.

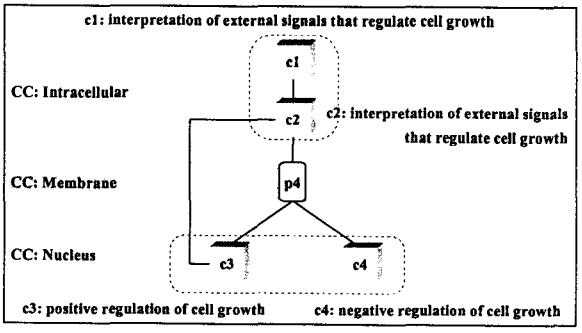
[그림 5]는 'Cell Growth'와 관련된 단백질들에 대한 상호작용 네트워크를 BP온톨로지를 통해 개념화하기 위한 1단계이다. p[1-9]는 단백질을 의미한다. 1단계에서 p[1]는 CC 개념 'Intracellular'과 BP 개념 'Cell surface receptor linked signal transduction', p[2-3]는 CC 개념 'Intracellular'과 BP 개념 'Interpretation of external signals that regulate cell growth', p4는 CC 개념 'Membrane', p[5-6]은 CC 개념 'Nucleus'과 BP 개념 'Positive regulation of

cell growth' 그리고 p[7-8]은 CC 개념 'Nucleus'과 BP 개념 'Negative regulation of cell growth'으로 각각의 단백질이 가지는 특성으로 네트워크 노드 레이블을 대치한다.



[그림 5] 온톨로지를 이용한 네트워크 개념화 1단계

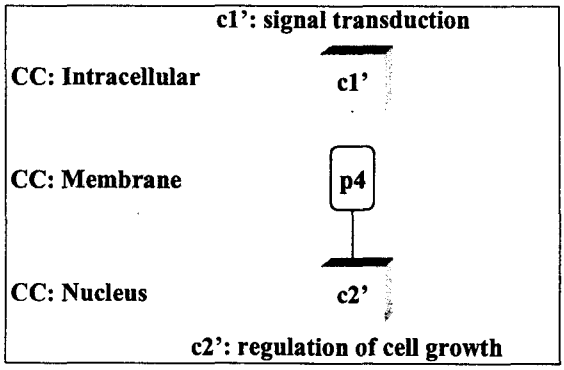
[그림 6]은 네트워크 개념화 2단계의 결과를 나타내고 있다. 여기서, c[1-4]은 개념 노드를 나타낸다. 즉, CC 온톨로지 개념 레벨을 고정시키고, BP 온톨로지를 이용하여 같은 BP 용어로 개념화된 단백질들을 하나의 개념 노드로 클러스터링한다. 만약, p2와 p5가 같은 BP 용어를 가지고 있을지라도 CC 용어가 서로 다르기 때문에 하나의 개념 노드로 클러스터링 될 수 없다.



[그림 6] 온톨로지를 이용한 네트워크 개념화 2단계

이 개념화된 네트워크는 BP 온톨로지 개념 레벨에 따라 보다 높은 상위 레벨로 개념화될 수 있다. 즉, BP 온톨로지에서 두 개념 'interpretation of external signals that regulate cell growth'과 'interpretation of external signals that regulate cell growth'을 동시에 포함하면서 가장 구체적이 상위 개념으 'signal transduction'이 존재하기 때문에 두 노드 c1과 c2는 다시 하나의 노드로 클러스터링 될 수 있다. 또한, 'positive regulation of cell growth'과 'negative regulation of cell growth'을 포함하는 가장 구체적인 상위 개념이 'regulation of cell growth'이기 때문에 c3과 c4 역시 하나로 클러스터링 될 수 있다. 이 과정은 [그림 7]과 같이 개념화 3단계를 통해 수행되며, 이 단계는 사용자가 요구에 따라 반복 수행될 수 있다.

이 개념화 단계에서 CC 개념 레벨을 적절한 수준에서 고정할 필요가 있다. 즉, BP용어에 의해 클러스터링되는 노드 수가 CC 개념 레벨이 너무 낮을 경우 매우 적을 수 있으며, 너무 클 경우 너무 많을 수 있다. 또한, CC와 BP의 레벨을 고정시키고 MF 온톨로지를 이용한 개념화 역시 같은 방법으로 수행될 수 있다.



[그림 7] 온톨로지를 이용한 네트워크 개념화 3단계

4. 결론 및 향후 연구

본 논문에서는 방대한 단백질들 사이의 복잡한 상화작용 관계 네트워크를 개념화하기 위한 방법을 제안하였다. 이를 위해 먼저 하나의 상호작용 관계 네트워크에 포함된 각각의 단백질들을 구성요소(CC), 생물학적 작용(BP) 그리고 분자 기능(MF) 3가지 특징으로 개념화한다. 다음으로 비슷한 특징으로 개념화된 단백질들은 서로 클러스터링될 수 있기 때문에 이 네트워크를 보다 일반적인 의미의 개념 네트워크로 표현할 수 있다. 이 단계를 반복 수행함으로써 복잡한 네트워크에 대한 개념화 레벨을 조절할 수 있다. 여기서, 구체적인 의미의 네트워크의 노드들을 일반적인 의미로 개념화하기 위해 유전자 온톨로지(Gene Ontology)를 이용하였다.

참고문헌

[1] C. L. Tucker, J. F. Gera, and P. Uetz, "Towards an Understanding of Complex Protein Interaction Maps," Trends in Cell Biology, Vol. 11, No. 23, 2001.

[2] S. Field, and O. Song, "A Novel Genetic System to Detect Protein-Protein Interactions," Nature 340: 245-247, 1989.

[3] P. Uetz, T. Ideker, and B. Schwikowski, "Visualization and Integration of Protein-Protein Interactions," Cold Spring Harbor Laboratory Press, pp. 623-646, 2002.

[4] B. Schwikowski, P. Uetz, and S. Fields, "A Network of Protein-Protein Interactions in Yeast," Nature Biotechnology, Vol. 18, No. 12, pp. 1257-1261, 2000.

[5] S. Oliver, "Guilt-by-Association Goes Global," Nature-News and Views, Vol. 403, pp. 601-603, 2000.