단백질 상호작용 네트워크를 위한 개념 기반 필터링

최재훈° 박종민 정재영 박선희
한국전자통신연구원
{jhchoi°, jmpark93, jjy72,psh}@etri.re.kr

A Concept-Based Filtering for Protein-Protein Interaction Networks

Jae-Hun Choi° , Jong-Min Park, Jae-Young Jeong, Seon-Hee Park
Electronic Telecommunication Research Institute(ETRI)

요 약

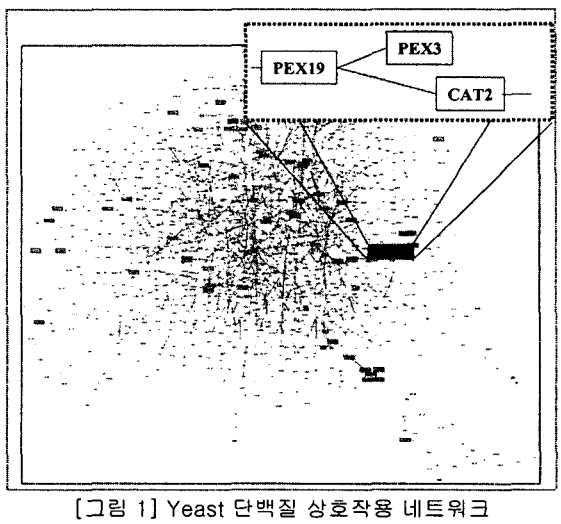
본 논문은 생명체 세포에 존재하는 방대한 단백질들 사이의 상호작용 관계들로 표현되는 네트워크에서 사용자가 관심있는 부분 네트워크를 개념적으로 필터링할 수 있는 방법을 설계하고 구현하였다. 이 방법은 1) 유전자 온톨로지를 이용하여 필터링 조건을 입력하고, 2) 이 조건을 만족하는 단백질들을 네트워크에서 필터링한 다음, 3) 이 단백질들 중 사용자가 관심이 있는 단백질만 선택하고, 4) 선택된 단백질들과 일정 거리에 있는 상호작용 관계들을 필터링함으로서 수행된다. 네트워크 필터링은 생물학자가 방대한 네트워크에서 자신이 관심이 있는 단백질들과 이들 사이의 관계에만 집중할 수 있도록 지원한다.

1. 서론

현재 대부분의 생명 과학 연구는 "하나의 유전자가 하나의 단백질을 만들고, 하나의 단백질이 하나의 기능을 수행한다."는 기존의 일차원적인 범위를 벗어나 복잡한 생물학적 기능을 단백질들 사이의 상호작용을 통해 규명하려는데 초점을 맞추고 있다. DNA에 포함된 유전자 정보가 발현되어 최종적으로 생성되는 물질로서 단백질은 다른 단백질과의 상호 유기적인 작용을 통해 신호전달(signal transduction), 세포 생명 주기(cell life cycle), 세포 분화(cell development), DNA 복제(replication), 물질대사(metabolism) 등과 같은 세포의 생리활성 반응을 조절하게 된다. 이 상호작용을 단백질들 사이의 관계로 나타내면 네트워크 형태로 표현된다.

일반적으로 단백질들 사이의 상호작용 관계는 이스트 투 하이브리드(yeast two hybrid)라는 생물학적 실험을 통해 추출되고 있다[1]. 이 실험에서 하나의 단백질(bait protein)은 보고 유전자(reporter gene)의 프로모터와 결합할 수 있는 DNA 결합 부위(DNA binding domain)을 갖도록 발현시키고, 다른 단백질(prey protein)은 이 보고 유전자를 발현시킬 수 있는 전사 활성화 부위(transcription activating domain)를 갖도록 발현시킨다. 따라서, 두 단백질의 상호작용을 하게 된다면 보고 유전자를 발현됨으로 상호관계를 알 수 있다. 현재, 이 실험을 통해 구축된 정보는 데이터베이스에 체계적으로 관리고 있으며, 대표적으로 PIM, BIND, DIP, GRID 등이 있다.

단백질 상호작용 네트워크는 매우 방대한 단백질들과 이들 사이의 복잡한 관계로 되어있어며, 생물체의 종(species), 조직(tissue) 또는 세포 주기(cell cycle) 에 따라 각각 참여하는 단백질뿐만 아니라 이들 사이의 관계가 다를 수 있다. [그림 1]은 [2]에서 제시한 2358개의 효모(yeast) 단백질 상호작용 네트워크를 나타내고 있다. 사용자들이 비록 전문가일지라도 이와 같이 방대한 효모 네트워크를 모두가 아닌 단지 일부분에만 관심을 가지고 생물실험을 한다[3]. 예를 들어, 단백질 PEX19, PEX3 그리고 CAT2들 사이이 상호작용 관계에만 관심을 가질 수 있다. 이를 위해 Osprey에서는 정확 일치에 의한 필터링 방법을 제공하고 있다. 그러나, 이 방법은 개념적으로 일치하는 부분 네트워크를 필터링할 수 없다는 단점을 가지고 있다.



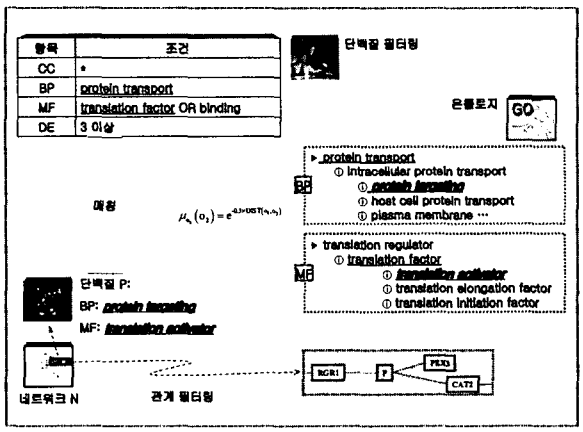
[그림 1] Yeast 단백질 상호작용 네트워크

따라서, 본 논문에서는 생물체의 세포에 존재하는 방대

한 단백질들 사이의 복잡한 상호작용 네트워크에서 사용자가 관심이는 단백질들과 상호작용 관계들을 개념적으로 필터링할 수 있는 방법을 설계하고 구현하였다.

2. 개념기반 네트워크 필터링

개념 기반 네트워크 필터링을 위해서는 네트워크에 포함된 단백질들이 통제 용어(Controlled Vocabulary)들을 포함하고 있어야 한다. 또한, 이 통제 용어들 사이의 의미적 관계를 정의한 온톨로지가 요구된다. 이를 위해 본 논문에서는 Swiss Prot에 있는 단백질 데이터베이스와 유전자 온톨로지(Gene Ontology)를 사용하였다. Swiss Prot는 공개된 단백질들에 관한 많은 정보들을 포함하고 있는 데이터베이스이며, 단백질 각각에 유전자 온톨로지 용어를 할당해 놓고 있다. 유전자 온톨로지는 통제 용어들을 3가지 관점(BP: Biological Process, CC: Cellular Component, MF: Molecular Function)에 따라 분류하고 이들 사이의 의미적 관계들을 계층적으로 표현하고 있다.



[그림 2] 네트워크 필터링 과정

네트워크 필터링은 단백질 필터링과 상호작용 관계 필터링으로 구성되어 있으며, 이들은 다음과 같은 처리 과정에 따라 수행된다. 첫째, 온톨로지를 참조하여 사용자가 직접 단백질 필터링 조건을 입력한다. 이때, 조건은 온톨로지 용어들의 불리언 연산자로 표현한다. 둘째, 시스템은 이 조건을 만족하는 단백질들을 하나의 네트워크에서 필터링한다. 이때, 단백질에 항당된 3 분류의 온톨로지 용어들과 필터링 조건의 온톨로지 용어들을 개념적으로 일치시킨다. 즉, 두 용어가 정확하게 일치하지 않더라도 온톨로지에서 개념적으로 유사하다고 판단되면 일치될 수 있다. 셋째, 사용자가 필터링된 단백질들의 정보를 참조하여 관심이 있는 단백질만 선택한다. 이를 위해서는 Swiss Prot 데이터베이스에 대한 상호참조가 가능해야 한다. 넷째, 선택된 단백질들과 일정 거리에 있는 상호작용 관계들을 필터링함으로서 수행된다. 하나의 단백질과 상호작용 관계를 가지는 다른 단백질들은 서로 기능이 유사할뿐만 아니라 서로 긴밀한 관계를 가지고 있다. 따라서, 시스템은 선택된 단백질과 일정 거리에 있는 단백질 및 이들 사이의 관계를 필터링하여 사용자에게 제공한다. [그림 2]는 이 과정을 설명하고 있다.

단백질 필터링 조건은 4개의 항목(CC, BP, MF, DE)으로 구성되어 있다. CC 조건은 무조건 항 '*'이고 MF과 BP은 온톨로지의 통제 용어들에 불리언 표현으로 각각 'translation factor' OR 'binding'과 'protein transport'이다. 이 개념 조건은 사용자가 온톨로지를 직접 참조하면서 구성할 수 있다. DE 조건은 네트워크에서 단백질의 차수(degree)를 설정하는 항목이다. 이 차수는 "n차", "n차 이상" 그리고 "n차 이하"와 같은 형식으로 입력할 수 있다. 이 예에서는 다른 단백질과 3개 이상 관계를 가지는 단백질만을 필터링하고 있다. 즉, 이 4개의 항목을 모두 만족하는 단백질들만을 네트워크에서 필터링하게 된다.

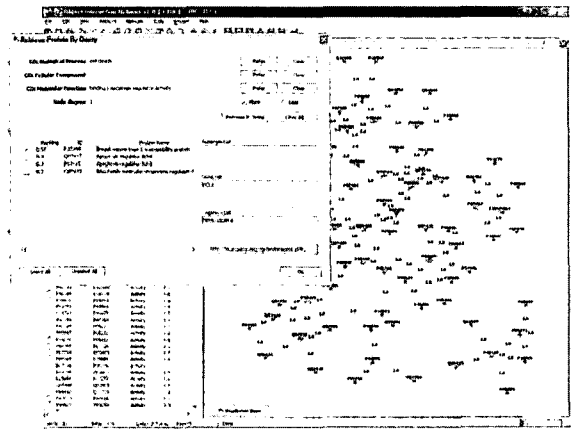
차수 조건에 의한 필터링은 네트워크에서 각각의 노드가 가지는 상호작용 개수에 따라 쉽게 수행될 수 있다. 이 예에서 네트워크 N의 단백질 P는 차수 조건 "3 이상"을 만족함으로 필터링 된다. 용어 조건에 의한 필터링은 단백질들이 가지는 용어들과의 개념 일치를 통해 수행된다. 이 예에서 MF 조건은 분자적 기능이 'translation factor' 또는 'binding'인 단백질들을 필터링 하게 된다. 기존의 정확 일치 방법을 사용한다면 이 조건으로 분자적 기능이 'translation activator'인 단백질 P를 필터링 할 수 없다. 그러나, 유전자 온톨로지에서 'translation factor'와 'translation activator'은 개념적으로 상당히 유사한 용어이기 때문에 본 논문에서 제시한 개념 일치 방법으로 P를 필터링 할 수 있다. 같은 방법으로 BP 조건의 'protein transport'와 P의 BP 용어 'protein targeting'이 온톨로지에서 개념적으로 일치하기 때문에 P는 BP에 의해서도 필터링된다.

두 용어 v1과 v2가 서로 의미적 관계를 가지기 위해서는 온톨로지에서 일정한 거리에 존재해야 한다. 이때, 이 두 용어 사이의 의미적 거리 $S_{1,2}$는 일반적으로 $s_{1,2}=e^{-0.3\times D_{1,2}}$ 에 의해 계산된다. 여기서, $D_{1,2}$는 온톨로지 계층에서 두 용어 사이의 거리로 'protein transport'와 'protein targeting'의 거리는 2이다.

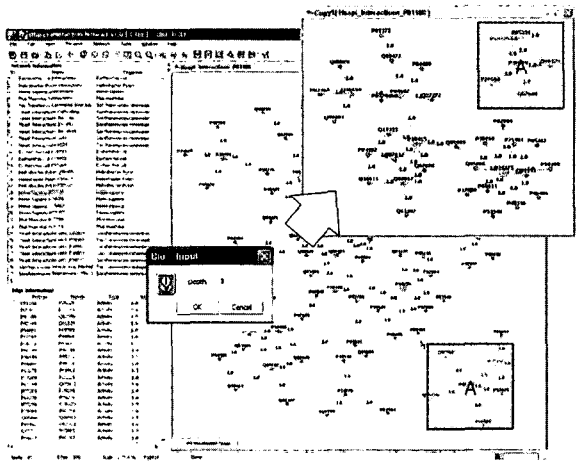
단백질 필터링 조건에 의해 단백질 P는 다음과 같이 평가된다. 즉, P는 BP 조건을 $S_{\text{'protein targeting', 'protein transport'}}= e^{-0.3\times2}=0.55$ 정도 만족시킨다. 또한, MF 조건을 $S_{\text{'translation activator', 'translation factor'}}= e^{-0.3*1}=0.74$ 또는 $S_{\text{'translation activator', 'binding'}}= e^{-0.3*\infty}=0$ 정도 만족시키기 때문에 max(0.74,0)으로 평가될 수 있다. 이때, AND, OR 연산은 min과 max 함

수로 계산할 수 있다. 따라서, P는 단백질 필터링 조건을 min(0.74, 0.55)=0.55 정도로 필터링 된다.

사용자는 개념 기반으로 필터링된 단백질들에 대한 Swiss Prot 정보를 참조함으로써 이들 중 관심이 있는 단백질만을 선택할 수 있다. [4]에 의하면 이 선택된 단백질들과 상호작용을 하는 다른 단백질들 역시 유사한 역할에 참여하고 있기 때문에 사용자의 관심은 이 단백질들과 상호작용 관계를 가지는 다른 단백질들로 확장될 수 있다. 따라서, 관계 필터링은 선택된 단백질들과 일정 거리 범위 안에 있는 상호작용 관계들을 추출함으로써 수행된다. 이 때, 사용자는 관심이 있는 다른 단백질이나 상호작용 관계를 직접 선택하여 추출할 수 도 있다.



[그림 3] 단백질 필터링 과정



[그림 4] 관계 필터링 과정

3. 구현

이 절에서는 네트워크 필터링의 구현 내용을 단백질 필터링과 관계 필터링으로 구분하여 설명한다. 먼저, [그림 3]은 조건을 입력하여 단백질을 필터링하는 과정을 나타내고 있다. 예를 위해 인간 단백질 상호작용 네트워크의 일부를 이용하였다.

여기서, 단백질 필터링에 대해 차수 조건은 "1 이상", BP 조건은 'cell death', 그리고 MF 조건은 "'bind' OR 'apoptosis regulator activity'"으로 각각 설정하였다. 이 결과로 4개의 단백질이 필터링 되었으며, 각각의 필터링 조건과의 개념적 관련 정도에 따라 순위화될 수 있다.

[그림 4]는 필터링된 하나의 단백질 P10415와 거리가 3 이하 범위 내에서 상호작용을 하는 모든 단백질 및 관계를 필터링한 결과이다. 여기서, 사용자는 여러 단백질을 대상으로 관계 필터링을 수행할 수 있으며, 관계 필터링 거리는 사용자가 직접 입력한다. 또한, "A" 지역과 같이 시스템에 의해 필터링되어 선택된 관계들에 사용자가 직접 필요한 다른 관계들을 추가적으로 선택할 수 있다. 이 선택된 관계들은 다른 윈도우에 적절하게 시각화함으로써 사용자가 관심이 있는 단백질들과 이들 사이의 관계에만 집중할 수 있도록 지원한다.

4. 결론

본 논문은 생명체 세포에 존재하는 방대한 단백질 상호작용 네트워크에서 사용자가 관심이 있는 부분의 단백질들과 이들사이의 관계를 개념적으로 필터링할 수 있는 방법을 설계하고 구현하였다. 이를 위해 사용자는 먼저 유전자 온톨로지를 이용하여 관심이 있는 단백질들을 개념적으로 필터링 한 다음, 이 단백질과 일정한 거리에 있는 관계들을 다시 필터링할 수 있다. 또한, 사용자는 직접 원하는 단백질과 관계들을 다양한 방법으로 선택함으로써 부분 네트워크들을 필터링할 수도 있다. 이때, 단백질 필터링 조건은 유전자 온톨로지의 용어들에 대한 불리언 표현으로 기술된다.

참고문헌

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

[2] J. J. Han etc, "Evidence for Dynamically Organized Modularity in the Yeast Protein-Protein Interaction Network," Nature 430: 88-94, 2004.

[3] 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.

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